Projektas

„Lietuvos HSM duomenų archyvo LiDA plėtra“

SFMIS Nr. VP1-3.1-ŠMM-02-V-02-001

TAIKOMOJI REGRESINĖ ANALIZĖ SOCIALINIUOSE TYRIMUOSE

Vydas Čekanavičius

(Paslaugų sutartis Nr. SA-2010-773 - 2, 2010-11-12)

Kaunas, 2011 m.

Turinys

Pratarmė **11**

1. ĮVADAS **14**

1.1. Kintamieji ir jų charakteristikos **14**

1.1.1. Kintamųjų matavimo skalės 14

1.1.2. Svarbiausios kintamųjų skaitinės charakteristikos 15

1.1.3. Normalieji kintamieji 18

1.2. Statistinės išvados **19**

1.2.1. Parametrų įverčiai ir jų pasikliautinieji intervalai 19

1.2.2. Statistinė hipotezė 19

1.2.3. Pastaba apie naudojamus terminus 21

2. TIESINĖ REGRESINĖ ANALIZĖ **22**

2.1. Tiesinės regresijos modelis **22**

2.1.1. Modelio lygtis 22

2.1.2. Prognozavimas 23

2.1.3. Reikalavimai duomenims 24

2.1.4. Modelio tinkamumas 25

2.1.5. Tiesinės regresinės analizės etapai 29

2.1.6. Modelio tobulinimas 29

2.1.7. Pastabos apie modelio taikymą 31

2.1.8. Standartiniai tiesinės regresijos modelio aprašymai 32

2.2. Tiesinė regresinė analizė su SPSS **33**

2.2.1. Duomenys 33

2.2.2. Pirminio modelio tyrimas 34

2.2.3. Pirminio modelio tyrimo rezultatai 36

2.2.4. Šapiro – Vilko testas SPSS 41

2.2.5. Modelio tobulinimas 42

2.2.6. Prognozės ir stebėjimų grafinis palyginimas 45

2.2.7. Rezultatų aprašymas ... 46
2.2.8. SPSS komandų sintaksė tiesinei regresinei analizei ... 47
2.2.9. Autokoreliacijos tyrimas ... 48
2.2.10. Multikolinearumo problema ... 49
2.2.11. Žingsninė regresija ... 49
2.3. Tiesinė regresinė analizė su STATA ... 52
2.3.1. Duomenys ... 52
2.3.2. Regresijos modelio lygtis ... 54
2.3.3. Regresijos prielaidų tikrinimas ... 55
2.3.4. Galutinis modelis ... 58
2.3.5. Prognozės ir stebėjimų grafinis palyginimas ... 59
2.3.6. Išvados ... 60
2.3.7. Autokoreliacija ... 61
2.3.8. Multikolinearumo problema ... 61
2.3.9. Žingsninė regresija ... 63
Uždaviniai ... 67
3. TIESINĖS REGRESIJOS ALTERNATYVOS ... 70
3.1. Tiesinei regresijai netinkanys duomenys ... 71
3.1.1. Heteroskedastiški ir ne visai normalūs duomenys ... 71
3.1.2. Duomenys su išskirtimis ... 74
3.1.3. Duomenų charakteristikų santrauka ... 75
3.2. Stabilizuotų liekamųjų paklaidų regresija ... 76
3.2.1. Modelio paskirtis ... 76
3.2.2. Liekamųjų paklaidų HC korekcijos su STATA ... 77
3.2.3. Savirankos regresija su STATA ... 79
3.2.4. Savirankos regresija su SPSS ... 80
3.3. Atsparioji regresija ... 83
3.3.1. Modelio paskirtis ... 83

3.3.2. Atsparioji regresija su STATA 84

3.4. Medianos regresija 86

3.4.1. Modelio paskirtis 86

3.4.2. Medianos regresija su STATA 87

3.5. Netiesinė regresija 93

3.5.1. Modelio paskirtis 93

3.5.2. Netiesinė regresija su SPSS 95

3.5.3. Netiesinė regresija su STATA 99

Uždaviniai 100

4. DVINARĖ LOGISTINĖ REGRESINĖ ANALIZĖ 103

4.1. Dvinarės logistinės regresijos modelis 103

4.1.1. Modelio lygtis 103

4.1.2. Galimybių santykis 104

4.1.3. Prognozavimas 106

4.1.4. Reikalavimai duomenims 107

4.1.5. Modelio tinkamumas 107

4.1.6. Dvinarės logistinės regresinės analizės etapai 109

4.1.7. Modelio tobulinimas 110

4.1.8. Pastabos apie logistinės regresijos taikymą 111

4.1.9. Standartiniai regresijos modelių aprašymai 112

4.2. Dvinarė logistinė regresinė analizė su SPSS 112

4.2.1. Duomenys 112

4.2.2. SPSS parinktys pradiniam regresijos modeliui 114

4.2.3. Rezultatai 115

4.2.4. Modelio tobulinimas 120

4.2.5. Rezultatų aprašymas 125

4.2.6. SPSS komandų sintaksė 126

4.2.7. Regresorių sąveika 126

4.2.8. Viena dominuojanti priklausomo kintamojo reikšmė 127
4.2.9. Multikolinearumas 127
4.2.10. Žingsninė logistinė regresija 128
4.2.11. Apibendrintas tiesinis modelis su logit jungtimi 132
4.3. Dvinarė logistinė regresinė analizė su STATA 138
4.3.1. Duomenys 138
4.3.2. Modelio tyrimas 141
4.3.3. Atsparioji logistinė regresija 145
4.3.4. Reikšmių prognozavimas 146
4.3.5. Kategoriniai regresoriai 147
4.3.6. Multikolinearumas 148
4.3.7. Žingsninė logistinė regresija 150
Uždaviniai 151
5. DAUGIANARĖ LOGISTINĖ REGRESINĖ ANALIZĖ 154
5.1. Daugianarės logistinės regresijos modelis 154
5.1.1. Modelio lygtys 154
5.1.2. Prognozavimas 156
5.1.3. Galimybių santykiai 156
5.1.4. Duomenys 157
5.1.5. Modelio tinkamumas 158
5.1.6. Daugianarės logistinės regresinės analizės etapai 159
5.1.7. Modelio tobulinimas 160
5.1.8. Pastabos apie multinominės regresijos taikymą 161
5.1.9. Standartiniai daugianarės logistinės regresijos modelių aprašymai 162
5.2. Daugianarė logistinė regresinė analizė su SPSS 162
5.2.1. Duomenys 162
5.2.2. SPSS parinktys daugianarei logistinei regresijai 163
5.2.3. Rezultatai 165

5.2.4. Prognozavimas 170
5.2.5. Rezultatų aprašymas 170
5.2.6. SPSS komandų sintaksė daugianarei logistinei regresinei analizei 171
5.2.7. Kitos kontrolinės kategorijos pasirinkimas 172
5.2.8. Kintamųjų sąveikos įtraukimas 173
5.2.9. Žingsninė daugianarė logistinė regresija 174
5.3. Daugianarė logistinė regresinė analizė su STATA 176
5.3.1. Duomenys 176
5.3.2. STATA parinktys 177
5.3.3. Prognozavimas 179
5.3.4. Klasifikacinė lentelė 179
5.3.5. Kategorinis kintamasis 180
Uždaviniai 182
6. RANGINĖ LOGISTINĖ REGRESINĖ ANALIZĖ 183
6.1. Ranginės logistinės regresijos modelis 183
6.1.1. Modelio lygtys 183
6.1.2. Galimybės ir jų santykis 184
6.1.3. Prognozavimas 185
6.1.4. Duomenys 186
6.1.5. Modelio tinkamumas 187
6.1.6. Modelio tyrimas ir tobulinimas 188
6.1.7. Pastabos apie ranginės regresijos taikymą 189
6.1.8. Standartiniai daugianarės logistinės regresijos modelių aprašymai 190
6.2. Ranginė logistinė regresinė analizė su SPSS 190
6.2.1. Duomenys 190
6.2.2. SPSS parinktys ranginei logistinei regresijai 192
6.2.3. Rezultatai 193
6.2.4. Klasifikacinė lentelė 196

6.2.5. Rezultatų aprašymas 197
6.2.6. SPSS komandų sintaksė daugianarei logistinei regresinei analizei 198
6.2.7. Ranginė logistinė regresija su kintamųjų sąveika 199
6.2.8. Ranginė regresija su kitomis jungties funkcijomis 200
6.3. Ranginė logistinė regresinė analizė su STATA 200
6.3.1. Duomenys 200
6.3.2. Modelio analizė 201
6.3.3. Prognozavimas 204
6.3.4. Klasifikacinė lentelė 206
Uždaviniai 207
7. PUASONO REGRESINĖ ANALIZĖ 208
7.1. Puasono regresijos modelis 208
7.1.1. Modelio lygtis ir jos taikymas prognozavimui 208
7.1.2. Modelis dažniui 209
7.1.3. Duomenys 210
7.1.4. Modelio tinkamumas 211
7.1.5. Puasono regresinės analizės etapai 212
7.1.6. Modelio tobulinimas 213
7.1.7. Pastabos apie Puasono regresijos taikymą 213
7.1.8. Standartiniai Puasono regresijos modelio aprašymai 214
7.2. Puasono regresinė analizė su SPSS 215
7.2.1. Duomenys 215
7.2.2. Preliminari analizė 216
7.2.3. SPSS parinktys Puasono regresijos modeliui 219
7.2.4. Rezultatai 220
7.2.5. SPSS komandų sintaksė Puasono regresinei analizei 223
7.2.6. Kategorinis regresorius 224
7.2.7. Puasono regresinė analizė dažniui 225

7.3. Puasono regresinė analizė su STATA 226

7.3.1. Duomenys 226

7.3.2. Puasono modelio taikymas 227

7.3.3. Prognozavimas 229

7.3.4. Kategorinis regresorius 230

Uždaviniai 230

8. NEIGIAMA BINOMINĖ REGRESIJA 232

8.1. Neigiamos binominės regresijos modelis 232

8.1.1. Modelio lygtis ir jos taikymas prognozavimui 232

8.1.2. Duomenys 233

8.1.3. Modelio tinkamumas 234

8.1.4. Neigiamos binominės regresinės analizės etapai 234

8.1.5. Modelio tobulinimas 235

8.1.6. Pastabos apie modelio taikymą 236

8.1.7. Standartiniai neigiamos binominės regresijos modelio aprašymai 236

8.2. Neigiama binominė regresinė analizė su SPSS 237

8.2.1. Duomenys 237

8.2.2. SPSS parinktys neigiamos binominės regresijos modeliui 238

8.2.3. Rezultatų aptarimas 239

8.2.4. Rezultatų aprašymas 242

8.2.5. SPSS komandų sintaksė 243

8.2.6. Neigiama binominė regresinė analizė dažniui 244

8.2.7. Neigiamos binominės regresinės analizės standartinė opcija SPSS 244

8.3. Neigiama binominė regresinė analizė su STATA 245

8.3.1. Duomenys 245

8.3.2. Modelio parametrų analizė 246

8.3.3. Prognozavimas 248

8.3.4. Kategorinis kintamasis 249

8.3.5. GLM alternatyva 250

Uždaviniai 251

9. MODIFIKUOTI SKAIČIUOJAMIEJI MODELIAI 252

9.1. Perteklinių nulių Puasono regresijos modelis 252

9.1.1. Modelio paskirtis 252

9.1.2. Perteklinių nulių Puasono regresija su STATA 254

9.2. Perteklinių nulių neigiama binominė regresija 256

9.2.1. Modelio paskirtis 256

9.2.2. Perteklinių nulių neigiama binominė regresija su STATA 257

9.3. Skaičiuojamų regresijos modelių palyginimas su STATA 258

9.4. Benulė Puasono regresija 261

9.4.1. Modelio paskirtis 261

9.4.2. Benulė Puasono regresija su STATA 262

Uždaviniai 264

10. PROBIT REGRESIJA 265

10.1. Probit regresijos modelis 265

10.1.1. Modelio lygtis 265

10.1.2. Reikalavimai duomenims 266

10.1.3. Modelio tinkamumas 267

10.1.4. Bendrosios probit regresinės analizės etapai 267

10.1.5. Modelio tobulinimas 268

10.1.6. Dozavimo modelis 269

10.1.7. Standartiniai regresijos modelių aprašymai 270

10.2. Probit regresinė analizė su SPSS 270

10.2.1. Duomenys 270

10.2.2. SPSS parinktys pradiniam regresijos modeliui 272

10.2.3. Rezultatai 274

10.2.4. Prognozavimas 277

10.2.5. Dozavimo modelio tyrimas ... 278
10.2.6. SPSS komandų sintaksė ... 282
10.3. Probit regresinė analizė su STATA ... 282
10.3.1. Duomenys ... 282
10.3.2. Modelio tyrimas ... 283
10.3.3. Reikšmių prognozavimas ... 286
Uždaviniai ... 286
11. LOGTIESINIAI MODELIAI ... 288
11.1. Logtiesinio modelio struktūra ir savybės ... 288
11.1.1. Modelio lygtis ... 288
11.1.2. Rezultatų interpretavimas ... 289
11.1.3. Galimybių santykis ... 290
11.1.4. Reikalavimai duomenims ... 291
11.1.5. Modelio tinkamumas ... 291
11.2. Logtiesinių modelių tyrimas su SPSS ... 292
11.2.1. Duomenys ... 292
11.2.2. SPSS parinktys hierarchiniam logtiesiniam modeliui ... 293
11.2.3. SPSS parinktys logtiesiniam modeliui ... 295
11.2.4. Rezultatai ... 297
11.2.5. Išvados ... 298
11.2.6. Intervalinis kintamasis ... 299
11.3. Logtiesiniai modeliai su STATA ... 301
11.3.1. Duomenys ... 301
11.3.2. Analizė ... 302
Uždaviniai ... 304
Duomenų šaltiniai ... 306
Literatūra ... 307

Pratarmė

Įsivaizduokime, kad nusipirko žmogus naują televizorių ir nori truputį pareguliuoti spalvas ir garsą. O vietoje pultelio gauna atsuktuvą ir elektroninę televizoriaus schemą. Skaityk, analizuok ir daryk! O jei neišmanai fizikos, tai ir nagų nekišk! Vargu ar toks pirkėjas būtų laimingas. O štai dauguma statistikos vadovėlių yra, kaip ta elektroninė schema – viskas juose matematiškai tikslu ir teisinga, bet praktines problemas spręsti nelabai padeda[1]. Kokie klausimai dažniausiai kyla, atliekant statistinę analizę? Tikrai ne apie įverčių gavimo metodo subtilybes, ar Gauso-Markovo teoremos įrodymą. Svarbiausias klausimas – ar šįkart gerą modelį gavau, ar (ir vėl) kažkokią nesąmonę? Išmintingos vadovėlinės tiesos – dviejų deviacijų skirtumas turi asimptotinį chi kvadrato skirstinį ir pan. – labiau ugdo neapykantą matematikai, nei padeda į jį atsakyti.

Šio studijų paketo tikslas – padėti įgyti tinkamus **praktinius** regresinio modeliavimo ir analizės įgūdžius. Vadovėlių[2], kuriuose išdėstyta regresinių modelių teorija, šie konspektai tikrai nepakeis, bet uždavinius spręsti pagelbės. Sąmoningai vietoje tikslių ir baisių (gražių) matematinių formulių pateikiami itin supaprastinti pagrindinių teorinių idėjų atpasakojimai.

Studijų paketas yra skirtas šiems regresijos modeliams ir jų apibendrinimams:

- **Tiesinė regresija**. Modelyje vieną intervalinį kintamąjį veikia vienas arba keli kiti kintamieji. Pavyzdžiui, tiriame, kaip kraujospūdį veikia per dieną surūkytų cigarečių skaičius ir kūno masės indeksas. Aiškinamės, kaip migracija priklauso nuo BVP augimo.
- **Stabilizuotų liekamųjų paklaidų regresija.** Tiesinės regresijos alternatyva, kai duomenys nėra tolygiai išsibarstę (duomenys heteroskedastiški).
- **Atsparioji regresija.** Tiesinės regresijos alternatyva, kai yra labai išsiskiriančių stebėjimų (išskirčių).
- **Medianos regresija.** Tiesinės regresijos alternatyva, kai duomenys nėra normaliai pasiskirstę.
- **Netiesinė regresija.** Įvairios tiesinės regresijos alternatyvos, kai kintamųjų priklausomybės nėra tiesinės.
- **Dvinarė logistinė regresija.** Modelyje dvireikšmis kintamasis priklauso nuo vieno arba kelių kitų kintamųjų. Pavyzdžiui, tiriame, kaip politiko apsisprendimą kandidatuoti veikia jo

[1] Kaip taisyklė, vadovėliuose aprašoma, kaip išspręsti uždavinius be kompiuterinių programų („rankomis“). Regresijos modeliavimas „rankomis“ yrai skirtas žmonėms, kurie ne tik patys darbą myli, bet ir darbas juos.

[2] Pavyzdžiui, vadovėlio V.Čekanavičius, G.Murauskas „Statistika ir jos taikymai II“, TEV, 2002 (2004, 2008) m.

turtinė padėtis ir „politinis stažas“. Dvinarė logistinė regresija padeda įvertinti tikimybę susirgti diabetu, atsižvelgus į kūno masės indeksą ir paveldimumo faktorius.

- **Daugianarė logistinė regresija.** Tai dvinarės logistinės regresijos apibendrinimas, kai modeliuojamas kintamasis gali įgyti daugiau nei dvi kategorines reikšmes. Pavyzdžiui, bandome išsiaiškinti nuo ko priklauso, kaip balsuos rinkėjas.
- **Ranginė logistinė regresija.** Modelis analogiškas daugianarei logistinei regresijai, tik kintamieji ranginiai. Pavyzdžiui, bandome nustatyti, kaip požiūris į baigtas studijas (labai palankus – palankus – nepalankus) priklauso nuo gaunamo atlyginimo.
- **Puasono regresija**. Modeliuojame retų įvykių skaičiaus arba įvykių dažnio priklausomybę nuo vieno arba kelių kitų kintamųjų. Pavyzdžiui, tiriame, kaip vaikų skaičius šeimoje susijęs su tėvų išsilavinimu bei brolių/seserų skaičiumi.
- **Neigiama binominė regresija.** Modeliuojame įvykių skaičiaus arba įvykių dažnio priklausomybę nuo vieno arba kelių kitų kintamųjų. Alternatyva Puasono regresijai.
- **Perteklinių nulių modeliai.** Puasono regresijos modifikacijos, kai duomenys daug nulinių reikšmių. Pavyzdžiui, modeliuojame respondento pavaldinių skaičių.
- **Probit regresija.** Alternatyva dvireikšmei logistinei regresijai. Klasikinis probit regresijos taikymas – dozavino modelis. Pavyzdžiui, bandome nustatyti, kiek pesticidų reikia, kad būtų išnaikinta 90% kenkėjų.
- **Logtiesiniai modeliai**. Keleto kategorinių kintamųjų tarpusavio priklausomybių analizė. Pavyzdžiui, norime nustatyti, kaip požiūris į mirties bausmę priklauso nuo respondento lyties, tikėjimo ir tautybės. Apibendrina chi kvadrato kriterijų.

Norime atkreipti dėmesį, kad konspektuose aptariama tik klasikinė situacija, kai pasirenkamas *vienas* mus dominantis reiškinys, ir tiriama kas jam daro įtaką. Statistikoje yra taikomi ir sudėtingesni regresijos modelių apibendrinimai. Kai tyrėjo sudarytame modelyje yra net keli vienas kitą veikiantys kintamieji (kai kurie iš jų gali būti latentiniai, t.y. tiesiogiai nestebimi), naudojamos struktūrinės lygtys (SEM). Pavyzdžiui, modeliuodami šešėlinę ekonomiką, nurodome nuo ko šis tiesiogiai nestebimas kintamasis priklauso (nuo mokesčių, nedarbo lygio), ir kam daro įtaką (BVP, užimtumui). Duomenims, turintiems griežtą hierarchinę struktūrą, taikomi hierarchiniai tiesiniai modeliai (HLM). Pavyzdžiui, modeliuodami mokinių matematikos egzamino rezultatus, atsižvelgiame ne tik į jų pačių socialinį – ekonominį satusą, bet ir į mokyklos charakteristikas (geri mokytojai, griežta drausmė ir pan.). Lietuviškos literatūros apie HLM ir SEM yra labai nedaug[3].

[3] V.Čekanavičius, G.Murauskas „Statistika ir jos taikymai III“, TEV, 2009.

Savo pobūdžiu šie konspektai yra iš serijos „Gaminame patys. Receptai žaliems“. Kokio žalumo skaitytojams ji skirta? Skaitytojams apie statistiką pakanka turėti tik labai bendrą (ir miglotą) supratimą. Vis dėlto, tikėtis, kad nieko apie statistiką nežinojau (tik girdėjau, kad žmonės ją taiko, rašydami disertacijas bei straipsnius), o štai dabar paskaitinėsiu knygelę ir tapsiu regresijos ekspertu, – nevertėtų. (Pavartysiu anatomijos vadovėlį ir – tik žybt – sėkmingai išpjausiu kaimynui apendicitą). Ekspertu tampama praktikuojantis. Šie konspektai padės įgyti *pradinius* modeliavimo įgūdžius. Siekiama, kad tie įgūdžiai iš tikro būtų praktiniai. Kaip analizę atlikti SPSS ir STATA programomis aprašyta itin detaliai. Sąmoningai vengta patarimų „pasidaryk pats“ .

Pavyzdžiams tirti naudotos IBM SPSS 19 ir STATA 10 (su SPost priedu)[4] programos. Kitos šių programų versijos skiriasi tik detalėmis. Todėl neturėtų kilti kliūčių, tiriant regresijos modelius senesnėmis ir (tikėkimės) naujesnėmis programų versijomis. Analizei ir uždaviniams naudoti ESS4[5] ir LiDA[6] duomenų bazės duomenys. Į pavyzdžiuose gautus rezultatus nereikia žiūrėti, kaip į didžiai rimtus mokslinius tyrimus. Nebandėme aiškinti, *kodėl* atsirado viena, ar kita kintamųjų priklausomybė.Tai jau būtų sociologų duona. O mūsų tikslas – kitas. Siekiame išmokyti tas priklausomybes *pastebėti*, kitaip tariant, išmokti statistinės tyrimo technikos. Maždaug – kaip taisyklingai kalti vinis, nenusimušant pirštų ir neišdaužant žiūrovų. O ką kalti – kėdę prie stalo, vinį į sieną, kryžių, prie kryžiaus, – čia jau skaitytojo reikalas (kartais ir problema).

Regresijos modelių yra įvairiausių. Vis dėlto, nereikia turėti iliuzijų, kad bet kokiems duomenims pavyks sudaryti tinkamą regresijos modelį. Net labai stengiantis, nepavyks regresijos modeliu susieti Lietuvos BVP augimą su kapibarų skaičiumi Argentinoje. Yra duomenų, kuriems jokie regresiniai modeliai netinka.

Nuoširdžiai dėkojame R. Rauleckui už pastabas, padėjusias pagerinti šiuos konspektus.

Pabaigai – labai svarbi žinia:

Tikrinant visas hipotezes, reikšmingumo lygmuo lygus 0,05.

O ką tai reiškia, paaiškinta įvade.

[4] Šį priedą, skirtą poregresinei analizei primygtinai rekomenduoja patys STATA kūrėjai. SPost autoriai yra Scott Long ir Jeremy Freese iš Indianos universiteto, žr. http://www.indiana.edu/~jslsoc/web_spost/sp_install.htm

[5] European Social Survey, http://www.europeansocialsurvey.org/

[6] Lietuvos HSM duomenų archyvas (LiDA), http://www.lidata.eu/

1. ĮVADAS

Įvade labai trumpai priminsime pagrindines statistikos sąvokas, naudojamas kituose skyriuose.

1.1. Kintamieji ir jų charakteristikos

1.1.1. Kintamųjų matavimo skalės

Norint tinkamai pasirinkti statistinį modelį, reikia, kad būtų tenkinami tam tikri formalūs reikalavimai. Aprašydami modelius, akcentuojame juose naudojamų kintamųjų matavimo skales. Visų pirma, kas apskritai yra tas kintamasis statistiniame tyrime? Faktiškai – tai klausimyno klausimas. Jeigu jau kažko klausiame, tai matyt nesitikime, kad visi atsakys vienodai.

Pagal matavimo skales visi kintamieji skirstytini į

a) nominalius (kategorinius);
b) ranginius,
c) intervalinius[7].

Jeigu atsakant į klausimą, galima rinktis tik iš kelių lygiaverčių galimybių, tai atitinkamas kintamasis yra *nominalusis* (*kategorinis*). Pavyzdžiui, tokie kintamieji yra rasė, tikyba, plaukų spalva. Jeigu kategorijos yra tik dvi (vyras – moteris, sutinku – nesutinku), sakysime, kad kintamasis yra *dvireikšmis* arba *dichotominis*. Jeigu atsakymai yra skaitiniai ir, galima pasakyti kiek kartų viena reikšmė didesnė už kitą bei suskaičiuoti reikšmių vidurkį, tai tiriame *intervalinį* kintamąjį. Pavyzdžiui, ūgis, svoris, reakcijos laikas, kraujospūdis yra intervaliniai kintamieji. Dažniausiai socialinių tyrimų klausimynuose naudojami Likerto skalės kintamieji, kai liepiama pasirinkti iš 1 (labai nepritariu),..., 5 (labai pritariu), ar pan. Tokie kintamieji vadinami *ranginiais*. Negalima pasakyti, kiek daug skiriasi rangai, bet galima konstatuoti, kuris didesnis. Pavyzdžiui, kiek kartų atsakymas „labai pritariu“ didesnis už „labai nepritariu“ nepasakysime (neteigsime gi, kad „labai pritariu“ rodo 5 kartus didesnį pritarimą vien todėl, kad toks atsakymo kodas). Vis dėlto nesuklysime, konstatavę, kad pasirinktas atsakymo kodas 4 rodo didesnį pritarimą nei kodas 2.

Praktinėje statistikoje atsiranda papildomų niuansų.

Jeigu kintamasis yra ranginis, bet turi mažai skirtingų atsakymo variantų (mažiau, kaip 5), tai dažniausiai belieka į jį žiūrėti, kaip į nominalų.

[7] Griežtai kalbant, yra atskiros intervalų ir santykių skalės, bet abiejų skalių kintamiesiems taikomi tie patys statistiniai metodai. Todėl šiuose konspektuose ir vienu ir kitu atveju sakysime, kad tiriame intervalinį kintamąjį.

Pavyzdžiui, kintamasis *savijauta*, įgyjantis tik tris reikšmes (1 – nors nusišauk, 2 – nebloga, 3 – skraidau) visuose statistiniuose modeliuose traktuotinas, kaip nominalus kintamasis.

Jeigu ranginio kintamojo reikšmių yra pakankamai daug, o jo vidurkis visai prasmingas, tai dažniausiai jis traktuojamas, kaip intervalinis kintamasis.

Pavyzdžiui, egzamino pažymys iš tikro yra *ranginis* kintamasis, nes negalima pasakyti, kad studentas gavęs dešimtuką žino *tiksliai* dukart daugiau už gavusį penketuką ir *tiksliai* trisdešimčia procentų daugiau už gavusį septintuką. Vis dėlto, visuose tyrimuose egzamino pažymys traktuojamas kaip intervalinis kintamasis. Iš didelės bėdos kartais, kaip intervalinius, tenka naudoti ir penkiarangius kintamuosius.

1.1.2. Svarbiausios kintamųjų skaitinės charakteristikos

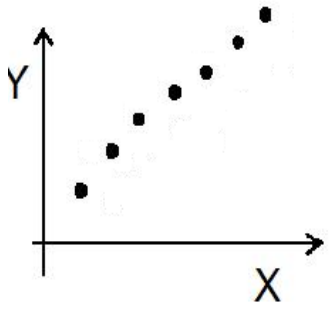
Statistiniuose tyrimuose, naudojamos specialios, tiriamus kintamuosius apibūdinančios, skaitinės charakteristikos. Aptarsime penkias tokias charakteristikas: vidurkį, dispersiją (ir jos ekvivalentą - standartinį nuokrypį), koreliaciją ir medianą.

Vidurkis yra vidutinė kintamojo reikšmė. Toks maždaug komunistinis pasidalijimo principas – kas būtų, jeigu visi savo turimus matavimus (pinigus, pragyventus metus, egzamino pažymius) sumestų į bendrą katilą ir draugiškai pasidalintų po lygiai. Formaliai – visų stebėjimų suma padalinta iš respondentų skaičiaus. Vidurkis – pati populiariausia skaitinė charakteristika statistikoje. Vis dėlto, kai duomenyse yra labai išsiskiriančių reikšmių, jis nėra itin informatyvus. Jeigu B.Geitsas savo gyvenama vieta deklaruos Adutiškį, kiekvieno adutiškiečio vidutinis turtas viršys milijoną dolerių, nors realiai jie nepraturtės.

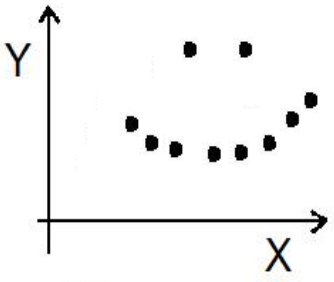
*Dispersij*a yra statistikoje naudojama skaitinė kintamojo reikšmių išsibarstymo charakteristika. Didelė dispersija reiškia, kad reikšmės labai skiriasi. Dispersija aktuali, kai norima pagal imtį spręsti apie visą populiaciją. Natūralu, kad patikimoms išvadoms apie populiacijas, kurių respondentai labai skiriasi, reikia didesnės imties. Jeigu išmatavus dešimties vyrų ūgius, radome ir 1,5 m, ir 2,5 m, ir 1,79 m ir dar kitokių pačių įvairiausių ūgių, tai tikrai padaryta išvada apie vidutinį to regiono vyrų ūgį bus mažiau patikima, nei tuo atveju, kai visi respondentai yra 1,60 m. Todėl, darant statistines išvadas, į dispersiją atsižvelgiama. Dažnai vietoje dispersijos naudojamas *standartinis nuokrypis*. Standartinis nuokrypis – tai kvadratinė šaknis ištraukta iš dispersijos.

Informatyvumo prasme abu rodikliai lygiaverčiai. Tradiciškai aprašomojoje statistikoje pateikiami standartiniai nuokrypiai, o statistinėse išvadose labiau taikoma dispersija.

Koreliacija (koreliacijos koeficientas) yra kintamųjų tiesinės priklausomybės matas. Kaip ir vidurkis ir dispersija, jis turi savo teorinį atitikmenį tikimybių teorijoje. Statistikoje koreliacija naudojama, kai reikia išmatuoti dviejų *intervalinių* kintamųjų priklausomybę. Koreliacijos koeficientas nematuoja netiesinės priklausomybės.



Tiesinė priklausomybė



Netiesinė priklausomybė

Koreliacijos koeficientas gali įgyti reikšmes nuo -1 iki 1. Kuo koreliacijos koeficientas absoliutiniu didumu didesnis (toliau nuo nulio), tuo priklausomybė stipresnė. Pažymėkime imties duomenims suskaičiuotą koreliaciją raide r. Sprendžiant apie koreliacijos stiprumą „iš akies“, dažniausiai laikomasi tokio vertinimo:

- $|r| < 0.3$ koreliacija labai silpna,
- $0.3 \leq |r| < 0.5$ silpna koreliacija,
- $0.5 \leq |r| < 0.7$ vidutinė koreliacija,
- $0.7 \leq |r| < 0.9$ stipri koreliacija,
- $0.9 \leq |r| \leq 1$ labai stipri koreliacija.

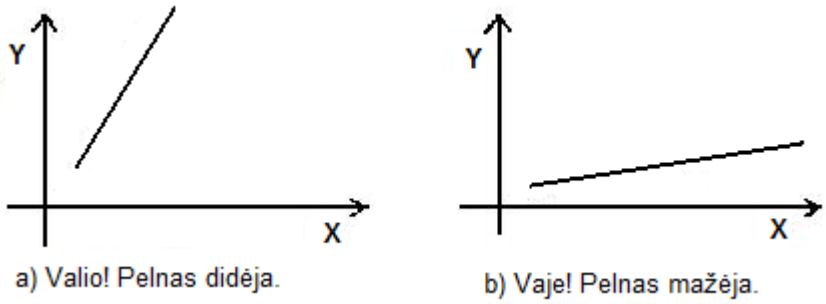
Koreliacijos koeficiento ženklas irgi informatyvus. Jeigu koreliacija tarp *X* ir *Y* teigiama, tai didesnes *X* reikšmes atitinka didesnės *Y* reikšmės. Jeigu neigiama – didesnes *X* reikšmes atitinka mažesnės *Y* reikšmės. Pavyzdžiui, jeigu augant darbo stažui, atlyginimas irgi augs, tai darbo stažo ir atlyginimo koreliacija bus teigiama. Koreliacija galima tarp visiškai skirtingos prigimties kintamųjų: tarp laiko ir pažymių, tarp amžiaus ir atlyginimo, tarp IQ ir TV žiūrėjimo ir pan.

Koreliacija nepaaiškina priežastingumo.

Koreliacija atsako į klausimą – ar tarp kintamųjų yra tiesinis ryšys, bet nepaaiškina, *kodėl* tas ryšys atsirado. Priežastingumui paaiškinti reikia remtis giliomis tiriamo dalyko žiniomis ir

vaizduote. Stipri koreliacija tarp chemijos ir fizikos pažymių rodo, kad mokiniai, geriau mokantys chemiją, geriau ir fiziką išmoksta. Bet tai dar nereiškia, kad fiziką išmoksta *todėl*, kad geriau moka chemiją.

Tarkime, kad įvertinome koreliaciją tarp išlaidų duonkepių reklamai ir gautų pajamų. Gavome r = 0,999. Koreliacija teigiama ir labai stipri. Daugiau išleidžiame pinigų reklamai – didesnės ir pajamos. Atrodytų, kad reklama labai naudinga duonkepių pardavėjams. Vis dėlto tokia išvada gali būti net labai klaidinga! Iš tikro, didėjant išlaidoms reklamai, didės ir gaunamos *pajamos*. Bet tai nereiškia, kad augs ir *pelnas* = *pajamos* – *išlaidos*. Pažvelkime į du galimus priklausomybės grafikus.



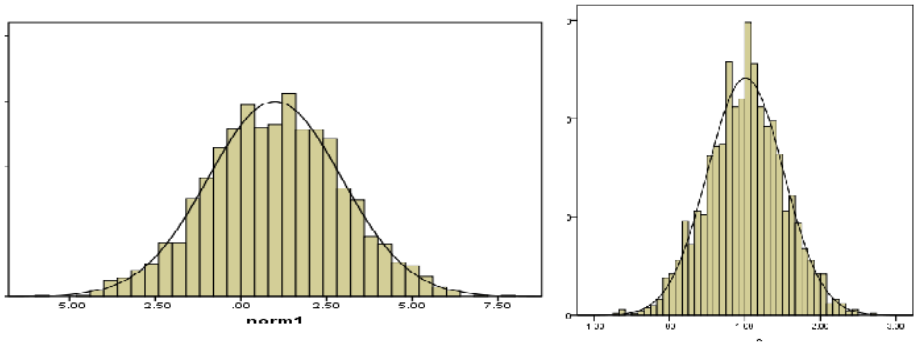
a) Valio! Pelnas didėja. b) Vaje! Pelnas mažėja.

Abiejuose grafikuose išlaidų reklamai (x-ašis) ir gautų pajamų (y-ašis) priklausomybę pavaizdavome ištisine tiese, nes koreliacija beveik 1. Matome, kad a) atveju kiekvienas reklamai papildomai išleistas litas duos keliolika litų papildomų pajamų ir pelnas augs. Tuo tarpu iš b) grafiko matyti, kad kiekvienas reklamai papildomai išleistas litas duos tik kelis centus papildomų pajamų, ir pelnas mažės (išleidžiame daugiau, nei uždirbame). Dar kartą norime pabrėžti, kad abiem atvejais koreliacija yra tokia pati – stipri ir teigiama. Šis pavyzdys rodo, kad tyrimuose vien koreliacijos gali ir nepakakti. Aptartuoju atveju visus reikiamus atsakymus (įskaitant ir tai, *kiek* padidės pajamos, išaugus išlaidoms reklamai vienu litu) padėtų gauti tiesinės regresijos modelis.

Mediana – tai vidurinė pagal didumą iš visų galimų reikšmių. Jeigu visas reikšmes išrikiuosime nuo mažiausios iki didžiausios, tai mediana bus tos eilutės viduryje esanti reikšmė (lyginės imties atveju – dviejų vidurinių reikšmių vidurkis). Mediana nėra nei tokia informatyvi, nei turi tokias geras matematines savybes, kaip vidurkis. Statistiniuose tyrimuose ji naudojama tada, kai duomenyse gali būti išskirčių. Mediana nepasikeis, jeigu vienas stebėjimas staiga taps milžinišku. Jeigu turtingiausią Adutiškio gyventoją pakeisime B.Geitsu, tai adutiškiečių turto mediana nepasikeis. Jeigu jau galima kalbėti apie vidurinę pagal didumą reikšmę, tai galima ir apie dešimtąją pagal didumą ir pan. Atitinkamos charakteristikos vadinamos *kvantiliais*.

1.1.3. Normalieji kintamieji

Visoje statistikoje dažnai sutinkamas terminas – normalusis kintamasis. Normalus kintamasis – tai intervalinis, pasiskirstęs pagal Gauso dėsnį, kintamasis. Jo įgyjamos reikšmės labiau sukoncentruotos apie vidurkį. Tikslus normalaus kintamojo elgesys aprašomas pakankamai bjauria formule, kurios čia nepateiksime[8]. Dažniausiai sakoma, kad tokio kintamojo histograma (grupuotų duomenų dažnių grafikas) primena varpą. Jeigu normalaus kintamojo dispersija didelė, tai tas varpas priplotas iš viršaus, jeigu maža – suplotas iš šonų. Žemiau pateikiame dvi kompiuteriu sugeneruotas normaliųjų kintamųjų histogramas: a) atveju vidurkis lygus 1, o dispersija yra 2; b) atveju vidurkis lygus 1, o dispersija yra 0,5. Ištisinės linijos rodo, kokie grafikai turėtų būti teoriškai.



a) b)

Praktiškai visos klasikinės statistinės išvados yra gautos, esant prielaidai, kad nagrinėjami normalieji kintamieji. O realiuose tyrimuose kintamieji retai kada būna normalūs (geriausiu atveju – beveik normalūs). Faktiškai statistikos taikymai primena situaciją su suknele, pasiūta manekenei. Manekenei (ideali figūra, 190 cm) tinka, kaip nulieta. O kai suknelę bandome užtempti ant eilinės pirkėjos (kartais 30 cm žemesnės, kartais 30 kg sunkesnės) rezultatai būna įvairūs. Kartais labai geri, kartais pakenčiami, kartais nepakenčiami. Taip ir su statistikos taikymais. Jeigu duomenys beveik normalūs – viskas beveik puiku. Jeigu visiškai nenormalūs – visiška tragedija. Jeigu normalumas modeliams labai svarbus, tai bent jau „iš akies“ jį reikėtų įvertinti.

Ranka rankon teorija ir praktika

[8] O kam ją pateikti? Mes tos formulės nenaudosime. Jei jau taip labai norisi ją pamatyti, tai tiks bet koks statistikos vadovėlis.

1.2. Statistinės išvados

Visų statistinių išvadų tikslas – ištyrus dalį respondentų (imtį), padaryti išvadą apie visą populiaciją. Akivaizdu, kad darant tokius apibendrinimus, galima ir suklysti. Statistikos taikymas neapsaugo nuo klaidų, bet užtikrina, kad jos bus daromos nedažnai.

1.2.1. Parametrų įverčiai ir jų pasikliautinieji intervalai

Kaip taisyklė, informacijos apie modelio parametrus (visokius jame esančius daugiklius, naudojamų kintamųjų vidurkius, dispersijas ir pan.) neturime. Norime juos įvertinti, pasinaudoję imties duomenimis. Taip gautos apytikslės nežinomų parametrų reikšmės vadinamos parametrų įverčiais. Yra įvairių parametrų įverčių radimo metodų. Visi jie sudėtingi ir skambiai vadinami – didžiausio tikėtinumo (angl. maximum likelihood), mažiausiųjų kvadratų (angl. ordinal least square) ir pan. Parametrų įverčių radimą patikėsime statistinėms programoms.

Kartais norima tikslesnės informacijos apie parametrą – jo galimų reikšmių intervalo. Informatyviau išgirsti, kad „šešėlinių“ pajamų vidurkis yra tarp 300 ir 500 litų per mėnesį, o ne vien tai, kad jis apytiksliai 400 litų. Tokią patikslintą informaciją suteikia parametro pasikliautinis intervalas. Kadangi, stebint imtį, suteikti šimtaprocentinės garantijos apie parametro reikšmę neišeina (imties atsitiktinumas gali pakišti koją), tai praktikoje naudojami 95% pasikliautinieji intervalai. Tie 95% reiškia, kad intervalo sudarymo taisyklė yra gera – 95 % jos taikymų baigiasi tuo, kad tikrieji parametrai į sudarytus intervalus pakliūva. Ar konkrečiai imčiai sudarytame intervale bus spėjamas parametras nežinome, bet intervalų konstravimo metodas apskritai yra neblogas. Čia panašiai, kaip skiriant vaistą, pacientui pasakyti: 95% ligonių jis padėjo. Ar konkrečiam ligoniui padės, neaišku, bet apskritai vaistas geras.

1.2.2. Statistinė hipotezė

Kartais užtenka nedidelės imties, kad galėtume padaryti išvadą apie visą populiaciją. Pavyzdžiui, virėjas paragauja sriubos ir nusprendžia – trūksta druskos. Visi šia išvada noriai patiki ir niekas nereikalauja išsrėbti viso katilo. Kartais net nemaža imtis neįtikina. TV žvaigždė paskelbia – visi vyrai tokie! (Ir paaiškina kokie…) Nesinori patikėti, kad taip jau ir visi. Gal vertėtų palaukti, kol tas vyrų tyrinėjimas bus pratęstas. Atkreipsime dėmesį, kad darant bet kokią išvadą apie populiaciją pagal jos dalį (imtį), visada lieka galimybė apsirikti (sriuboje gali būti nugrimzdęs dar neištirpęs druskos gabalas, kitas vyras bus ne toks). Todėl teks susitaikyti su neišvengiama tiesa – jokia

statistinė išvada nėra šimtaprocentinė. Belieka pasistengti, kad klaidingos išvados tikimybė būtų nedidelė. Labai grubiai visas statistines hipotezes galim įsivaizduoti, kaip du teiginius:

H_0: parametras (parametrų skirtumas) yra lygus nuliui,
H_1: parametras (parametrų skirtumas) nelygus nuliui.

Dažniausiai tyrėjo tikslas yra patvirtinti H_1, todėl vadinamuoju **reikšmingumo lygmeniu** iš anksto nusistatoma, koks turėtų būti tokio sprendimo patikimumas. Tiksliau kalbant, reikšmingumo lygmuo – tai teorinis, iš anksto nusistatytas leistinų *apsirikimų* procentas, priimant H_1. Visuotinai naudojamas 0,05 reikšmingumo lygmuo. Jeigu jau skelbsime, kad kažkokie parametrai skiriasi, tai klaidų bus ne daugiau 5% (garantuosime sprendimo teisingumą bent 95 %).

Reikšmingumo lygmuo – tai mūsų iš anksto pasirenkamas leistinų neteisingų sprendimų procentas. Praktiniuose skaičiavimuose atsižvelgiama į ***p reikšmes***. *P* reikšmė – tai tikimybė, klaidingai nuspręsti, kad parametrai skiriasi (atmesti nulinę hipotezę) konkrečios imties duomenims. Skamba panašiai, kaip ir reikšmingumo lygmuo. Esminis skirtumas tas, kad reikšmingumo lygmuo – tai teorinis iš anksto nusistatytas nekintantis etalonas, o *p* reikšmė – tai, ką pavyko konkrečiu atveju gauti. Ji kiekvienai imčiai vis kita. Sprendimas priimamas, *p* reikšmę lyginant su etalonu – reikšmingumo lygmeniu. Jeigu *p* reikšmė mažesnė už 0,05, tai nulinė hipotezė atmetama. Tada daroma išvada apie kažkokius statistiškai reikšmingus skirtumus (kokius priklauso nuo tikrinamos hipotezės).

Jeigu $p < 0{,}05$, tai H_1 priimame.

Jeigu $p \geq 0{,}05$, tai H_1 atmetame.

P reikšmės mažėja, didėjant imties didumui. Tai visai logiška – kuo daugiau duomenų, tuo patikimesnės išvados. Kita vertus, tai reiškia, kad nedera pasitikėti vien *p* reikšmėmis. Labai didelėms imtims galime gauti visiškai tyrimo prasme neįdomius statistiškai reikšmingus skirtumus. Pavyzdžiui, nustatyti, kad statistiškai reikšmingai varanopsai miega visa minute ilgiau nei varanopsės (atitinkamai septynias valandas trisdešimt tris minutes ir septynias valandas trisdešimt dvi minutes). Ir, matyt, būtent tą minutę sapnuoja vyriškai šovinistinius sapnus.

Pati populiariausia šio konspekto frazė susijusi su statistinėmis hipotezėmis – parametras statistiškai reikšmingas. Išvertus į šnekamąją kalbą, tai reiškia, kad labai tikėtina (95% garantija), jog jis nelygus nuliui. Norisi pabrėžti, kad frazė – parametras nėra statistiškai reikšmingas – dar *nereiškia* spėjimo, jog jis nulis. Tiesiog negalime duoti 95 % garantijos, kad nelygus nuliui (nors galbūt galime duoti 90% garantiją). Mes reiklūs sau, ir mažesnė nei 95 % garanatija, mūsų netenkina.

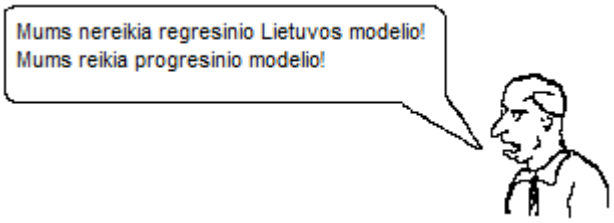
Parametras statistiškai reikšmingas = yra bent 95% garantija, kad jis nelygus 0.

Parametras statistiškai nereikšmingas = garantija, kad nelygus 0, yra mažesnė už 95% .

Visos imties duomenų funkcijos vadinamos statistikomis (kaip ir pati disciplina, toks jau nemalonus sutapimas). Jos turi pačius įvairiausius pavadinimus. Pavyzdžiui, chi kvadrato statistika, Voldo statistika, Fišerio statistika. Neretai modelio gerą tikimą duomenims parodo tai, kad atitinkama statistika yra statistiškai reikšminga (labai nepanaši į nulį). Tuo atveju, atitinkama *p* reikšmė yra mažesnė už pasirinktąjį reikšmingumo lygmenį 0,05. Šis faktas būtinai paminimas, aprašant išvadas.

1.2.3. Pastaba apie naudojamus terminus

Šiek tiek pastabų apie konspektuose naudojamus statistikos terminus. Tokie terminai, kaip *regresija, p reikšmė, pasiklautinis intervalas, heteroskedastiškumas* yra tradiciniai ir plačiai naudojami. Kiti – *kaupiamoji žingsninė regresija, atsparioji regresija, betos pokyčio statistika DFB* – kur kas menkiau paplitę. Nesiimame teigti, kad jie labiausiai vykę ir vieninteliai teisingi. Tikimės, kad, esant reikalui, skaitytojas sugebės juos aprašymuose pakeisti į tuos, kurie nusistovės lietuviškoje terminijoje. Beje, terminas *regresiniai* modeliai yra istorinis ir jokios neigiamos prasmės neturi.

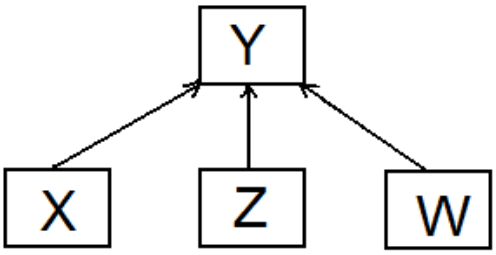


2. TIESINĖ REGRESINĖ ANALIZĖ

Tipinės situacijos, kai taikoma regresinė analizė yra šios:

- Tiriame nusikaltimų skaičiaus priklausomybę nuo gyventojų tankumo ir regiono urbanizuotumo laipsnio.
- Aiškinamės, ar galima prognozuoti užsienio prekybos mastus, atsižvelgiant į BVP, bankinių palūkanų normą ir vidutinį darbo užmokestį.
- Nagrinėjame, ar per egzaminą surinktų balų skaičius priklauso nuo socialinio statuso.

Visais atvejais manome, kad intervalinis kintamasis priklauso nuo vieno ar kelių kitų kintamųjų. Modelio schema:



Kintamasis *Y* vadinamas *priklausomu* (arba regresuojamu) kintamuoju, kintamieji *X, Z, W* – *aiškinamaisiais* kintamaisiais arba *regresoriais*. Visos spėjamos priklausomybės yra tiesinės.

2.1. Tiesinės regresijos modelis

2.1.1. Modelio lygtis

Matematinis formalus regresijos modelio užrašas yra toks:

$$Y = C+b_1 X+b_2Z+b_3W+e.$$

Čia *e* žymi liekamąją paklaidą, t.y. viską nuo ko dar gali priklausyti *Y*. Nei konstantos C, b_1, b_2, b_3, nei liekamoji paklaida nėra žinomos. Konstantų įverčiai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$ gaunami, panaudojus imties duomenis. Gaunama lygtis apytikslei *Y* reikšmei:

$$\hat{Y} = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W \ .$$

Ši lygtis naudojama kokybinei ir kiekybinei kintamųjų priklausomybių analizei. Koeficientų ženklai nurodo, ar regresoriams didėjant, *Y* didės, ar mažės.

Jeigu $\hat{b}_1 > 0$, tai *X* didėjant, *Y* didėja.

Jeigu $\hat{b}_1 < 0$, tai *X* didėjant, *Y* mažėja.

Ką tas didėjimas/mažėjimas reiškia, priklauso nuo kintamųjų prasmės ir kodavimo. Beje, jeigu koeficientų $\hat{b}_i$ ženklai yra ne tokie, kaip tikėjomės sudarydami modelį, tai labai galimas daiktas, kad susidūrėme su kintamųjų multikolinearumu. Ši problema aptarta kituose skyreliuose.

Koeficientas $\hat{b}_1$ parodo, kiek pasikeis $\hat{Y}$ reikšmė, vienu vienetu padidėjus *X* ir fiksavus visų kitų kintamųjų reikšmes.

Pavyzdys. Tarkime, kad gavome tokį regresijos modelį

$$Egzamino\ balas = 70 + 1{,}1\ h\_su\_repetitoriumi - 2\ h\_socialiniame\_tinklalapyje.$$

Daugiklis 1,1 prie kintamojo *h_su_repetitoriumi* rodo, kad kiekviena papildomai su repetitoriumi praleista valanda padidina egzamino rezultatą vidutiniškai 1,1 balo. Analogiškai, kiekviena valanda, sugaišta skaitinėjant/rašinėjant socialinį tinklalapį, sumažina egzamino rezultatus vidutiniškai 2 balais.

Be modelio koeficientų dar naudojami *standartizuotieji beta koeficientai.* Jie atsiranda, regresijos modelį taikant regresorių standartizuotosioms z reikšmėms. Faktiškai tai reiškia, kad suvienodiname regresorių matavimo skales ir nebesvarbu, kad vieną kintamąjį matuojame tonomis, o kitą centimetrais. Šie koeficientai reikalingi, norint palyginti regresorių santykinę įtaką priklausomam kintamajam.

Kuo absoliutiniu didumu standartizuotasis beta koeficientas didesnis, tuo atitinkamas kintamasis (regresorius) modelyje svarbesnis.

Pavyzdžiui, jeigu kintamojo *X* standartizuotasis beta koeficientas lygus -0,8, tai jis modelyje daug svarbesnis, nei kintamasis *Z*, kurio standartizuotasis beta koeficientas lygus 0,3. Beta koeficientas neaptariamas, jeigu modelyje yra tik vienas regresorius.

2.1.2. Prognozavimas

Prognozuojama tiesiog, įstatant konkrečias regresorių reikšmes į gautąją regresijos lygtį

$$\hat{Y} = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W.$$

Pavyzdys. Tarkime, kad atlikę regresnę analizę, gavome tokį modelį

$$\hat{Y} = 7 - 1.3X + 2Z + 3W.$$

Kokią *Y* reikšmę prognozuosime, kai *X=10, Z=0, W=1*? Prognozuosime $Y = 7 - 13 + 3 = -3$.

Vis dėlto, reikia nepamiršti, kad prognozės daromos tik toms *X, Z, W* reikšmėms, kurios nedaug skiriasi nuo turimų duomenų. Kvaila daryti prognozes apie pensininkų išgeriamo pieno kiekį, jeigu tyrėme tik ikimokyklinio amžiaus vaikus. Kvaila, pasitikint vien regresijos modeliu, prognozuoti 2050 metų BVP, kai turime tik 1995—2007 metų BVP duomenis.

2.1.3. Reikalavimai duomenims

Visi duomenys skaitiniai. Net ir nominaliųjų kintamųjų reikšmės užkoduotos skaičiais. Be to:

a) Priklausomas kintamasis *Y* yra normaliai pasiskirstęs. Visi kiti kintamieji yra intervaliniai, išskyrus dalį dvireikšmių kintamųjų.

b) Klasikiniame modelyje tariama, kad regresoriai matuojami be paklaidų ir yra neatsitiktiniai. Pastarasis reikalavimas praktikoje dažniausiai nėra logiškas, todėl visuotinai priimta regresorius irgi laikyti atsitiktiniais dydžiais. Regresoriai tuo geriau tinka modeliui, kuo jie panašesni į normaliuosius atsitiktinius dydžius. Faktiškai, tai reikalavimas, kad visų kintamųjų histogramos būtų „varpo formos“. Neretai, siekiant didesnio kintamųjų panašumo į normaliuosius, kintamieji transformuojami. Pavyzdžiui, ekonominiuose regresijos modeliuose itin dažnai taikoma logaritminė transformacija, t.y. surandami kintamųjų logaritmai ir regresijos modelis sudaromas šiems logaritmams. Pradinis modelis *Darbo užmokestis* $= C + b_1 BVP + e$ gali būti keičiamas modeliu *log*(*Darbo užmokestis*) $= C + b_1 log(BVP) + e$. Regresorių normalumo nereikia suabsoliutinti. Socialiniuose moksluose masiškai taikomi regresiniai modeliai, kai joks kintamųjų normalumas netiriamas ir pasitenkinama tuo, kad jie yra intervaliniai.

c) Kartais į modelį įtrukiami ir kategoriniai kintamieji, vadinami *pseudokintamieji.* Visi jie turi būti perkoduoti taip, kad įgytų tik dvi reikšmes – 0 ir 1. Pavyzdžiui, galima į modelį įtraukti pseudokintamąjį lytis (1 – vyr., 0 – mot.). Bendrojoje regresijos modelio lygtyje įstatę vietoje lyties reikšmę 1, gausime regresijos modelio lygtį vyrams. Įstatę 0, gausime regresijos modelio lygtį moterims. Vis dėlto pseudokintamieji naudotini tik iš didelės bėdos, kai duomenų mažoka. Daug tikslesni regresijos modeliai bus gauti, kai atskirai tirsime vyrus ir moteris. Jeigu kategorinis kintamasis įgyja tris reikšmes (pvz., lietuvis, latvis, estas), tai įvedami du pseudokintami P_1 ir P_2 (lietuviams $P_1 = 0$, $P_2 = 0$, latviams $P_1 = 0$, $P_2 = 1$, o estams $P_1 = 1$, $P_2 = 0$). Kodavimo sistema tokia: vieną kategoriją atitinka visi nuliai, kiekvieną kitą kategoriją atitinka *vienas* vienetas ir nuliai. Kaip taisyklė, regresijos modeliai, kai visi regresoriai yra pseudokintamieji, nenagrinėjami.

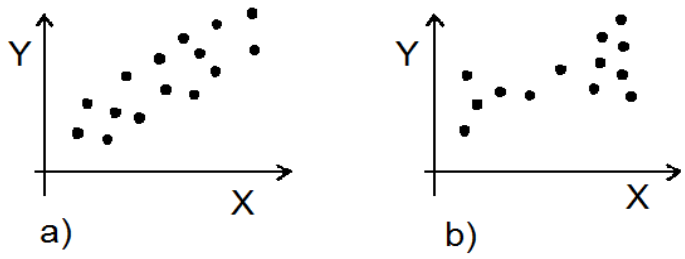
d) Skirtingų stebėjimų liekamosios paklaidos *e* neturi koreliuoti. Praktikoje tai reiškia, kad stebėjimai nesusiję. Dažniausiai tai labai natūralus reikalavimas – mes nemanome, kad Petro

IQ priklauso nuo Jono IQ, arba, kad pirmojo respondento ištikimybė Europos Sąjungai kažkaip paveikia kitų respondentų požiūrį.

e) Regresoriai neturi stipriai koreliuoti. Priešingu atveju iškyla vadinamoji multikolinearumo problema. Tada modelis tampa nestabiliu, t.y. keli papildomi stebėjimai gali radikaliai pakeisti vertinamų koeficientų $\widehat{C}, \widehat{b}_1, \widehat{b}_2, \widehat{b}_3$ reikšmes. Sudarius regresijos modelį 100 stebėjimų, natūralu tikėtis, kad vienas papildomas stebėjimas radikaliai modelio nepakeis. Priešingu atveju darosi problematiškas regresijos lygties koeficientų reikšmių komentavimas. Kita problema, kuri gali kilti dėl regresorių multikolinearumo, yra neteisingi jų ryšiai modelyje. Pavyzdžiui, gali būti taip, kad iš tikro koreliacija tarp *X* ir *Y* yra teigiama (t.y. *X* didėjant, didėja ir *Y*), o regresijos modelis neteisingai rodo, kad didėjant *X*, kintamasis *Y* mažėja.

f) Duomenyse neturi būti išskirčių. Išskirtis – tai tokia *Y, X, Z* arba *W* reikšmė, kuri stipriai skiriasi nuo kitų stebėjimų. Modelis, sudarytas duomenims su išskirtimis nėra patikimas.

g) Duomenys turi būti homoskedastiški. Reikalaujama, kad liekamosios paklaidos dispersija nepriklausytų nuo regresorių reikšmių. Jeigu taip nėra, tai sakome, kad iškilo *heteroskedastiškumo* problema. Praktiškai hetroskedastiškumas pasireiškia tuo, kad vienoms regresorių reikšmėms priklausomas kintamasis *Y* įgyja labai skirtingas reikšmes, o kitoms – ne. Nubraižykime *Y* priklausomybės nuo kiekvieno regresoriaus grafikus (angl. scatter plot) ir pažiūrėkime, ar gauti *Y* reikšmių išsibarstymo „debesėliai“ yra daugmaž vienodo storio visoms *X* reikšmėms. Jeigu taip, tai duomenys homoskedastiški (a) pav.). Jeigu ne, tai – heteroskedastiški (b) pav.). Modelis, sudarytas labai heteroskedastiškiems duomenims nėra patikimas.



2.1.4. Modelio tinkamumas

Regresinį modelį aprašo tokie rodikliai:

a) *Determinacijos koeficientas (R kvadratas).* Tai svarbiausia modelio tikimo duomenims charakteristika, kuri privaloma visuose regresijos modelių aprašymuose. Determinacijos koeficientas lygina skirtumus tarp *Y* reikšmių, kai atsižvelgiama į regresijos modelį, su

skirtumais tarp *Y* reikšmių, kai į modelį neatsižvelgiama. Labai apytikslė R^2 interpretacija, padedanti geriau suvokti jo prasmę, yra tokia – kiek procentų *Y* elgesio paaiškina kintamųjų *X, Z, W* elgesys. Determinacijos koeficientas įgyja reikšmes iš intervalo [0, 1]. Kuo koeficiento reikšmė didesnė, tuo modelis geriau tinka duomenims. Blogai, kai $R^2 < 0{,}20$. Apskritai nėra ko labai džiaugtis ir tada, kai $R^2 = 0{,}25$ (modelis tinka tik iš bėdos). O štai, jei $R^2 = 0{,}89$, tai modelis labai gerai aprašo duomenis. Tiesa, tai dar neužtikrina, kad visi kintamieji jame būtini, o pats modelis yra prasmingas.

b) *Koreguotas determinacijos koeficientas* (angl. Adjusted R square). Tai alternatyva determinacijos koeficientui, kai modelyje yra daug regresorių ir mažai stebėjimų. Jeigu respondentų yra bent 7 kartus daugiau, nei regresorių, tai nereikia jokių koreguotų determinacijos koefecientų, nes visiškai pakanka standartinio R^2. Pavyzdžiui, jeigu turime 15 respondentų duomenis, o modelyje jų IQ modeliuojame pagal socialinį statusą, tėvo IQ, motinos IQ, brolių ir seserų IQ, mokytojų IQ vidurkį, respondento amžių ir perskaitytų knygų per metus skaičių, tai tikrai reikia koreguoto R^2. O jeigu apsiribosime tik tėvų IQ, tai užteks įprasto determinacijos koeficiento. Rezultatų aprašyme pateikiamas tik vienas (koreguotas arba paprastas) determinacijos koeficientas.

c) *ANOVA p-reikšmė*. Ji parodo, ar modelyje yra su priklausomu kintamuoju susijusių regresorių. Jeigu *p* reikšmė didesnė už 0,05, tai regresijos modelio tinkamumas labai abejotinas (faktiškai gauname, kad *Y* nepriklauso nuo *X, Z* ir *W*). Jeigu *p* reikšmė mažesnė už 0,05, tai gavome patvirtinimą, jog modelis nėra beviltiškas (o gal ir visai geras – reikia tirti toliau). Dažniausiai apie ANOVA *p* reikšmę regresijos modelio aprašyme neužsimename. Jeigu jau regresijos modelis pristatomas, kaip duomenims tinkamas, tai visi ir taip supranta, kad ši reikšmė buvo maža.

d) *T (Stjudento) testai atskiriems regresoriams*. Padeda nuspręsti ar kintamasis šalintinas iš modelio. Jeigu atitinkamo testo *p* reikšmė $< 0{,}05$, tai sakome, kad *kintamasis yra statistiškai reikšmingas* ir dažniausiai (jei nėra multikolinearumo) jį modelyje paliekame. Jeigu *p* reikšmė $\geq 0{,}05$, tai kintamasis yra statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais (žr. modelio tobulinimą žemiau). Dažniausiai modelio konstanta *C* paliekama net ir tada, kai ji statistiškai nereikšminga.

Ar patenkintos regresijos modelio prielaidos parodo:

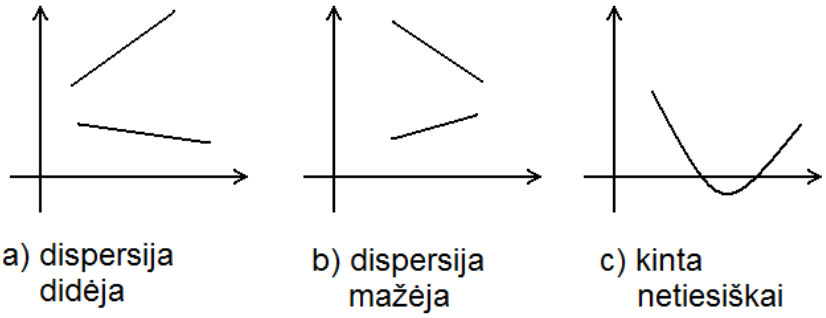
1) *Dispersijos mažėjimo daugiklis (VIF)*. Parodo ar regresoriai stipriai tarpusavyje koreliuoja (yra multikolinearumo problema). VIF skaičiuojamas kiekvienam regresoriui.

Multikolinearumas yra, kai VIF > 4. Vietoje VIF galima naudoti per jį išsireiškiantį rodiklį – toleranciją. Kintamojo *tolerancija = 1/VIF*. Blogai, kai *tolerancija < 0,25*.

2) *DFB (betos pokyčio,* angl. difference in beta) statistika. Parodo, ar duomenyse yra išskirčių. DFB skaičiuojama kiekvienam regresoriui ir konstantai atskirai. Jeigu imties didumas *n*, o regresorių yra *k*, tai DFB statistikų bus $n\ (k + 1)$. Blogai (konkreti regresoriaus reikšmė yra išskirtis), kai DFB > 1.
3) *Kuko matas.* Parodo, ar duomenyse yra išskirčių. Alternatyva DFB. Skaičiuojama kiekvienam regresorių rinkiniui. Jeigu imties didumas *n*, tai ir Kuko matų bus *n*. Blogai, kai Kuko matas viršija 1. Itin griežti (ir negausūs) statistikai mano, kad blogai, kai viršija 4/n. Skirtingai nei DBF, nenurodo išsiskiriančios regresoriaus reikšmės. Norodo tik kurio respondento duomenys išsiskiria.
4) *Durbino-Vatsono statistika*. Norime nustatyti, ar skirtingų stebėjimų liekamosios paklaidos koreliuoja (yra autokoreliacija). Faktiškai tai reiškia, kad vieno respondento stebėjimai daro įtaką kito respondento atsakymams. Durbino – Vatsono statistika tiriama *tik tada*, kai kyla toks įtarimas (pvz. respondentai atsakinėdami žiūrėjo kaimynams per petį; matavome biržos indeksus kelias dienas iš eilės ar pan.). Visais kitais atvejais šios statistikos nereikia. Praktiškai Durbino – Vatsono statistika reikalinga tik tada, kai duomenys sudaro laiko eilutę (duomenys yra laikiniai, t.y. stebime kintamųjų elgesį, bėgant laikui), o norime taikyti paprastą tiesinę regresiją, o ne sudėtingą laiko eilučių modelį. Socialiniuose tyrimuose tokios situacijos retos ir Durbino – Vatsono statistikos taikymas juose yra gryna egzotika. Jeigu Durbino – Vatsono statistika yra tarp 1,5 ir 2,5, tai dažniausiai tariama, kad auto-koreliacijos nėra.
5) *Standartizuotosios liekamosios paklaidos* (angl. standardized residuals) naudojamos patikrinimui, ar *Y* normaliai pasiskirstęs. Dažniausiai tiriama histograma, kuri lyginama su normaliąja kreive (turi nedaug skirtis) ir standartizuotojų liekamųjų paklaidų ir normaliojo atsitiktinio dydžio santykiniai procentinai dažniai (angl. P-P plot). Kuo taškai arčiau nubrėžtos tiesės (idealiu atveju visi taškai yra *ant* tiesės), tuo duomenys normalesni.
6) *Šapiro – Vilko (Shapiro – Wilk) testo p reikšmė* ≥ 0,05 rodo, kad standartizuotosios paklaidos yra normalios. Reikia nepamiršti, kad didelėms imtims šis kriterijus gali normalumo hipotezę nepagrįstai atmesti. Didelėms imtims grafikai (žr. ankstesnį punktą) informatyvesni.
7) *Liekamųjų paklaidų diagramos kiekvienam kintamajam* parodo, kurie kintamieji (regresoriai) modelyje reikalingesni. Geri tie kintamieji, kuriems diagramos rodo tiesinę priklausomybę. Nereikia stebėtis, kad x ir y ašys įgyja keistas reikšmes. Mums svarbi tik

grafiko *forma*. Grafike *neatidedamos* kintamųjų reikšmės. Iš tikro vienoje ašyje atidedamos liekamosios paklaidos, gautos iš regresijos modelio be tikrinamojo kintamojo, o kitoje ašyje – liekamosios paklaidos iš regresijos modelio, kai pats kintamasis modeliuojamas pagal likusius regresorius.

8) *Standartizuotųjų prognozuojamų reikšmių* (angl. standardized predicted values) grafikas, kai y-ašyje atidedamos standartizuotosios liekamosios paklaidos, naudojamos homoskedastiškumui tikrinti. Gerai, kai grafikas primena vienodo storio juostelę. Blogai, kai grafiko taškų išsibarstymas rodo, kažkokį kitą reguliarumą. Galimos blogos situacijos pateiktos a) – c) paveikslėliuose.



9) *Breušo – Pagano (Breusch – Pagan) testo p reikšmė* $\geq$ 0,05 rodo, kad homoskedastiškumo prielaida patenkinta. SPSS šio testo (kol kas) nėra. Reikia nepamiršti, kad didelėms imtims šis kriterijus gali homoskedastiškumo prielaidą nepagrįstai atmesti. Didelėms imtims grafikai (žr. ankstesnį punktą) informatyvesni.

Apibendrindami galime nusakyti, kokie turi būti gero regresijos modelio rodikliai.

Geram regresijos modeliui:

- $R^2 \geq 0{,}20$.
- ANOVA $p < 0{,}05$.
- Visų t testų $p < 0{,}05$.
- Visi VIF ≤ 4 (nėra multikolinearumo problemos).
- Visos Kuko mato (arba DFB) reikšmės $\leq$ 1.
- Koeficientų ženklai atitinka koreliacijas.
- Sprendžiant pagal histogramą ir P-P grafiką, liekamosios paklaidos normalios.
- Šapiro – Vilko testo $p \geq 0{,}05$ (liekanos normalios).
- Sprendžiant pagal liekamųjų paklaidų grafiką (ir Breušo – Pagano testo $p \geq 0{,}05$) nėra heteroskedastiškumo problemos

2.1.5. Tiesinės regresinės analizės etapai

Regresinė analizė susideda iš preliminaraus kintamųjų tinkamumo tyrimo ir modelio tinkamumo duomenims tikrinimo. Išsami analizė susideda iš tokių etapų:

1) Patikriname ar priklausomas kintamasis Y panašus į normalųjį kintamąjį. Idealiu atveju normalūs ir visi regresoriai.
2) Nustatome, ar priklausomas kintamasis Y koreliuoja su regresoriais. Jeigu regresoriai blogai koreliuoja su Y, tai jų įtraukimas į modelį yra labai abejotinas. Galima nusibraižyti Y priklausomybės nuo kiekvieno regresoriaus grafikus (angl. scatter plot). Jie turi rodyti tikėtiną tiesinę priklausomybę ir homoskedastiškumą.
3) Sudarome regresijos modelį ir pažiūrime, kokio didumo yra R^2 (arba, jeigu reikia, koreguotas R^2). Jeigu $R^2 < 0{,}20$ modelis nelabai tinkamas.
4) Patikriname ar ANOVA p reikšmė $< 0{,}05$. Jeigu ne – modelis netinkamas.
5) Patikriname ar visi regresoriai statistiškai reikšmingi (visos t kriterijaus p reikšmės $< 0{,}05$). Jeigu ne – modelis taisytinas. Konstantai p reikšmės nežiūrime.
6) Patikriname, ar nėra multikolinearumo, t.y. , ar visi VIF ≤ 4 ir ar modelio koeficientai b_j yra to paties ženklo, kaip ir atitinkamos 2) punkte rastos koreliacijos.
7) Patikriname ar nėra išskirčių.
8) Pagal standartizuotas liekamas paklaidas sprendžiame ar tenkinamos modelio normalumo ir homoskedastiškumo prielaidos ir, ar visi kintamieji modelyje reikalingi.
9) Nustatome modelio koeficientų įverčius ir (jeigu yra daugiau nei vienas regresorius) standartizuotuosius beta koeficientus.
10) Jeigu visi rodikliai geri – modelį aprašome. Jeigu ne – tobuliname.

2.1.6. Modelio tobulinimas

Beveik visada modelio tobulinimas reiškia kintamųjų šalinimą arba jų transformaciją.

> Pašalinus bent vieną regresorių, keičiasi visos modelio charakteristikos: R^2, p reikšmės, VIF, Kuko mato reikšmės.

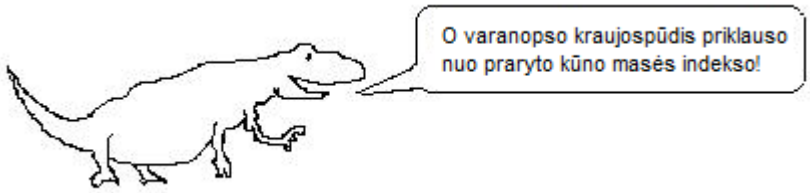
Modelis tobulinamas, jeigu yra problemų:

- $R^2 < 0{,}20$. Regresorių šalinimas šiuo atveju nelabai padeda. Juos prasmingiau keisti kitais regresoriais. Dalis statistikų mano, kad ir mažiems R^2 (0,15 – 0,19) dar galima modelio neatmesti. Ypač, jeigu modelio ryšiai puikiai atitinka ekonominę (ar kitokią) logiką.

- *Yra statistiškai nereikšmingas kintamasis* (t testo *p* reikšmė didesnė už 0,05). Kintamasis pašalinamas iš modelio ir regresinė analizė pakartojama be jo. Viskas gerai, jeigu R^2 daug nesumažėjo (ką reiškia daug – sprendžia pats tyrėjas). Pristatydami galutinį modelį, tada rašysime, kad modelyje liko tik statistiškai reikšmingi kintamieji. Jeigu R^2 stipriai sumažėjo (pvz., nuo 0,35 iki 0,20), tai kintamąjį į modelį grąžiname, o aprašyme konstatuojame, kad „kintamasis buvo statistiškai nereikšmingas. Vis dėlto, jis paliktas modelyje, nes be jo labai sumažėja determinacijos koeficientas“.
- *Yra multikolinearumo problema* (yra VIF > 4 arba modelis neteisingai atspindi *Y* ir regresoriaus priklausomybę). Tada, arba kintamąjį šaliname ir tikriname R^2 sumažėjimą, arba kelis, stipriai koreliuojančius regresorius, keičiame jų vidurkiu. Žinoma, keičiame tik tada, jeigu toks vidurkis prasmingas. Pavyzdžiui, jeigu modeliuosime mokinio valstybinio egzamino pažymį atsižvelgdami į keturių lygiaverčių ekspertų nuomones, tai multikolinearumo neišvengsime (nebus taip, kad tas pats mokinys vienam ekspertui atrodys Einšteino reinkarnacija, o kitam – alternatyviai protingas). Tada logiškiau sudaryti modelį, atsižvelgiant į visų keturių ekspertų nuomonių vidurkį, o ne į kažkurio vieno eksperto nuomonę.
- *Imtyje yra išskirčių* (yra didelių Kuko mato arba DFB reikšmių). Gali būti, kad duomenyse tiesiog pasitaikė duomenų įvedimo klaida, arba sugebėjome praleistos reikšmės kodą (pvz. atvejį, kai kintamojo vaikų skaičius reikšmė 99 tereiškia, kad respondentui klausimas netaikytinas nes jis vienuolis, o ne itin daugiavaikis tėvas) traktuoti, kaip eilinį stebėjimą. Šiaip jau išskirties šalinti vien todėl, kad ji netinka regresijos modeliui negalima. Rekomenduojama, radus išskirtį, pakartoti regresinę analizę be jos. Jeigu gavome labai panašų modelį, tai išskirtis nėra kenksminga, ir dėl jos nėra ko sukti galvos. Kitu atveju, išeitimi gali tapti atsparioji regresija (žr. atitinkamą skyrelį).
- *Duomenys heteroskedastiški*. Gero heteroskedastiškumo problemos sprendimo nėra. Neretai duomenys transformuojami, pavyzdžiui, visos kintamųjų reikšmės logaritmuojamos. Dar dažniau (ypač socialiniuose tyrimuose) duomenų heteroskedastiškumas paprasčiausiai ignoruojamas. Net netiriamas.
- *Yra autokoreliacija.* Jeigu buvo galima įtarti, kad stebėjimai susiję ir Durbino – Vatsono statistikos reikšmė artima nuliui arba keturiems, tai turime autokoreliuotus stebėjimus. Tada regresijos modelis netinka ir reikia taikyti sudėtingesnius laiko eilučių metodus, kurių čia neaptarsime.

2.1.7. Pastabos apie modelio taikymą

- Terminas *daugialypė tiesinė regresija* naudotinas tik tada, kai regresorių yra bent du. Gali būti ir vienas regresorius. Visada pakanka užrašo „*tiesinė regresija*“, nes prireikus skaitytojas pats suskaičiuos, kiek tų regresorių modelyje yra.
- T testai dar vadinami Stjudento testais. Tiesog toks trumpinys.
- Įprasta modelyje palikti visus statistiškai reikšmingus kintamuosius. Vis dėlto pasitaiko (žr. 2.2.4 skyrelį), kai net statistiškai reikšmingas kintamasis modelyje nereikalingas.
- Kai turime vieną priklausomą kintamąjį ir vieną regresorių, rašyti apie multikolinearumo nebuvimą (VIF) nereikia. Nes išskyrus priklausomą kintamąjį, neturi daugiau su kuo tas vienišas regresorius koreliuoti. Sakinys „buvo vienas regresorius ir multikolinearumo problemos nėra (VIF = 1)“, žinoma, yra teisingas, bet skaitosi analogiškai sakiniui „Visi tyrime dalyvavę vyrai nebuvo nėšti“.
- Kai turime vieną priklausomą kintamąjį ir vieną regresorių, rašyti apie standartizuotąjį beta koeficientą nereikia. Juk nebeturime su kuo to vienišo regresoriaus lyginti.
- Šiame konspekte išsamiai aptariame tik tiesioginę tiesinę regresiją, kai visus modelio tobulinimus atlieka pats tyrėjas. Tyrimuose dar taikoma žingsninė (angl. stepwise) regresija, kai modelyje automatiškai paliekami patys statistiškai reikšmingiausi iš pradinių regresorių. Kaip ją atlikti parodysime atitinkame skyrelyje. Vis dėlto, mes manome, kad regresijos modelius reikia sudaryti sąmoningai, o ne formaliai. Tyrėjas turi nusimanyti, ką jis tiria. Dažnai ekonominė logika (sociologijos, psichologijos, biologijos, medicinos ir pan. teorija) apsprendžia, kuriuos kintamuosius modelyje palikti. Tarkime, formaliai modeliuodami sistolinį kraujospūdį pagal diastolinį kraujospūdį ir kūno masės indeksą, galime gauti statistiškai patikimą modelį, iš kurio bus pašalintas kūno masės indeksas. Tokio modelio vertė labai abejotina, nes mes tik konstatuojame faktą – žmonėms, kurių diastolinis kraujospūdis padidėjęs, turi padidėjusį ir sistolinį kraujospūdį. Na ir kas iš to? Turbūt neimsime teigti, kad sistolinio kraujospūdžio padidėjimą lemia diastolinis kraujospūdis. Šiuo atveju, kur kas įdomesnis modelis, kai nustatoma sistolinio kraujospūdžio priklausomybė nuo kūno masės indekso.



- Reikia nepamiršti, kad darydami prognozes mes, prognozuojame vidutinę reikšmę. Prognozuodami, kad studentas, kurio IQ = 100, išlaikys statistiką labai gerai (gaus 9), mes iš tikrųjų darome prognozę, kad visų tų respondentų, kurių IQ = 100, statistikos pažymių vidurkis bus 9. Vis geriau, negu nieko. Nereikia tikėti stebuklinga regresijos prognozių galia ir neteks skaudžiai nusivilti.
- Kartais spėjamame modelyje *Y* tiesiškai priklauso nuo kintamųjų laipsnių arba jų sandaugų Tuo atveju sakoma, kad spėjame esant kintamųjų tarpusavio sąveiką (interakciją). Tada reikia sukurti naujus kintamuosius, kuriems jau taikoma įprastinė tiesinė regresinė analizė. Pavyzdžiui, norint patikrinti modelio $Y = C + b_1XZ + e$ tinkamumą, pakanka atlikti kintamųjų transformaciją $V = XZ$, o po to nagrinėti modelį $Y = C + b_1V + e$.
- Nereikia taikyti Durbino-Vatsono statistikos, jeigu nėra pagrindo įtarti stebėjimų priklausomumo. Pasakymas, kad apklausus po vieną respondentą iš 30 skirtingų rajonų, Durbino-Vatsono statistika rodo duomenų tikimą regresijai, akivaizdžiai demonstruoja tyrėjo liguistą nepasitikėjimą respondentais – "O gal jie, atsakinėdami į klausimus, bendravo". Matyt, telepatiškai.

2.1.8. Standartiniai tiesinės regresijos modelio aprašymai

Svarbiausia visų regresijos modelių aprašymo taisyklė yra tokia: visada pateikiamas determinacijos koeficientas, beveik visada – modelio lygtis, o apie kitas tinkamumo charakteristikas rašoma tik tada, jeigu buvo problemos. Maždaug taip: buvo tokios ir tokios problemos, o mes jas išsprendėme taip ir taip. Visada aprašomas galutinis modelis. Skaitytojui neįdomu, kiek daug sunkaus triūso įdėjo tyrėjas, tobulindamas savo pradinį modelį. Visus domina, kas iš to tobulinimo išėjo. Pateiksime keletą galimų aprašymo fragmentų.

- *Sudarėme regresinį modelį, atspindintį Y priklausomybę nuo X. Gautasis R^2=0,87. Modelis* $Y = 7{,}34 + 2{,}1X$.
- *Sudarėme regresinį modelį, atspindintį Y priklausomybę nuo X, W,Z. Modelyje buvo palikti tik statistiškai reikšmingi kintamieji W ir Z. Determinacijos koeficientas R^2 = 0,66. Galutinis regresijos modelis* $Y = -2 + 3W + 3Z$. *Daug didesnę įtaką Y turi kintamasis W (jo standartizuotoji beta koeficiento reikšmė 0,68), nei Z (jo standarizuotoji beta koeficiento reikšmė 0,39).*
- *Sudarėme regresinį modelį, atspindintį Y priklausomybę nuo X, W, Z. Iš modelio buvo pašalintas statistiškai nereikšmingas kintamasis X. Nors kintamasis Z ir buvo statistiškai*

nereikšmingas (p = 0,059), jį pašalinus iš modelio, labai sumažėja determinacijos koeficiento reikšmė. Todėl jį modelyje palikome....

- *Dėl kintamųjų W ir Z multikolinearumo (VIF > 4) į modelį įtraukėme jų vidurkį...*
- *Modelyje buvo išskirtis, kurią pašalinus, regresijos modelis praktiškai nepasikeitė.....*

2.2. Tiesinė regresinė analizė su SPSS

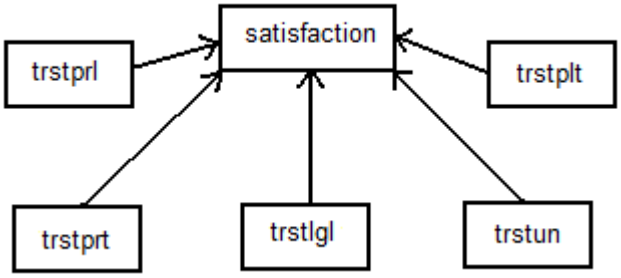
2.2.1. Duomenys

Kaip atlikti daugialypę tiesinę regresinę analizę, parodysime, tirdami 2008 metų Europos Sąjungos socialinio tyrimo (European Social Survey) Estijos duomenis ESS4EE. Tirsime 20 – 40 metų vyrus. Tyrime naudosime tokius kintamuosius:

- *stfgov* (kiek respondentas yra patenkintas savo šalies vyriausybe), *stfdem* (kiek respondentas yra patenkintas demokratijos būsena šalyje),
- *trstprl* (pasitikėjimas šalies parlamentu), *trstlgl* (pasitikėjimas teisine sistema), *trstplt* (pasitikėjimas politikais), *trstprt* (pasitikėjimas partijomis), *trstun* (pasitikėjimas Jungtinėmis Tautomis),
- *happy* (laimingumas).

Kintamieji *stfgov* ir *stfdem* matuojami 10 balų skale nuo 0 (itin nepatenkintas) iki 10 (itin patenkintas). Kintamieji *trstprl, trstlgl, trstplt, trstprt, trstun* matuojami 10 balų skale nuo 0 (itin nepasitikiu) iki 9 (itin pasitikiu). Kintamasis *happy* matuojamas 10 balų skale nuo 0 (itin nelaimingas) iki 9 (itin laimingas).

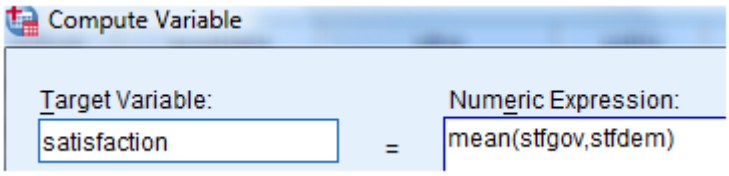
Modeliuosime kintamųjų *stfgov* ir *stfdem* vidurkį (pavadinkime jį *satisfaction*). Šis kintamasis atspindi respondentų pasitikėjimą politiniu šalies klimatu. Manome, kad *satisfaction* gali priklausyti nuo likusiųjų kintamųjų. Kitaip tariant, manome, kad pasitikėjimas įvairiomis šalies institucijomis veikia ir pasitikėjimą politiniu klimatu. Schematiškai spėjamas modelis atrodo taip:



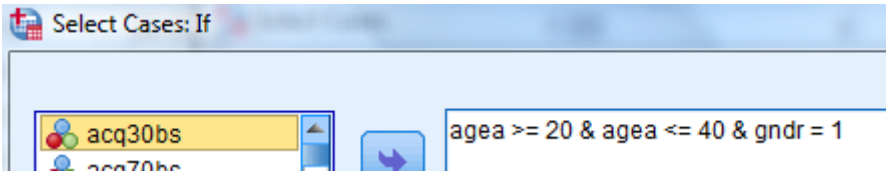
Norime nustatyti, ar duomenys neprieštarauja spėjamam modeliui, ar visi kintamieji modelyje reikalingi ir kurie iš jų svarbesni. Sąlyginai modelį galima užrašyti taip:

satisfaction = f(trstprl, trstlgl, trstplt, trstprt, trstun, happy).

Reikiamus respondentus atsirinksime, atsižvelgę į *agea* (amžių) ir *gndr* (lytį, 1-vyr., 2-mot.). Su komanda **Compute** sukuriame naują kintamąjį *satisfaction*, kuris yra kintamųjų *stfgov* ir *stfdem* vidurkis.

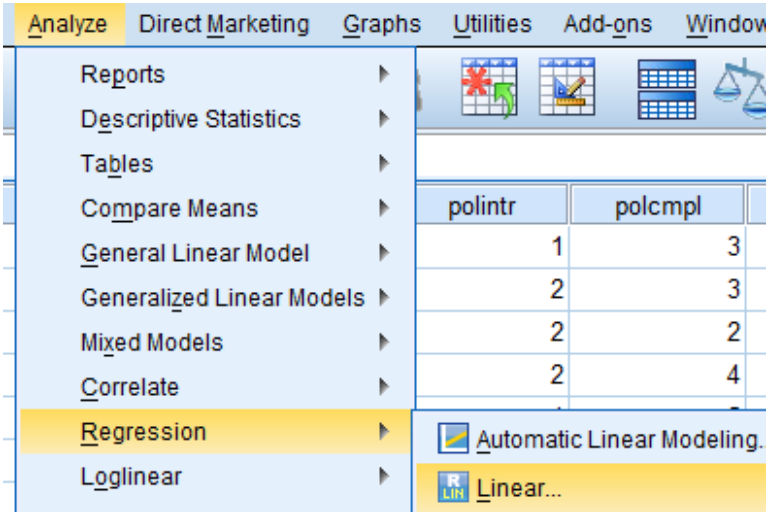


Su komanda **Select Cases** atrenkame reikiamus respondentus. Tai daroma taip: meniu **Select Cases** pažymime **If condition is satisfied** ir spaudžiame klavišą **If**. Naujajame meniu suformuluojame sąlygą *agea ≥ 20 & agea <= 40 & gndr = 1*. Tada spaudžiame **Continue** ir **OK.**

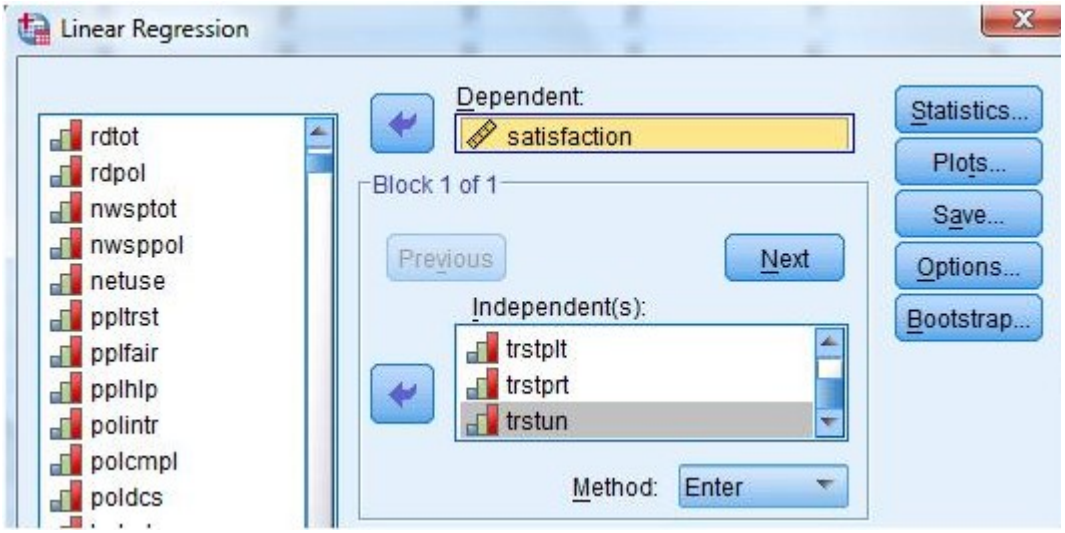


2.2.2. Pirminio modelio tyrimas

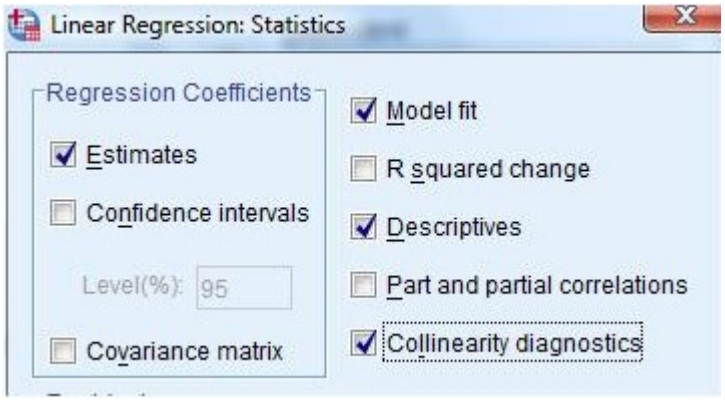
Atsidarome langelį **Analyze** ir renkamės opciją **Regression → Linear:**



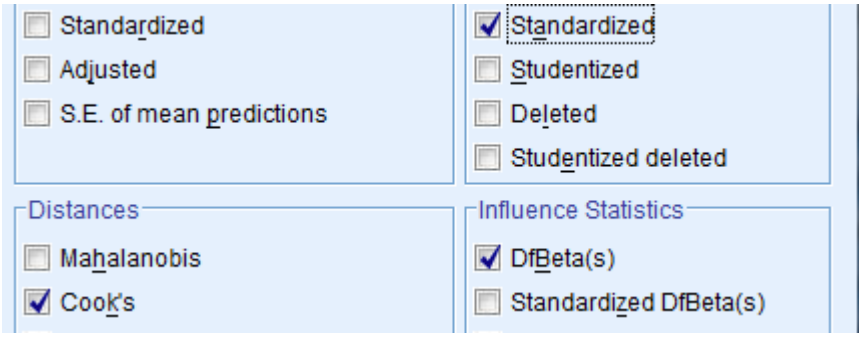
Į laukelį **Dependent** įkeliamas kintamasis *satisfaction*. Kintamieji *trstprl, trstlgl, trstplt, trstprt, trstun, happy* sukeliami į laukelį **Independent(s).** Po to spaudžiame **Statistics.**



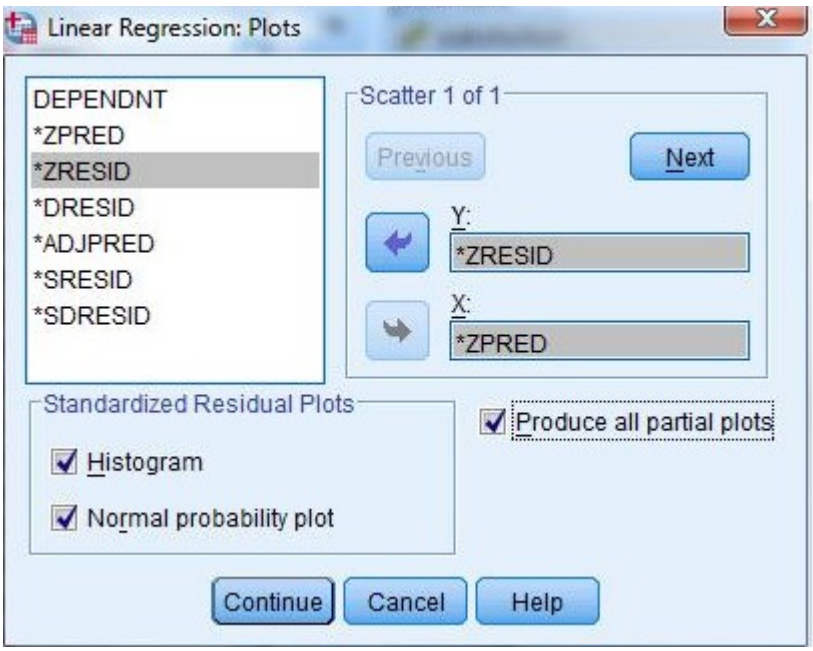
Atsidariusiame lange pažymime **Descriptives** ir **Colinearity diagnostics.** Visos kitas, pagal nutylėjimą pažymėtas opcijas (**Estimates, Model fit**), paliekame ramybėje.



Spaudžiame **Continue**, grįžtame į ankstenįjį langą, ir renkamės **Save.** Pažymime **Cook's** ir **DfBeta(s)** ir **Standardized.**



Spaudžiame **Continue**, grįžtame į ankstenįjį langą, ir renkamės **Plots**. Atsidariusiame lange pažymime laukus **Histogram**, **Normal probability plot** ir **Produce all partial plots**. Į laukelį **X** įkeliame ZPRED, o į laukelį **Y** -- ZRESID. Spaudžiame **Continue** ir **OK**.



2.2.3. Pirminio modelio tyrimo rezultatai

Rezultatų išklotinė prasideda nuo lentelėje **Descriptive Statistics** pateikiamos informacijos apie visų modelio kintamųjų vidurkius is standartinius nuokrypius. Matome, kad imtyje buvo 212 respondentų. Prisiminę kintamųjų kodus, galime konstatuoti, kad respondentai nelabai pasitiki visomis institucijomis, išskyrus Jungtines tautas.

Descriptive Statistics

	Mean	Std. Deviation	N
satisfaction	4.1439	2.22849	212
trstprl	3.93	2.322	212
trstlgl	5.19	2.542	212
trstplt	3.31	1.980	212
trstprt	3.36	1.926	212
trstun	5.44	2.237	212
happy	6.71	1.890	212

Tiesinės regresijos modelyje visi intervalinai regresoriai turi koreliuoti su priklausomu modeliujamuoju kintamuoju. Kita vertus, tarpusavyje regresoriai neturėtų stipriai koreliuoti. Koreliacijų lentelė **Correalations** rodo, kad visi regresoriai su kintamuoju *satisfaction* koreliuoja statistiškai reikšmingai, o koreliacijos vidutinio stiprumo. Be to, visos koreliacijos yra teigiamos. Koreliacijų ženklai svarbūs tuo, kad nurodo, kokios turi būti priklausomybės pačiame regresijos modelyje. Tiriamu atveju, visi regresijos modelio koeficientai turi būti teigiami. Vis dėlto, matome, kad regresoriai stipriai koreliuoja vieni su kitais. Todėl galima tikėtis, kad gali kilti multikolinearumo problema (regresoriai pradės vienas kitam trukdyti).

Correlations

		satisfaction	trstprl	trstlgl	trstplt	trstprt	trstun	happy
Pearson Correlation	satisfaction	1.000	.688	.623	.640	.624	.432	.421
	trstprl	.688	1.000	.648	.782	.746	.362	.288
	trstlgl	.623	.648	1.000	.544	.579	.402	.330
	trstplt	.640	.782	.544	1.000	.842	.446	.232
	trstprt	.624	.746	.579	.842	1.000	.453	.270
	trstun	.432	.362	.402	.446	.453	1.000	.189
	happy	.421	.288	.330	.232	.270	.189	1.000
Sig. (1-tailed)	satisfaction	.	.000	.000	.000	.000	.000	.000
	trstprl	.000	.	.000	.000	.000	.000	.000
	trstlgl	.000	.000	.	.000	.000	.000	.000
	trstplt	.000	.000	.000	.	.000	.000	.000
	trstprt	.000	.000	.000	.000	.	.000	.000
	trstun	.000	.000	.000	.000	.000	.	.003
	happy	.000	.000	.000	.000	.000	.003	.

Lentelėje **Model Summary** randame determinacijos koeficientą $R^2 = 0,596$. Primename, kad modelis nelabai tinkamas, jeigu $R^2 < 0,20$. Darome išvadą, kad determinacijos koeficientas pakankamai didelis ir tiesinės regresijos modelio atmesti, kaip netinkamo, dar negalima. Koreguotasis R^2 šiam modeliui neaktualus, nes respondentų yra keliasdešimt kartų daugiau, nei regresorių.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.772[a]	.596	.584	1.43749

a. Predictors: (Constant), happy, trstun, trstprl, trstlgl, trstprt, trstplt

b. Dependent Variable: satisfaction

ANOVA kriterijaus *p* reikšmė yra mažesnė už 0,05 (mažesnė net už 0,01). Todėl darome išvadą, kad modelyje yra bent vienas regresorius, nuo kurio priklauso *satisfaction*.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	624.254	6	104.042	50.350	.000[a]
	Residual	423.608	205	2.066		
	Total	1047.862	211			

a. Predictors: (Constant), happy, trstun, trstprl, trstlgl, trstprt, trstplt

b. Dependent Variable: satisfaction

Lentelėje **Coefficients** pateikiama pagrindinė informacija apie modelio koeficientus.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.856	.407		-2.103	.037		
	trstprl	.286	.077	.298	3.705	.000	.305	3.278
	trstlgl	.183	.054	.208	3.383	.001	.520	1.923
	trstplt	.190	.104	.169	1.832	.068	.232	4.319
	trstprt	.040	.101	.034	.390	.697	.257	3.891
	trstun	.113	.051	.113	2.199	.029	.747	1.338
	happy	.231	.056	.196	4.137	.000	.875	1.142

a. Dependent Variable: satisfaction

Patys koeficientų įverčiai yra stulpelyje **Unstandardized Coefficients B**. Visų pirma išsiaiškinsime, kaip užrašyti matematinį regresijos modelio pavidalą, kurį galima naudoti prognozėms. Reikia koeficientus padauginti iš šalia parašytų kintamųjų ir sudėti. Konstanta pridedama tokia, koks yra jos įvertis. Gauname tokį modelį:

$$satisfaction = -0{,}856 + 0{,}286trstprl + 0{,}183trstgl + 0{,}190trstplt + 0{,}040trstprt + 0{,}113trstun + 0{,}231happy.$$

Visų koeficientų ženklai teigiami, kaip ir turėjo būti, sprendžiant pagal koreliacijas. Norėdami suprasti, kuris kintamasis modelyje svarbesnis, pasižiūrime į **Standardized Coefficients Beta**. Matome, kad mažiausiai modelyje svarbus kintamasis *trstprt*, nes jo standartizuotasis beta koeficientas lygus 0,034. Stulpelyje **t** pareikiamos Stjudento kriterijaus reikšmės, o stulpelyje **Sig.** – Stjudento kriterijaus *p* reikšmės kiekvienam koeficientui. Jeigu $p < 0{,}05$, tai sakome, kad atitinkamas kintamasis statistiškai reikšmingas. Matome, kad kintamieji *trstprl, trstlgl, trstun* ir *happy* yra statistiškai reikšmingi, o kintamieji *trstplt* ir *trstun* statistiškai nereikšmingi. **Todėl modelį reikės tobulinti**. Vis dėlto parodysime, kur pateikiama likusi modelio informacija. Stulpelyje **VIF** yra surašyti dispersijos mažėjimo daugikliai. Kadangi kintamajam *trstplt* gauta, kad VIF > 4, tai darome išvadą, kad modelyje yra multikolinearumo problema. Ši išvada tik patvirtina įtarimus, kurie kilo peržiūrėjus pradines regresorių koreliacijas. Gavome dar vieną patvirtinimą, kad modelį reikia tobulinti.

Išskirčių buvimą nustatome peržiūrėję Kuko matus ir DFB kiekvienam respondentui. Informaciją apie Kuko matą galima rasti lentelėje **Residual Statistics.** Priešpaskutinėje eilutėje yra informacija apie tai, kad Kuko mato maksimumas yra 0,091. Iš to išplaukia, kad ir visos Kuko mato reikšmės neviršja vieneto. Sprendžiant pagal Kuko matą, išskirčių nėra. Šis būdas naudotinas, jeigu DFB netiriame.

Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-.3928	9.1071	4.1439	1.72004	212
-- -- -- -- -- -- --	-- --	-- --	-- --	-- -- --	---
Cook's Distance	.000	.091	.006	.011	212
Centered Leverage Value	.002	.122	.028	.021	212

Kiekvienam respondentui Kuko mato ir DFB reikšmės atsirado duomenyse: nauji stulpeliai COO_1 ir DFB0_1 – DFB6_1 (nes buvo konstanta ir penki kintamieji) .

COO_1	DFB0_1	DFB1_1	DFB2_1	DFB3_1	DFB4_1	DFB5_1	DFB6_1
.01525	.06058	-.00825	-.00082	-.00410	.02005	-.01089	-.00479
.02551	.06139	-.00940	-.00813	.00992	.01697	-.00380	-.00831

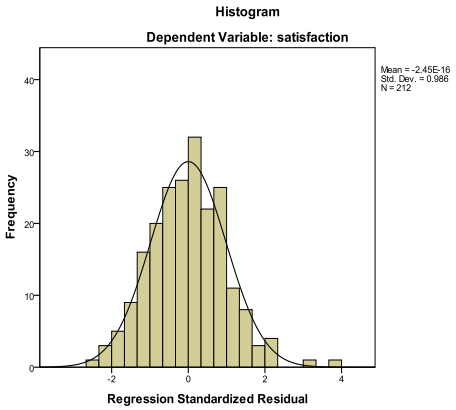
Galima tiesiog peržiūrėti visas eilutes, ieškant reikšmių viršijančių vienetą. Vis dėlto, tai nelabai patogu, nes eilučių daug. Alternatyviai galima panaudoti komandą **Analyze → Descriptive Statistics → Descriptives.** Sukeliame visus kintamuosius į **Variable(s)**. Pasirenkame **Options** ir paliekame pažymėtą tik parinktį **Maximum**. Renkamės **Continue** ir **OK**. Matome, kad visi maksimumai buvo mažesni už vienetą. Todėl darome išvadą, kad duomenyse išskirčių nebuvo.

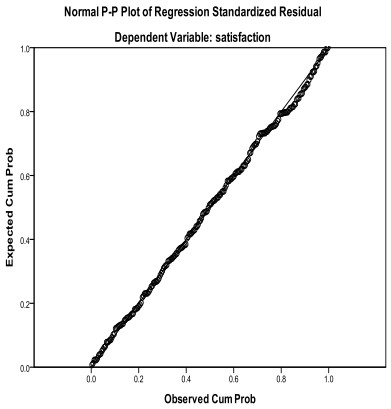
Descriptive Statistics

	N	Maximum
COO_1 Cook's Distance	212	.09110
DFB0_1 DFBETA Intercept	212	.16573
DFB1_1 DFBETA happy	212	.01580
DFB2_1 DFBETA trstlgl	212	.01754
DFB3_1 DFBETA trstplt	212	.05593
DFB4_1 DFBETA trstprt	212	.04814
DFB5_1 DFBETA trstprl	212	.02556
DFB6_1 DFBETA trstun	212	.01556
Valid N (listwise)	212	

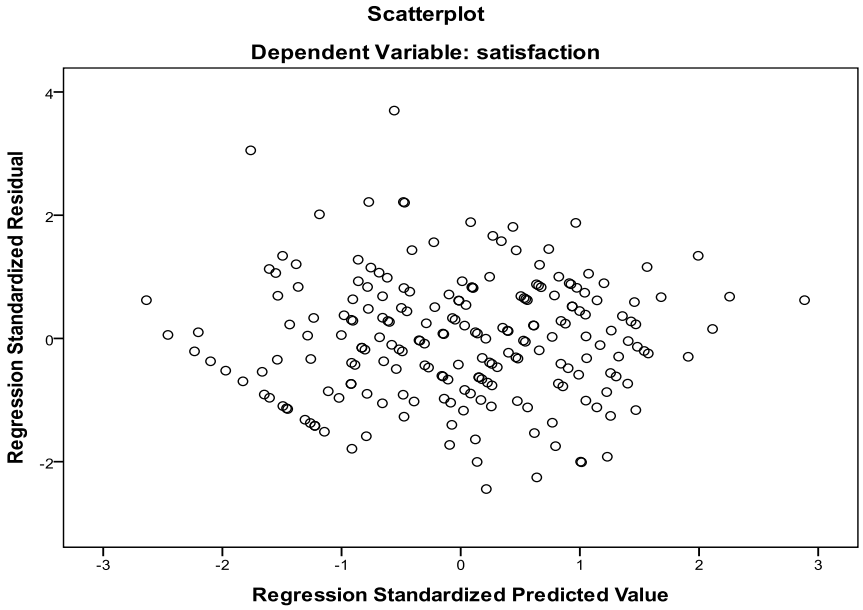
Informacija apie modelio standartizuotuojų liekamųjų paklaidų normalumą pateikiama dviejuose grafikuose. Histogramos grafike dar nubraižyta ir normalioji kreivė, t.y. tai, į ką turėtų būti panaši histograma, jeigu liekamosios paklaidos būtų normalios. Matome, kad histograma

nedaug skiriasi nuo normaliosios kreivės. Antrajame grafike atidėti standartizuotojų liekamųjų paklaidų ir normaliojo atsitiktinio kintamojo santykiniai procentinai dažniai (angl. P-P plot). Kuo taškai arčiau nubrėžtos tiesės (idealiu atveju visi taškai yra *ant* tiesės), tuo duomenys normalesni. Antrasis grafikas irgi rodo, kad standartizuotojų liekamųjų paklaidų normalumo reikalavimas tenkinamas.

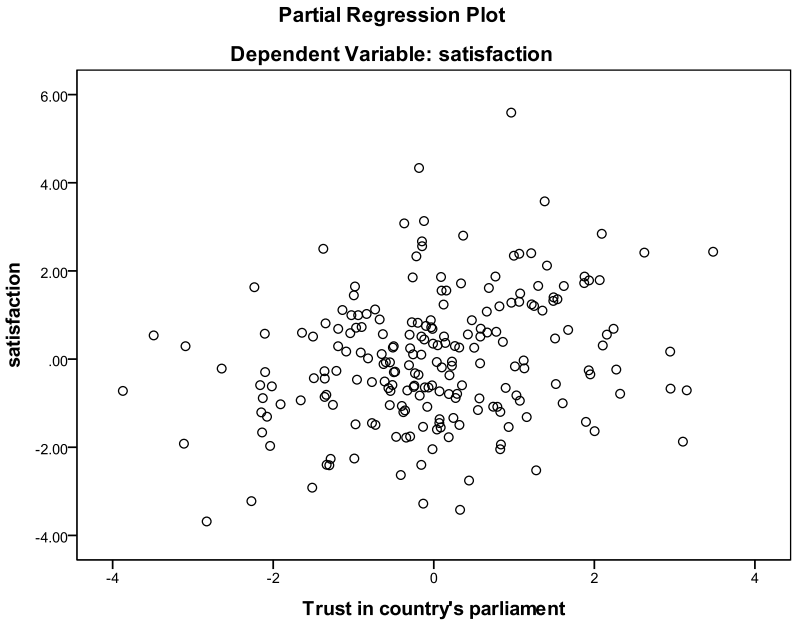




Ar duomenys nėra heteroskedastiški sprendžiame, žiūrėdami į grafiką **Scatterplot.** Nors grafiko kraštuose ir mažiau taškų, negalima įžiūrėti kažkokio standartizuotųjų liekamųjų paklaidų reguliarumo, o pats grafikas primena tolygio storio debesėlį. Tėra tik keletas nuo bendro debesėlio nutolusių taškų. Darome išvadą, kad duomenys pakankamai homoskedastiški.



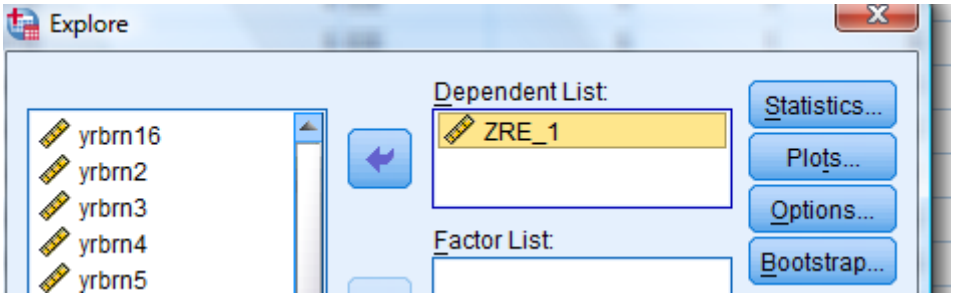
Papildoma informacija apie kiekvieno kintamojo tinkamumą modeliui yra grafikuose **Partial Regression Plot.** Kintamasis tuo geriau tinka modeliui, kuo grafike taškų išsibarstymo „debesėlis“ labiau primena tiesę. Dėl didelio regresorių tarpusavio koreliuotumo šie grafikai nėra patikimi, todėl pateikiame tik vieną pavyzdį. Ką nors nuspręsti apie kintamojo *trstprl* tikimą modeliui pagal pateiktąjį grafiką yra sunku.



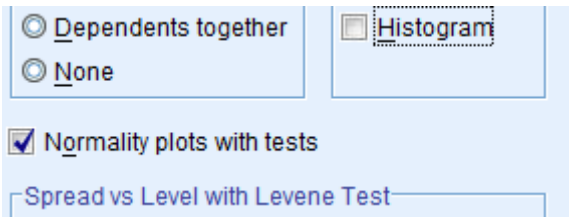
Galutinė pirminio modelio tyrimo išvada: nors determinacijos koeficientas ir pakankamai didelis, modelyje yra statistiškai nereikšmingų ir multikolinearių kintamųjų, todėl modelis netinkamas. Jį reikia tobulinti.

2.2.4. Šapiro – Vilko testas SPSS

Jau žinome, kad modelį reikės tobulinti, ir vizualiai įsitikinome, kad normalumo prielaida nėra (grubiai) pažeista. Vis dėlto, susipažinsime, kaip patikrinti liekamųjų paklaidų normalumą Šapiro – Vilko testu. Renkamės **Analyze → Descriptives → Explore** . Į laukelį **Dependent List** įkeliame *ZRE_1*. Šis standartizuotuojų liekamųjų paklaidų kintamasis buvo sukurtas vykdant regresiją (atitinkamą opciją pasirinkome, kai buvome lange **Save**).



Spaudžiame **Plots**. Pažymime **Normality plots with tests**. Spaudžiame **Continue** ir **OK**.



Rezultatų išklotinėje randame **Test of Normality**. Matome, kad Šapiro – Vilko testo p = 0,234 > 0,05. Taigi, darome išvadą, kad pagal Šapiro – Vilko testą normalumo prielaidos atmesti negalime.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_1 Standardized Residual	.036	212	.200*	.991	212	.234

a. Lilliefors Significance Correction

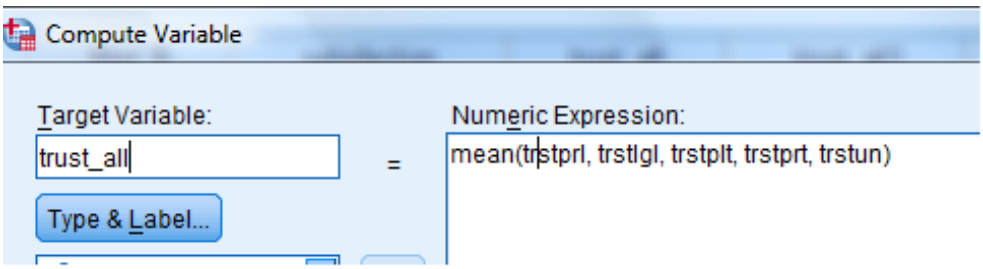
*. This is a lower bound of the true significance.

Greta dar pateiktas Komogorovo – Smirnovo testas, kuris irgi tikrina normalumą (nulinė hipotezė – kintamasis normalus). Kadangi p = 0,20 > 0,05, šis kriterijus irgi neprieštarauja normalumo prielaidai.

Reikia nepamiršti, kad visi statistiniai kriterijai jautrūs stebėjimų skaičiui. Normalumo prielaida gali būti atmesta vien todėl, kad imtyje stebėjimų labai daug. Todėl tikėti šiais testais reikia atsargiai. Laimei tiesinė regresija yra pakankamai atspari saikingiems prielaidų pažeidimams (plačiau žr. sekantį 3 skyrių). Todėl tikrinant regresijos prielaidas nereikėtų nusiristi iki paranojos. Jei labai priekabiai tikrinsim visus reikalavimus duomenims – liksim be nieko. Pasižiūrėkit į senbernius. Beje, kartu su Šapiro – Vilko testo rezultatais SPSS pateikė grafikus, kurie didelio normalumo nerodo.

2.2.5. Modelio tobulinimas

Dažniausiai modelio tobulinimas reiškia statistiškai nereikšmingų kintamųjų šalinimą. Bet ne visada. Tiriamo pavyzdžio atveju regresoriai tarpusavyje dar ir pakankamai stipriai koreliavo (buvo multikolinearumas). Todėl paprasčiausiai pakeisime kintamuosius *trstprl, trstlgl, trstplt, trstprt, trstun* jų vidurkiu, kurį pavadinsime *trust_all.* Šį kintamąjį sukuriame naudodami **Transform** → **Compute** opciją:



Po to tikriname regresijos modelio *satisfaction = f(trust_all, happy)* tinkamumą. Peržiūrėję rezultatų išklotinę, įsitikiname, kad determinacijos koeficientas praktiškai nepasikeitė: R^2= 0,590, ANOVA *p* reikšmė < 0,01, išskirčių nėra (visiems stebėjimams Kuko matas < 1). Koeficientų lentelė atrodo taip:

Coefficients[a]

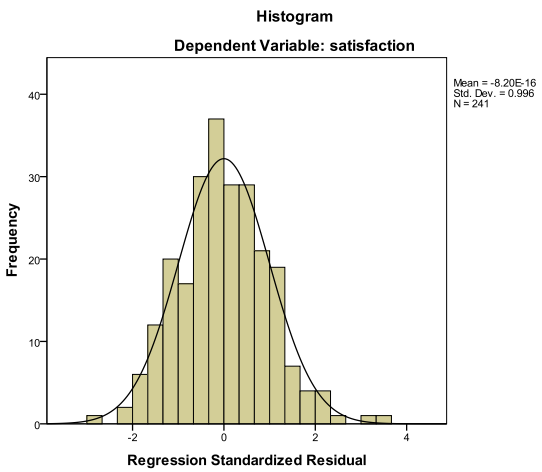
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.847	.359		-2.358	.019		
	trust_all	.824	.052	.689	15.841	.000	.911	1.098
	happy	.226	.052	.191	4.385	.000	.911	1.098

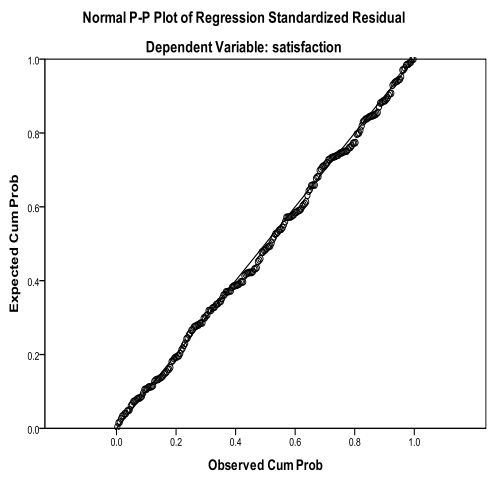
a. Dependent Variable: satisfaction

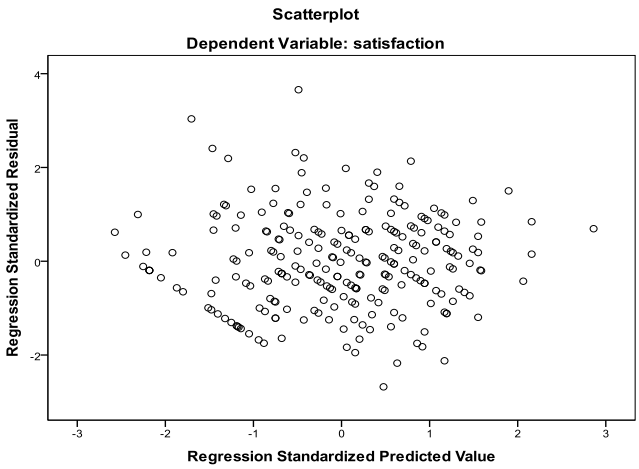
Matome, kad abudu kintamieji yra statistiškai reikšmingi, o multikolinearumo problemos nėra. Be to, iš standartizuotųjų beta koeficientų matyti, kad *trust_all* yra daug svarbesnis modelio kintamasis, nei *happy*. Regresijos modelis:

$$satisfaction = -0{,}847 + 0{,}824\ trust\_all + 0{,}226 happy.$$

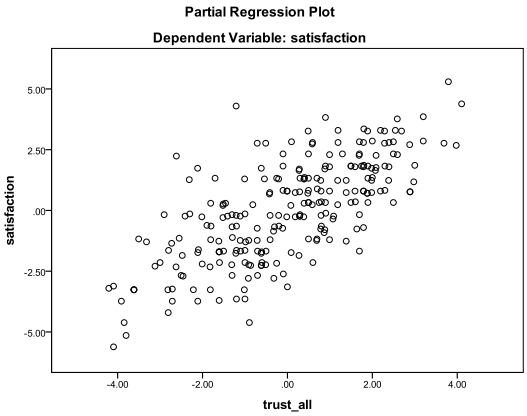
Standartizuotųjų liekamųjų paklaidų grafikai rodo, kad normalumo ir heteroskedastiškumo prielaidos patenkintos:

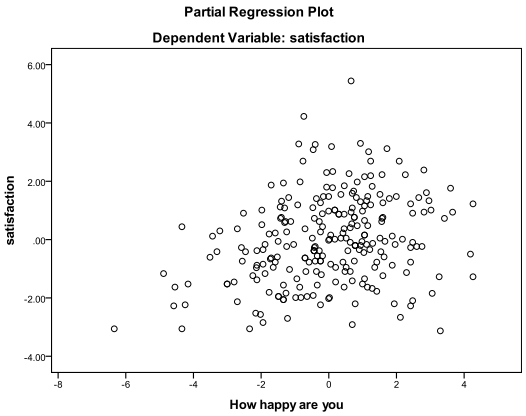






Daliniuose liekamųjų paklaidų grafikuose matyti aiški tiesinė priklausomybė kintamajam *trust_all*. Tuo tarpu kintamajam *happy*, tokios priklausomybės nesimato.





Todėl, dėl visa ko, patikriname regresijos modelį be kintamojo *happy*. Naujojo modelio determinacijos koeficientas $R^2 = 0,558$.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.747[a]	.558	.556	1.49917

a. Predictors: (Constant), trust_all

b. Dependent Variable: satisfaction

Nors kintamasis *happy* ir buvo statistiškai reikšmingas, jo pašalinimas iš modelio sumažino determinacijos koeficiento reikšmę tik per 0,032.

Visos kitos charakteristikos (ANOVA *p* reikšmė, Kuko matas, liekamųjų paklaidų grafikai) rodo, kad modelis yra tinkamas. Todėl jį ir laikysime galutiniu modeliu. Beje, į VIF nežiūrime (vis tiek jis lygus 1), nes liko tik vienas regresorius ir jam nebėra su kuo multikolinearuoti (žr. pastabas).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.379	.232		1.628	.105		
	trust_all	.894	.051	.747	17.409	.000	1.000	1.000

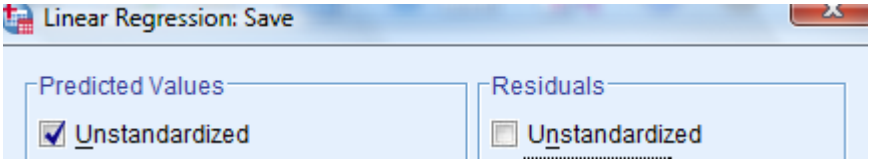
a. Dependent Variable: satisfaction

Pats tiesinės regresijos modelis atrodo taip:

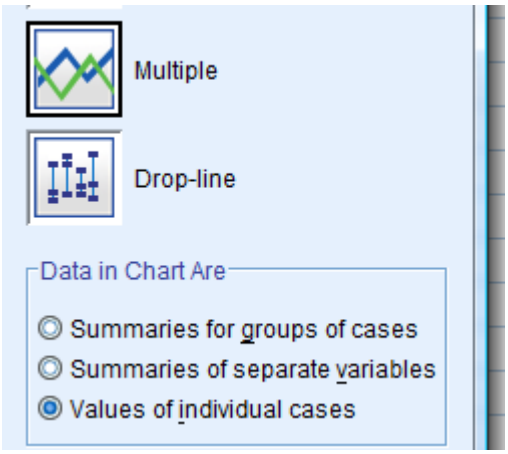
$$satisfaction = 0{,}379 + 0{,}894 trust\_all.$$

2.2.6. Prognozės ir stebėjimų grafinis palyginimas

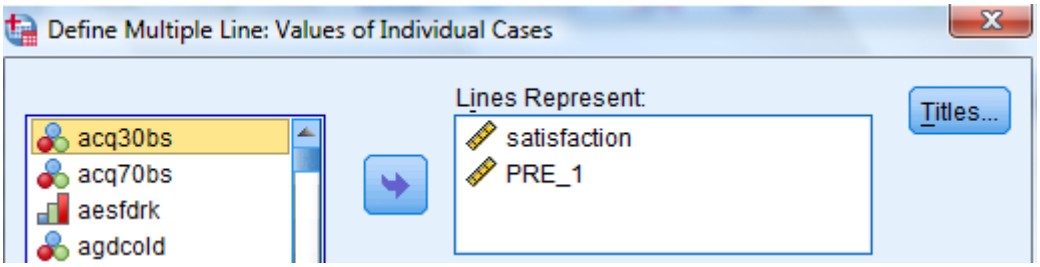
Galima grafiškai palyginti stebėtas *satisfaction* reikšmes su reikšmėmis, gautomis prognozei pritaikius aukščiau aprašytąjį regresijos modelį. Visų pirma reikia išsaugoti prognozuojamas *satisfaction* reikšmes. Tam pakanka regresijos **Save** lange pažymėti **Unstandardized**.



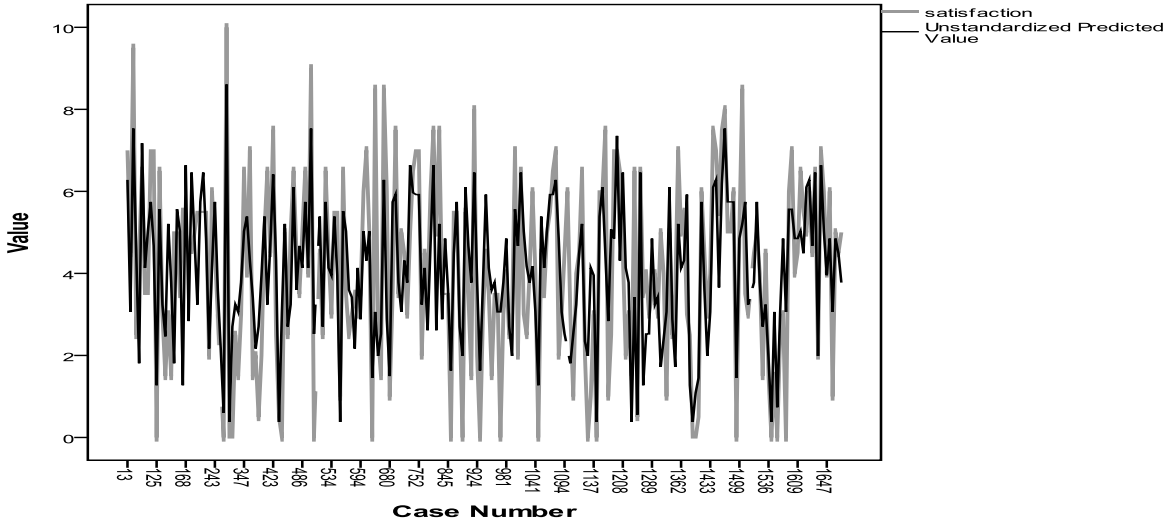
Duomenyse atsiras naujas stulpelis *PRE_1*. Prognozuojamųjų ir stebimų reikšmių grafikas gaunamas, pasirinkus **Graphs → Legacy Dialogs → Line**. Pažymime **Multiple** ir **Values of Individual Cases.** Spaudžiame **Define.**



Perkeliame *satsfaction* ir *PRE_1* į langelį **Lines Represent**. Spaudžiame **OK**.



Gautasis grafikas (kiek paredagavus, kad geriau matytųsi) atrodo taip:



Idealiu atveju, kai prognozė tiksli, abi linijos turi sutapti. Matome, kad tokių sutapimų yra gana daug. Grubiai kalbant, visai pakenčiamai tas modelis prognozuoja. Vis dėlto, grafikas palieka labai daug erdvės subjektyvumui. Kitam tyrėjui gali atrodyti, kad pagal grafiką modelio tikimas ne itin geras. Tokie grafikai beveik visada pridedami, kai tiriami laikiniai duomenys (pvz., BVP kitimas ketvirčiais). Socialiniuose tyrimuose panašūs grafikai braižomi retai. Čia jau, kaip tyrėjo širdis geidžia.

2.2.7. Rezultatų aprašymas

Visų pirma, patys sau trumpai reziumuojame, ką gi gavome. Tikėjomės, kad pasitikėjimas įvariomis šalies institucijomis siesis su požiūriu į šalies politinę sistemą. Tą ir pavyko parodyti. Tiesa, tirdami pradinį modelį, susidūrėme su matematinėmis-statistinėmis problemomis (multikolinearumu, statistišku regresorių nereikšmingumu). Šias problemas įveikėme, regresorius pakeitę jų vidurkiu.

Aprašydami rezultatus, nurodome ką tyrėme, kokius kintamuosius naudojome, ir kaip juos perkodavome. Tada kompaktiškai surašome visa tai, ką išsamiai aptarėme ankstesniuose skyreliuose. Dažnai dar pridedame vieną kitą grafiką. Pateikiame vieną iš daugybės galimų tokio aprašymo variantų .

Tyrėme modelį *satisfaction = f(trstprl, trstlgl, trstplt, trstprt, trstun, happy)*. Regresoriai *trstprl, trstlgl, trstplt, trstprt, trstun* rodo pasitikėjimą šalies parlamentu, teisine sistema, politikais, partijomis ir Jungtinėmis Tautomis, regresorius *happy* atspindi respondento laimingumo laipsnį. Pradinis modelio tyrimas parodė, kad yra multikolinearumo problema (keletui regresorių VIF > 4). Regresorius *trstprl, trstlgl, trstplt, trstprt, trstun* modelyje pakeitėme jų vidurkiu *trust_all.* Nors naujojo modelio R^2 buvo pakankamai didelis, o abudu kintamieji buvo statistiškai reikšmingi, tačiau standartizuotieji beta koeficientai ir liekamųjų paklaidų grafikai rodė, kad kintamojo *happy* įtaka kintamajam *satisfaction* nėra stipri. Todėl regresorių *happy* iš modelio pašalinome. Determinacijos koeficiento reikšmė po kintamojo *happy* pašalinimo sumažėjo tik per 0,032. Galutinis tiesinės regresijos modelis atrodo taip:

$$satisfaction = 0{,}379 + 0{,}894 trust\_all.$$

Modelio R^2 = 0,558. Liekamųjų paklaidų grafikai rodo, kad modelio normalumo ir heteroskedastiškumo prielaidos tenkinamos.

2.2.8. SPSS komandų sintaksė tiesinei regresinei analizei

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam pradiniam tiesinės regresijos modeliui. Jas galima naudoti sintaksės lange, pasirenkant opciją **Run**.

Duomenų atranka:

```
USE ALL.
COMPUTE filter_$=(agea >= 20 & agea <= 40 & gndr = 1).
VARIABLE LABELS filter_$ 'agea >= 20 & agea <= 40 & gndr = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

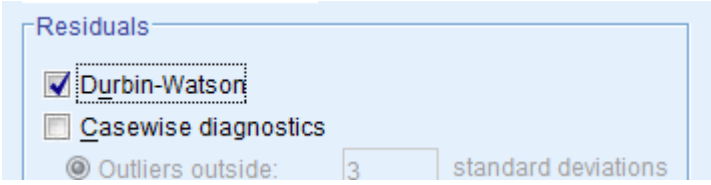
Tiesinės regresijos modelis:

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
```

```
/CRITERIA=PIN(.05) POUT(.10)   /NOORIGIN   /DEPENDENT satisfaction
/METHOD=ENTER trstprl trstlgl trstplt trstprt trstun happy
/PARTIALPLOT ALL   /SCATTERPLOT=(*ZRESID ,*ZPRED)
/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
/SAVE COOK.
```

2.2.9. Autokoreliacijos tyrimas

Ar yra autokoreliacija reikia tirti, tik esant logiškam paaiškinimui, kodėl vieno respondento atsakymas gali priklausyti nuo kito respondento atsakymo. Mūsų nagrinėtam pavyzdžiui toks spėjimas būtų visiškai nelogiškas. Nepanašu, kad antrojo respondento požiūris į politines institucijas priklauso nuo pirmojo respondento požiūrio. Vis dėlto parodysime, kaip Durbino-Vatsono kriterijus taikomas, naudojant SPSS programą. Pasirinkę **Analyze** → **Regression**→ **Linear** → **Statistics**, pažymime **Durbin-Watson**.



Tarp rezultatų suradę lentelę **Model Summary**, joje rasime ir Durbino-Vatsono statistikos reikšmę. Žemiau patekiame šios statistikos reikšmę galutiniam, aukščiau nagrinėto pavyzdžio modeliui. Matome, kad statistikos reikšmė 2,043 praktiškai nesiskiria nuo dviejų, t.y. iš tikro vieno respondento atsakymai nesusiję su kito respondento atsakymais.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.747[a]	.558	.556	1.49917	2.043

a. Predictors: (Constant), trust_all

b. Dependent Variable: satisfaction

Primename, kad autokoreliacijos nėra, jeigu Durbino-Vatsono statistika nedaug skiriasi nuo 2 (nedaug – reiškia, kad statistikos reikšmė pakliuvo į intervalą [1,5; 2,5].) Autokoreliaciją reikia tirti, kai turime stebėjimus laike. Pavyzdžiui, ketvirtinius BVP augimo duomenis.

2.2.10. Multikolinearumo problema

Primename, kad labai tarpusavyje koreliuojantys (multikolinearūs) regresoriai gali sukelti problemų. Modelis tampa nestabiliu, gali atsirasti neteisingų priklausomybių. Pateiksime vieną, kiek dirbtinį pavyzdį, parodantį, kaip multikolinearumas gali iškreipti tikrąsias kintamųjų priklausomybes. Sudarysime tiesinės regresijos modelį *trstlgl = f(trstprl, trust_all)*. Koreliacija tarp *trstlgl* ir *trstprl* yra 0,659, t.y. priklausomybė tiesioginė – kuo vieno kintamojo reikšmė didesnė, tuo didesnė ir kito kintamojo reikšmė. Kitaip sakant, tie respondentai, kurie labiau pasitiki parlamentu, yra labiau linkę pasitikėti ir teisine sistema. Į modelį įtraukėme ir stipriai su abiem kintamaisiais koreliuojantį (visai natūralu, nes abudu kintamieji buvo naudoti, jį sudarant) kintamąjį *trust_all*. Koreliacija tarp *trust_all* ir *trstprl* yra labai stipri ir lygi 0,877.

Gautojo modelio $R^2 = 0,66$, ANOVA *p* reikšmė < 0, 05. Pažvelkime į lentelę **Coefficients**. Pagal Stjudento kriterijų abudu kintamieji statistiškai reikšmingi (p=0,007 ir p=0,000). Taip pat matome, kad yra multikolinearumo problema (abudu VIF > 4). Kas įvyktų, jeigu į multikolinearumą nekreiptume dėmesio? Atsižvelgę į lentelėje pateiktus koeficientų įverčius, gautume tokį modelį:

$$trstlgl = 0,33 - 0,23trsprl + 1,372\ trust\_all.$$

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
´(Constant)	.330	.247		1.335	.183		
trstprl	-.230	.085	-.214	-2.708	.007	.231	4.338
trust_all	1.372	.109	.995	12.598	.000	.231	4.338

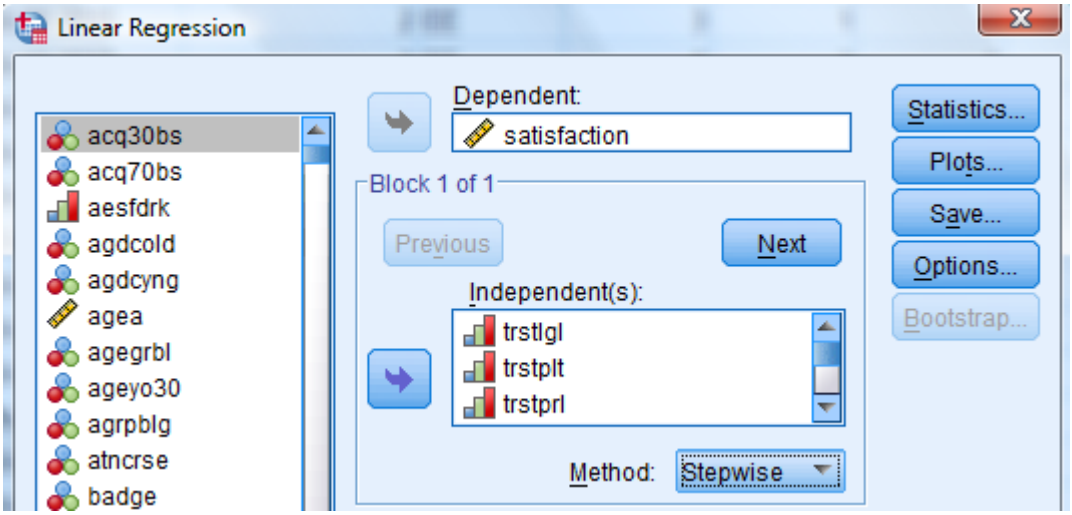
a. Dependent Variable: trstlgl

Neigiamas koeficiento prie *trstprl* ženklas reiškia, kad *trstprl* didėjant, *trstlgl* mažėja. Kitaip tariant, kuo labiau respondentas pasitiki parlamentu, tuo mažiau pastiki teisine sistema. Ši išvada prieštarauja aukščiau aptartai kintamųjų priklausomybei ir jų koreliacijos ženklui. Taigi, nekreipdami dėmesio į stiprias regresorių tarpusavio koreliacijas, galime padarytti labai klaidingas išvadas.

2.2.11. Žingsninė regresija

Primename, kad žingsninė (keliaetapė) regresija automatiškai palieka tik statistiškai reikšmingus regresorius, kuriuos atrenka pagal formalius kriterijus. Yra net keletas žingsninės regresijos rūšių.

Parodysime, kaip atliekama žingsninė regresija, nagrinėdami pirminį modelį. Tirsime modelį *satisfaction = f(trstprl, trstlgl, trstplt, trstprt, trstun, happy)* 20 - 40 m vyrams. Visos SPSS parinktys yra tokios pačios, kaip ir tikrinant pirminio modelio tinkamumą. Vienintelis skirtumas – pagrindiniame meniu pasirenkame **Method Stepwise.**



Pasirinkome tokį žingsninės regresijos metodą, kai regresoriai po vieną įtraukiami į modelį. Rezultatų išklotinėje dalis skirta aprašomajai statistikai liko tokia pati, kaip ir tiesioginės tiesinės regresijos atveju. Lentelėje **Model Summary** pateikti penkių modelių determinacijos koeficientai. Po lentele nurodyta, kokie regresoriai modeliuose naudojami. Matome, kad galutiniame modelyje yra visi regresoriai, išskyrus *trsprt*. Vis dėlto, R^2 reikšmės kito nedaug. Pavyzdžiui, neturime rimtų priežasčių teigti, kad penktasis modelis yra iš esmės geresnis už ketvirtąjį.

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.688[a]	.473	.471	1.62132
2	.726[b]	.528	.523	1.53888
3	.751[c]	.564	.558	1.48232
4	.765[d]	.585	.577	1.44882
5	.772[e]	.595	.586	1.43453

a. Predictors: (Constant), trstprl

b. Predictors: (Constant), trstprl, trstlgl

c. Predictors: (Constant), trstprl, trstlgl, happy

d. Predictors: (Constant), trstprl, trstlgl, happy , trstplt

e. Predictors: (Constant), trstprl, trstlgl, happy , trstplt, trstun

f. Dependent Variable: satisfaction

Lentelėje **ANOVA** yra visų penkių regresijos modelių ANOVA *p* reikšmės. Jos visos yra mažesnės už 0,05 . Todėl šis rodiklis nerodo, kad kuris nors modelis geresnis, nei kiti.

ANOVA[f]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	495.843	1	495.843	188.629	.000[a]
	Residual	552.020	210	2.629		
	Total	1047.862	211			
2	Regression	552.917	2	276.458	116.740	.000[b]
	Residual	494.945	209	2.368		
	Total	1047.862	211			
3	Regression	590.831	3	196.944	89.631	.000[c]
	Residual	457.031	208	2.197		
	Total	1047.862	211			
4	Regression	613.351	4	153.338	73.050	.000[d]
	Residual	434.511	207	2.099		
	Total	1047.862	211			
5	Regression	623.939	5	124.788	60.639	.000[e]
	Residual	423.923	206	2.058		
	Total	1047.862	211			

Lentelėje **Coefficients** yra informacija apie visų regresijos modelių koeficientų reikšmes, VIF ir beta koeficientus. Matome, kad visi modeliai tankina regresoriams keliamus reikalavimus: visi regresoriai yra statistiškai reikšmingi (konstantai statistiško reikšmingumo nežiūrime), koeficientų ženklai sutampa su koreliacijų ženklais iš koreliacijų lentelės (nepateikėme, bet ji sutampa su 2.2.3 skyrelio lentele), multikolinearumo problemos nėra. Ištyrę beta koeficientus, įsitikiname, kad svarbiausias regresorius yra *trstprl* (beta = 0,303).

Galutinis SPSS siūlomas modelis:

$$Satisfaction = -0{,}86 + 0{,}291 trstprl + 0{,}185 trstlgl + 0{,}233 happy + 0{,}214\ trstplt + 0{,}115\ trstun.$$

Kitų rezultatų (jie yra tik penktajam modeliui) nebepateiksime. Tik pažymėsime, kad regresijos prielaidos yra tenkinamos. Liekamųjų paklaidų grafikai rodo *satisfaction* normalumą ir pakankamai gerą homoskedastiškumą. Papildomai ištyrę Kuko mato ir DFB statistikų reikšmes, įsitikiname, kad jos visos neviršija vieneto. Darome išvadą, kad duomenyse išskirčių nėra. Nors žingsninė regresija siūlo naudoti penktąjį modelį, panašu, kad ketvirtasis modelis gali būti toks pat geras. Žodžiu, vis tiek viskas susiveda į tai, kad galutinį sprendimą teks priimti tyrėjui, remiantis savo teoriniais samprotavimais.

Coefficients[a]

Model		Unstandard. Coefficients		Standard. Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.550	.219		7.071	.000		
	trstprl	.660	.048	.688	13.734	.000	1.000	1.000
2	(Constant)	.904	.246		3.669	.000		
	trstprl	.470	.060	.489	7.842	.000	.580	1.723
	trstlgl	.269	.055	.306	4.909	.000	.580	1.723
3	(Constant)	-.373	.388		-.960	.338		
	trstprl	.445	.058	.464	7.672	.000	.574	1.742
	trstlgl	.225	.054	.256	4.177	.000	.558	1.793
	happy	.239	.058	.203	4.154	.000	.881	1.135
4	(Constant)	-.528	.382		-1.381	.169		
	trstprl	.277	.076	.289	3.634	.000	.317	3.159
	trstlgl	.211	.053	.241	4.001	.000	.554	1.805
	happy	.240	.056	.203	4.263	.000	.881	1.135
	trstplt	.265	.081	.236	3.275	.001	.386	2.589
5	(Constant)	-.860	.406		-2.119	.035		
	trstprl	.291	.076	.303	3.839	.000	.315	3.179
	trstlgl	.185	.053	.212	3.474	.001	.530	1.888
	happy	.233	.056	.197	4.178	.000	.879	1.138
	trstplt	.214	.083	.190	2.571	.011	.358	2.793
	trstun	.115	.051	.115	2.268	.024	.759	1.318

a. Dependent Variable: satisfaction

2.3. Tiesinė regresinė analizė su STATA

2.3.1. Duomenys

Priešingai SPSS programai, dirbant su STATA, komandas paprasčiau surašyti patiems, nei naudotis meniu. Tirsime tuos pačius duomenis, kaip ir SPSS atveju, tiktai pasinaudosime jau sukurtais kintamaisiais. Tiriame 2008 metų Europos Sąjungos socialinio tyrimo (European Social Survey) Estijos duomenis ESS4EE. Mus domins, kaip pasitenkinimas šalies politiniu klimatu susijęs su pasitikėjimu politinėmis institucijomis ir bendruoju žmogaus laimingumu. Tirsime 20 – 40 metų vyrus. Tyrime naudosime tokius kintamuosius:

- *satisfaction* (pasitenkinimas šalies politiniu klimatu),

- *trust_all* (pasitikėjimas politinėmis institucijomis),
- *happy* (laimingumas).

Kintamasis *satisfaction* sukurtas paėmus kintamųjų *stfgov* (kiek respondentas yra patenkintas savo šalies vyriausybe) ir *stfdem* (kiek respondentas yra patenkintas demokratijos būsena šalyje) vidurkį. Jis įgyja reikšmes nuo 0 (itin nepatenkintas) iki 10 (labai patenkintas). Kintamasis *trust_all* sukurtas, suskaičiavus kintamųjų *trstprl* (pasitikėjimas šalies parlamentu), *trstlgl* (pasitikėjimas teisine sistema), *trstplt* (pasitikėjimas politikais), *trstprt* (pasitikėjimas partijomis), *trstun* (pasitikėjimas Jungtinėmis Tautomis) vidurkį ir įgyja reikšmes nuo 0 (itin nepasitiki) iki 9 (itin pasitiki). Kintamasis *happy* matuojamas 10 balų skale nuo 0 (itin nelaimingas) iki 9 (itin laimingas).

Reikiamus respondentus atsirinksime, atsižvelgę į *agea* (amžių) ir *gndr* (lytį, 1-vyr., 2-mot.). SPSS *.*sav* formato failas buvo išsaugotas STATA *.*dta* formatu. Atidarę failą STATA programa, atsirenkame reikiamus respondentus:

```
keep if agea >= 20 & agea <= 40 & gndr == 1 & happy < 11 & satisfaction < 11
```

Reikia turėti omenyje, kad neatsakyti klausimai dažnai užkoduojami didelėmis skaitinėmis reikšmėmis. Norint, kad jos nebūtų naudojamos, rikėtų jas tiesiog pašalinti iš tyrimo. Būtent todėl STATA komandoje ir atsirado paskutinės sąlygos (STATA tuščią vietą traktuoja, kaip begalybę).

Tiesinė regresija turi prasmę tik tada, kai intervaliniai regresoriai pakankamai stipriai koreliuoja su priklausomu kintamuoju, ir pakankamai silpnai tarpusavyje. Patikriname kinatmųjų koreliacijas komanda

```
pwcorr satisfaction trust_all happy, sig
```

Gauta koreliacijų lentelė atspindi stiprią teigiamą *satisfaction* ir *trust_all* koreliaciją (r = 0,7461) ir vidutinę teigiamą *satisfaction* ir *happy* koreliaciją (r = 0,3967). Patys regresoriai *happy* ir *trust_all* koreliuoja silpnai (r = 0,2988). Multikolinearumo problema neturėtų iškilti. Visos koreliacijos yra statistiškai reikšmingos (visos *p* reikšmės lygios 0,000 < 0,05) .

	satisf~n	trust_~l	happy
satisfaction	1.0000		
trust_all	0.7461 0.0000	1.0000	
happy	0.3967 0.0000	0.2988 0.0000	1.0000

2.3.2. Regresijos modelio lygtis

Dabar patikrinsime regresijos modelį *satisfaction = f(trust_all, happy).* Įvykdome komandą

```
regress satisfaction happy trust_all, beta
```

Gauname pagrindinę informaciją apie modelį.. Determinacijos koeficientas R^2= 0,5898 yra pakankamai didelis.

Source	SS	df	MS
Model	716.420078	2	358.210039
Residual	498.181582	238	2.09319992
Total	1214.60166	240	5.06084025

Number of obs =	241
F(2, 238) =	171.13
Prob > F =	0.0000
R-squared =	0.5898
Adj R-squared =	0.5864
Root MSE =	1.4468

satisfaction	Coef.	Std. Err.	t	P>|t|	Beta
happy	.2260815	.0515548	4.39	0.000	.1907632
trust_all	.8241485	.0520248	15.84	0.000	.6891185
_cons	-.8467996	.3590563	-2.36	0.019	.

Įstatę į regresijos modelį koeficientų įverčius, gauname:

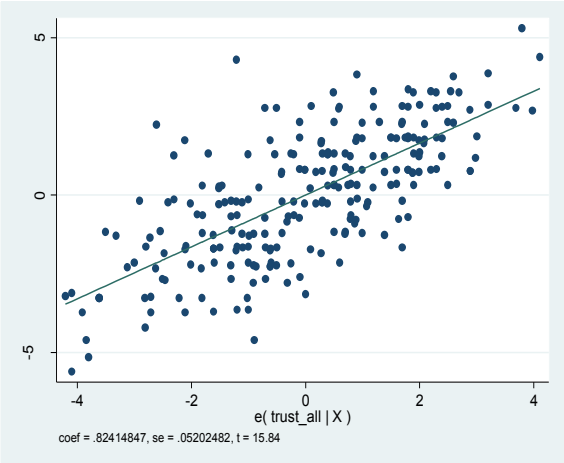
satisfaction = -0,847+0,824 *trust_all + 0,226happy*

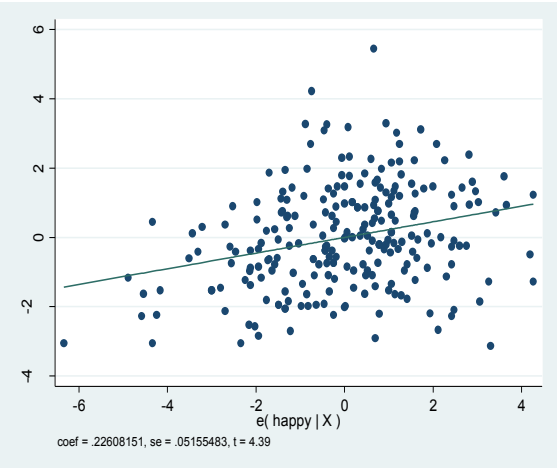
ANOVA *p* reikšmė yra eilutėje **Prob > F.** Ji mažesnė už 0,05. Taigi modelyje yra statistiškai reikšmingų kintamųjų. Iš lentelės matyti, kad abudu kintamieji yra statistiškai reikšmingi, nes jiems t kriterijaus *p* reikšmės (jos yra stulpelyje **P>|t|**) mažesnės už 0,05.

Norėdami nustatyti, kuris kintamasis modelyje yra svarbesnis, pažiūrime į standartizuotuosius beta koeficientus. Kintamasis *trust_all* (jo beta koeficientas lygus 0,689) yra daug svarbesnis nei *happy* (beta lygus 0,191). Papildomai nubraižysime dalinius liekamųjų paklaidų grafikus abiems regresoriams. Tai galima atlikti viena komanda apps. Vis dėlto, aiškesnis vaizdas yra, kai grafikus braižome kiekvienam regresoriui atskirai. Paeiliui įvykdome dvi komandas:

```
avplot trust_all
avplot happy
```

Gautuosiuose daliniuose liekamųjų paklaidų grafikuose matyti aiški tiesinė priklausomybė kintamajam *trust_all*. Tuo tarpu kintamajam *happy*, tokios akivaizdžios priklausomybės nesimato.





Taigi, pagrindinės modelio charakteristikos rodo gerą modelio tikimą duomenims. Vis dėlto, kintamojo *happy* reikalingumas modelyje yra abejotinas.

2.3.3. Regresijos prielaidų tikrinimas

Dar reikia patikrinti modelio prielaidas:

- ar liekamosios paklaidos homoskedastiškos,
- ar liekamosios paklaidos normalios,
- ar nėra išskirčių,
- ar regresoriai nėra multikolinearūs.

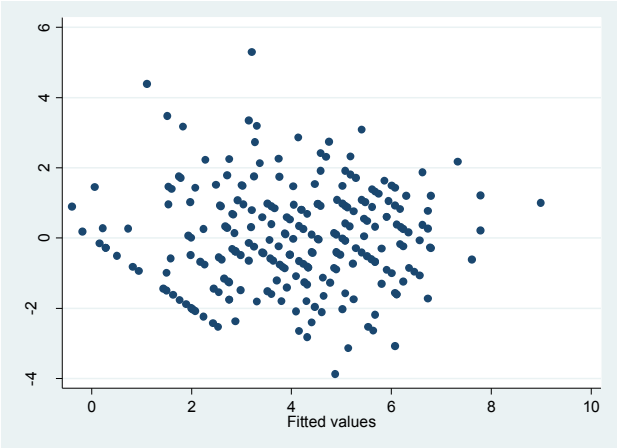
Norėdami patikrinti, ar tenkinama homoskedastiškumo prielaida, nubraižome liekamųjų paklaidų ir prognozuojamųjų *satisfaction* reikšmių išsibarstymo grafiką. Tai galima padaryti, tiesiog įvykdžius komandą:

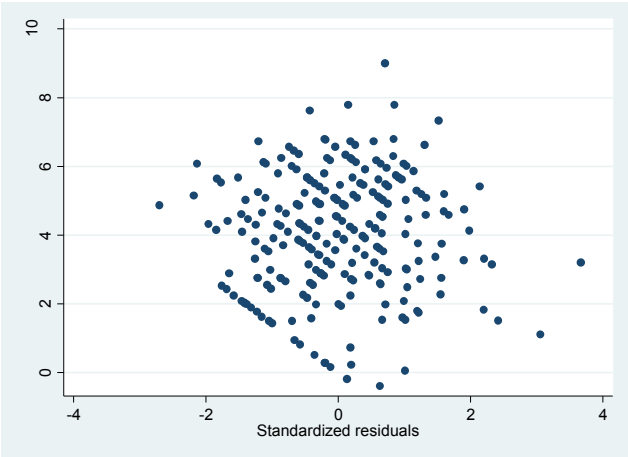
```
rvfplot
```

Alternatyviai galima nubraižyti standartizuotųjų liekamųjų paklaidų ir prognozuojamųjų *satisfaction* reikšmių išsibarstymo grafiką. Tai galima padaryti, paeiliui įvykdžius tris komandas

```
predict a
predict r1, rstandard
scatter a r1
```

Gautieji grafikai skiriasi nedaug ir rodo, kad duomenys nėra labai heteroskedastiški, liekanos išsibarsčiusios daugiau mažiau tolygiai (grafikas primena juostelę).





Galima įtarti, kad homoskedastiškumo prielaida tenkinama. Patikrinsime, ar duomenys tam statistiškai reikšmingai neprieštarauja, taikydami Breušo – Pagano kriterijų. Nulinė šio kriterijaus hipotezė teigia, kad homoskedastiškumo prielaida galioja. Homoskedastiškiems duomenims *p* reikšmė turi būti **ne mažesnė** už 0,05. Įvykdome komandą

```
estat hettest
```

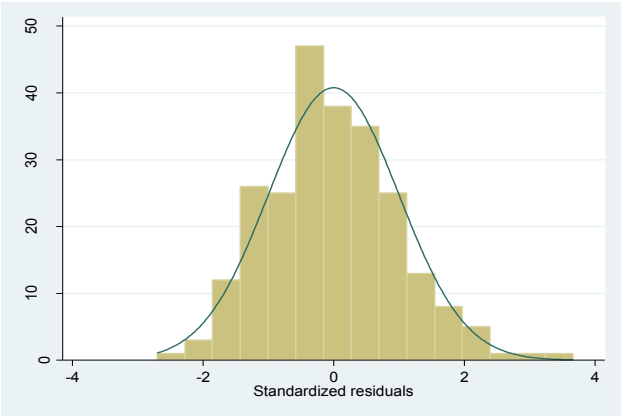
Gautoji chi kvadrato statistikos reikšmė yra 2,19, o $p = 0{,}1391 \geq 0{,}05$. Darome išvadą, kad duomenys statistiškai reikšmingai neprieštarauja homoskedastiškumo prielaidai.

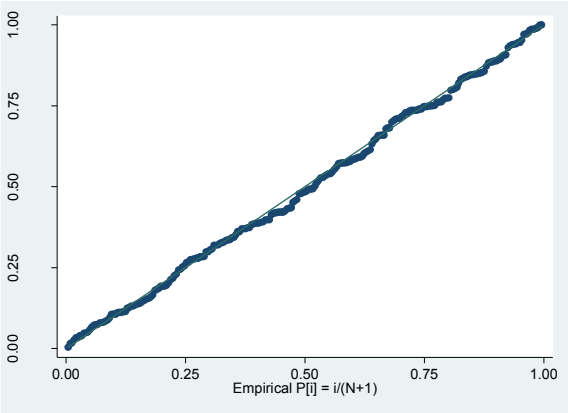
```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of satisfaction

         chi2(1)      =     2.19
         Prob > chi2  =   0.1391
```

Norime, atkreipti dėmesį, kad *p* reikšmės gali būti mažos vien dėl imties didumo. Todėl visada reikia atsižvelgti ir į chi kvadrato statistikos reikšmę (homoskedastiškiems duomenims ji maža) ir į liekamųjų paklaidų grafikus.

Tikrindami liekamųjų paklaidų normalumą, visų pirma nubraižysime standartizuotųjų liekamųjų paklaidų histogramą (ji turėtų būti panaši į normaliąją kreivę) ir kvantilių P-P grafiką (grafiko taškai turi nedaug skirtis nuo tiesės). Standartizuotųjų liekamųjų paklaidų kintamąjį *r1* sukūrėme komanda predict r1, rstandard. Paeiliui įvykdome dvi komandas:

```
histogram r1, frequency normal
pnorm r1
```





Matome, kad ir histograma ir kvantilių grafikas rodo liekamųjų paklaidų normalumą. Dabar patikrinsime, ar paklaidų normalumą Šapiro – Vilko testu. Nulinė hipotezė teigia, kad paklaidos normalios. Normalumo prielaida nepatenkinta, jeigu $p < 0{,}05$. Įvykdome komandą

```
swilk r1
```

```
                  Shapiro-Wilk W test for normal data
    Variable |   Obs        W          V          z     Prob>z
-------------+--------------------------------------------------
          r1 |   241    0.99257      1.305      0.617   0.26847
```

Įsitikiname, kad $p = 0{,}268 > 0{,}05$. Todėl galima teigti, kad sprendžiant pagal Šapiro – Vilko testą, liekamosios paklaidos turi normalųjį skirstinį. Kaip ir tikrinant homogeniškumą, reikia nepamiršti, kad p reikšmės gali būti mažos vien dėl imties didumo. Todėl visada reikia atsižvelgti ir į grafikus nes didelėms imtims jie informatyvesni.

Patikrinsime, ar duomenyse nebuvo nebuvo išskirčių. Atsižvelgsime į regresorių DFB reikšmes ir į Kuko matą. Sukuriame DFB kiekvienam regresoriui (naujieji kintamieji turės vardus Dftrust_all ir Dfhappy) ir pasižiūrime jų maksimalias reikšmes:

```
dfbeta  trust_all happy

summarize DFtrust_all  DFhappy
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
 DFtrust_all |       241   -.0000299     .060555  -.2955649   .2377913
     DFhappy |       241   -2.38e-06    .0634298  -.3248331   .2693256
```

Išskirtimi laikomas tas duomuo, kuriam DFB > 1. Kadangi maksimalios DFB reikšmės mažesnės už vienetą, tai tokių duomenų imtyje nėra.

Kuko mato reikšmes kiekvienam respondentui sukuriame (pavadiname naująjį stulpelį Coo) ir pažiūrime maksimalią rekšmę analogiškai:

```
predict Coo , cooksd
summarize Coo
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Coo	241	.0039244	.0066371	7.96e-08	.0515792

Įsitikiname, kad ir Kuko mato maksimali reikšmė neviršija vieneto.Darome išvadą, kad duomenyse išskirčių nėra.

Galimą kintamųjų multikolinearumą patikrinsime suradę dispersijos mažėjimo daugiklius VIF. Įvykdome komandą

```
estat vif
```

Multikolinearumo problema kyla, jeigu yra VIF > 4. Abiems regresoriams VIF < 1. Todėl darome išvadą, kad kintamųjų multikolinearumo nėra.

Variable	VIF	1/VIF
happy	1.10	0.910707
trust_all	1.10	0.910707
Mean VIF	1.10	

2.3.4. Galutinis modelis

Ankstesniuse skyreliuose įsitikinome, kad regresijos modelis labai gerai aprašo duomenis. Vis dėlto nedidelė *happy* ir *satisfaction* koreliacija, maža beta koeficiento reikšmė kintamajam *happy* ir netiesinis liekanų išsibarstymo grafikas kelia abejonių dėl šio regresoriaus reikalingumo. Todėl patikriname regresijos modelį be jo.

```
regress satisfaction  trust_all, beta
```

Matome, kad naujojo modelio R^2= 0,5567 nuo ankstesniojo R^2= 0,5898 skiriasi tik per 0,0331, t.y. determinacijos koeficientas sumažėjo labai nedaug.

Source	SS	df	MS
Model	676.16673	1	676.16673
Residual	538.43493	239	2.25286581
Total	1214.60166	240	5.06084025

Number of obs =	241
F(1, 239) =	300.14
Prob > F =	0.0000
R-squared =	0.5567
Adj R-squared =	0.5548
Root MSE =	1.501

satisfaction	Coef.	Std. Err.	t	P>|t|	Beta
trust_all	.8923218	.0515065	17.32	0.000	.7461222
_cons	.38224	.2328408	1.64	0.102	.

Darome išvadą, kad regresorių *happy* galima iš modelio pašalinti. *Nepaisant to, kad jis buvo statistiškai reikšmingas.* Galutinis modelis:

satisfaction = 0,382 +0,892 *trust_all.*

Naujajame modelyje liko tik vienas regresorius, todėl nebereikia tikrinti multikolinearumo, nebereikia aprašinėti standartizuotosios beta reikšmės, o ANOVA ir t testo *p* reikšmės sutaps. Dar reikėtų ištirti, ar tenkinamos regresijos prielaidos, t.y. įvykdyti komandas

```
avplot trust_all
rvfplot
predict r2, rstandard
estat hottest
histogram r2, frequency normal
pnorm r2
```

```
swilk r2
dfbeta  trust_all
summarize DFtrust_all
predict Coo1 , cooksd
summarize Coo1
```

Nebepateiksime gautųjų rezultatų, tik konstatuosime, kad visos regresijos modelio prielaidos buvo tenkinamos.

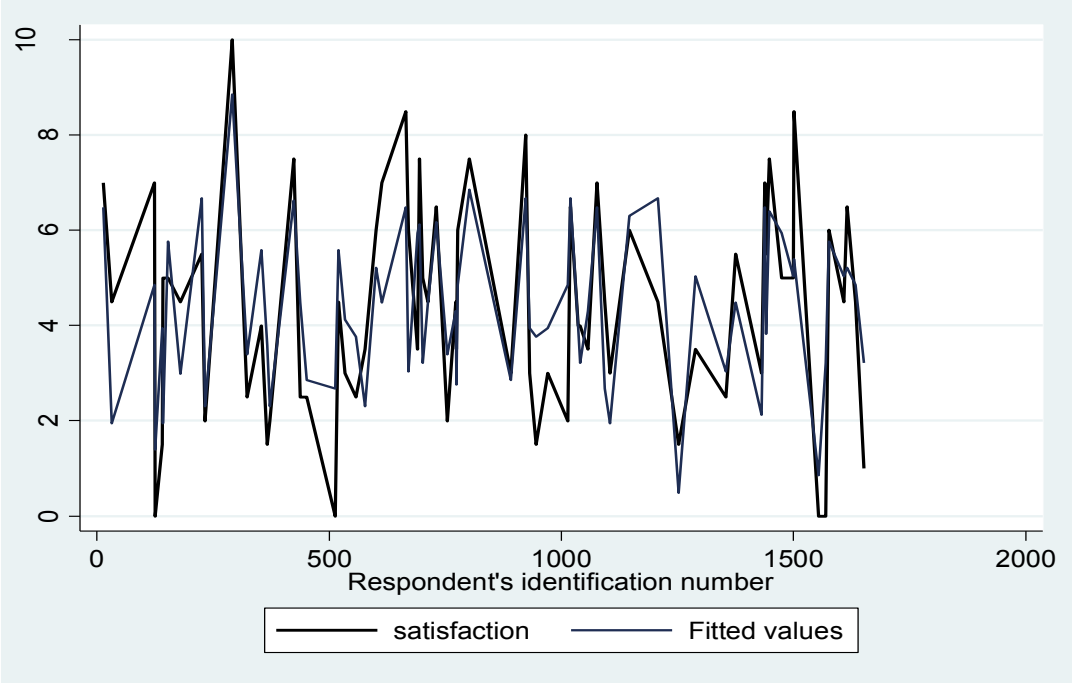
2.3.5. Prognozės ir stebėjimų grafinis palyginimas

Sukuriame prognozuojamų *satisfaction* reikšmių stulpelį duomenyse (pavadiname jį *PRED*) įvykdydami komandą:

```
predict PRED
```

Braižydami grafiką, pasinaudosime duomenyse esančiu respondento identifikaciniu numeriu *idno*. Įvykdome:

```
graph twoway line  satisfaction PRED  idno
```



Kuo labiau sutampa linijos, tuo modelio prognozės tikslesnės. Matome, kad modelis gana geras.

2.3.6. Išvados

Regresijos aprašymas iš esmės nesiskiria nuo pateiktojo 2.2.5 skyrelio pabaigoje. Tik jį reikia papildyti informacija apie Breušo – Pagano ir Šapiro – Vilko testus. Pilnas aprašymas turėtų apimti naujajam modeliui pakartotą visą tyrimą. Lakoniškas (straipsniui tinkamas) aprašymas gali skambėti, kad ir taip:

Tyrėme modelį *satisfaction = f(trust_all, happy).* Nors R^2 buvo pakankamai didelis, o abudu kintamieji buvo statistiškai reikšmingi, tačiau standartizuotieji beta koeficientai ir liekamųjų paklaidų grafikai rodė, kad kintamojo *happy* įtaka kintamajam *satisfaction* nėra stipri. Todėl regresorių *happy* iš modelio pašalinome. Determinacijos koeficiento reikšmė po kintamojo *happy* pašalinimo sumažėjo tik per 0,03. Galutinis tiesinės regresijos modelis atrodo taip: *satisfaction* = 0,382 +0,892 *trust_all.* Modelio R^2 = 0,558.

Visi duomenims keliami reikalavimai buvo tenkinami. Šapiro – Vilko testas parodė, kad liekamųjų paklaidų normalumo prielaida tenkinama. Pritaikę Breušo – Pagano testą, įsitikinome, kad galioja ir homoskedastiškumo prielaida. DFB ir Kuko mato reikšmės, leidžia teigti, kad duomenyse išskirčių nebuvo.

2.3.7. Autokoreliacija

Tirtiesiems pavyzdžiams tai visiškai neaktualus klausimas nes nebuvo jokio įtarimo, jog duomenyse gretimų respondentų atsakymai kažkaip susiję. Vis dėlto, parodysime, kaip su STATA apskaičiuoti Durbino – Vatsono statistiką. Tam reikia turėti vadinamąjį laiko kintamąjį (dažnai tai metai, kada buvo atliktas matavimas, galimi ir niuansai, jeigu duomenys ketvirtiniai). Kadangi tokio kintamojo tiriamame faile nėra, tai dirbtinai apsiribojame pirmaisiais šimtu respondentu (keep if idno < = 100) ir pavadiname kintamąjį *idno* laiko kintamuoju:

```
tsset idno
```

Atlikę regresiją (reg satisfaction trust_all), įvykdome komandą

```
dwstat
```

Gautoji Durbino – Vatsono statistikos reikšmė (1,7496) yra pakankamai arti 2, todėl manome, kad autokoreliuotumo nėra. Primename, kad autokoreliuotumas mažai tikėtinas, jeigu Durbino – Vatsono statistika yra tarp 1,5 ir 2,5.

```
Number of gaps in sample:  4

Durbin-Watson d-statistic(  2,    96) =  1.749634
```

Beje, **Number of gaps in sample** rodo, kad ne visos eilutės buvo panaudotos regresijoje (ne visoms turėjome duomenų ir apie *satisfaction* ir apie *trust_all*).

2.3.8. Multikolinearumo problema

Pateiksime vieną, kiek dirbtinį pavyzdį, parodantį, kaip multikolinearumas gali iškreipti tikrąsias kintamųjų priklausomybes. Sudarysime tiesinės regresijos modelį, regresuodami požiūrį į vyriausybę (*stfgov*) pagal požiūrį į parlamentą (*trstprl*) ir politikus (*trstplt*). Visiems kintamiesiems didesnė reikšmė rodo didesnį pasitikėjimą. Tiriame tuos pačius 20 – 40 metų Estijos vyrus, kaip ir ankstesniuose skyreliuose. Sąlyginis regresijos modelis: *stsfgov = f(trstprl, trstplt)*. Įvykdome:

```
reg stfgov trstprl trstplt
```

Source	SS	df	MS
Model	13050.5097	2	6525.25483
Residual	16357.8577	242	67.5944532
Total	29408.3673	244	120.526096

Number of obs =	245
F(2, 242) =	96.54
Prob > F =	0.0000
R-squared =	0.4438
Adj R-squared =	0.4392
Root MSE =	8.2216

stfgov	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
trstprl	.6721899	.060272	11.15	0.000	.5534652	.7909147
trstprt	-.1205474	.0554678	-2.17	0.031	-.2298087	-.0112861
_cons	1.821313	.5783163	3.15	0.002	.6821367	2.960489

Atrodytų, kad modelis visai gerai tinka: determinacijos koeficientas R^2 = 0,44, ANOVA *p* reikšmė = 0,000; abudu regresoriai statistiškai reikšmingi (atitinkamos t kriterijaus *p* reikšmės yra 0,031 ir 0,000) . Patikriname multikolinearumą:

```
estat vif
```

Variable	VIF	1/VIF
trstprl	2.04	0.489249
trstprt	2.04	0.489249
Mean VIF	2.04	

Abiejų regresorių VIF < 4. Taigi, formaliai žiūrint, jokio kintamųjų multikolinearumo nėra.

Užrašykime gautąją regresijos lygtį:

$$stfgov = 1{,}82 + 0{,}67\ trstprl - 0{,}12\ trstplt.$$

Koeficientas prie *trstprl* yra teigiamas. Kuo labiau respondentas pasitiki savo parlamentu, tuo labiau ir savo šalies vyriausybe. Visiškai natūrali priklausomybė. Koeficientas prie *trstplt* yra neigiamas. Kuo labiau respondentas *nepasitiki* politikais apskritai, tuo labiau pasitiki savo vyriausybe. Labai keista priklausomybė nes vyriausybėje irgi sėdi politikai. Prieš pradėdami kurti naują teoriją apie specifinį Estijos gyventojų mentalitetą (vyriausybės nesieja su politikais, o sieja su dievų duota valdžia, pasąmonėje tupi noras turėti monarchą), pažiūrėkime į kintamųjų koreliacijas:

```
corr stfgov trstprl trstplt
```

	stfgov	trstprl	trstplt
stfgov	1.0000		
trstprl	0.6580	1.0000	
trstplt	0.4869	0.8872	1.0000

Visos koreliacijos *teigiamos!* Labiau pasitiki vyriausybe tie respondentai, kurie *labiau* pasitiki politikais. Tiesiog regresijoje buvo multikolinearumas (koreliacija tarp *trstprl* ir *trstplt* yra 0,887), nors VIF to ir neparodė.

> Tikrinant multikolinearumą reikia pažiūrėti, ar modelio priklausomybės atitinka koreliacijas. Vien VIF nepakanka.

2.3.9. Žingsninė regresija

Parodysime, kaip atliekama žingsninė (pažingsninė, keliaetapė) regresija, nagrinėdami ESS4 Lietuvos duomenis. Tiriame visus 20 – 30 metų respondentus. Norime išsiaiškinti, nuo ko priklauso pasitikėjimas Seimu. Tirsime modelį

trstprl = f(stfeco, stflife, stfgov, stfdem, stfedu, stfhlth).

Čia *trstprl* yra pasitikėjimas parlamentu, *stfeco, stflife, stfgov, stfdem, stfedu, stfhlth* atitinkamai – pasitenkinimas ekonomika, gyvenimu apskritai, vyriausybe, demokratijos padėtimi, švietimo būkle, sveikatos apsauga. Visi kintamieji intervaliniai (0-9 arba 10), didesnė reikšmė rodo palankesnį požiūrį. Norima sužinoti, kuri sritis daro didesnę įtaką parlamento autoritetui. Panaudosime žingsninę regresiją. Žingsniuoti galima ir į priekį ir atgal (akumuliuojant, kaupiant kintamuosius arba eliminuojant, šalinant kintamuosius). Todėl ir atitinkamų regresijų yra ne viena. Išbandysime du tokių regresijų modelius.

Duomenis *dta* formatu išsaugojome, naudodami SPSS programą. Todėl, norėdami išvengti neteisingo praleistų reikšmių traktavimo paliekame tik reikiamus respondentus:

```
keep if agea >= 20 & agea <= 30 & stfeco < 11 & trstprl < 11&  stflife < 11
& stfgov < 11 &  stfdem < 11 & stfedu < 11 &  stfhlth < 11
```

Tyrimą pradėsime nuo regresijos, kai „žingsniuojama pirmyn ir tik pirmyn“, t.y., kai į modelį įtraukiami vis nauji statistiškai reikšmingi regresoriai (akumuliujančioji, kaupiamoji regresija, angl. forward stepwise regression). Procesas baigiamas, kai neįtraukti lieka tik statistiškai nereikšmingi regresoriai. Įvykdome komandą:

```
sw reg  trstprl  stflife stfeco stfgov stfdem stfedu stfhlth, pe(0.05) beta
```

Parinktis pe(0.05) reiškia, kad į modelį bus įtraukti tik tie regresoriai, kuriems t kriterijaus *p* reikšmė mažesnė už 0,05.

```
. sw reg  trstprl  stflife stfeco stfgov stfdem stfedu stfhlth, pe(0.05) beta
                      begin with empty model
p = 0.0000 <  0.0500  adding   stfgov
p = 0.0009 <  0.0500  adding   stfdem
p = 0.0415 <  0.0500  adding   stflife
p = 0.0252 <  0.0500  adding   stfhlth

      Source |       SS       df       MS              Number of obs =     298
-------------+------------------------------           F(  4,   293) =   50.18
       Model |  436.987469     4  109.246867           Prob > F      =  0.0000
    Residual |  637.848101   293  2.17695598           R-squared     =  0.4066
-------------+------------------------------           Adj R-squared =  0.3985
       Total |  1074.83557   297  3.61897498           Root MSE      =  1.4755

------------------------------------------------------------------------------
     trstprl |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      stfgov |   .4374456   .0605192     7.23   0.000                  .444568
      stfdem |   .1679884   .0465637     3.61   0.000                 .2221915
     stflife |  -.0944838   .0400954    -2.36   0.019                -.1156208
     stfhlth |    .093558   .0415772     2.25   0.025                 .1134544
       _cons |   .7043779   .2359976     2.98   0.003                        .
------------------------------------------------------------------------------
```

Matome, kad modelis vis buvo pildomas, pridedant naują regresorių. Į galutinį modelį nepakliuvo *stfedu* ir *stfeco.* Determinacijos koeficientas $R^2 = 0{,}41$. Galutinis modelis:

$$trstprl = 0{,}70 + 0{,}44\ stfgov + 0{,}17\ stfdem - 0{,}09\ stflife + 0{,}09\ stfhlth.$$

Tūlas politikas tokį rezultatą (neteisingai) gali interpretuoti taip: norint, kad jauno amžiaus rinkėjai labiau gerbtų parlamentą, reikia rūpinantis sveikatos apsauga, puoselėti demokratiją ir prižiūrėti, kad gerai dirbtų vyriausybė (koeficientai prie šių regresorių yra teigiami) . Be to, nėra ko labai stengtis, kad žmonės jaustųsi laimingi (koeficientas prie *stflife*) neigiamas. Na, o ekonomikos ir švietimo būklė požiūrio į parlamentą neveikia (kintamieji buvo statistiškai nereikšmingi). Prie šios interpretacijos dar grįšime, o dabar tiems patiems duomenims pritaikykime kitą modelį.

Pasirinksime regresiją, kai „žingsniuojama atgal ir tik atgal“, t.y., kai pradedama nuo modelio su visais regresoriais, palaipsniui po vieną vis šalinant statistiškai nereikšmingus iš jų (eliminuojančioji, šalinamoji regresija, angl. backward stepwise regression). Procesas baigiamas, kai modelyje lieka tik statistiškai reikšmingi regresoriai. Įvykdome komandą:

```
sw reg  trstprl  stflife stfeco stfgov stfdem stfedu stfhlth, pr(0.05) beta
```

Parinktis pr(0.05) reiškia, kad iš modelio bus šalinami tie regresoriai, kuriems t kriterijaus *p* reikšmė nemažesnė už 0,05.

```
. sw reg  trstprl  stflife stfeco stfgov stfdem stfedu stfhlth, pr(0.05) beta
                      begin with full model
p = 0.3668 >= 0.0500  removing stfeco
p = 0.1597 >= 0.0500  removing stfedu

      Source |       SS       df       MS              Number of obs =     298
-------------+------------------------------           F(  4,   293) =   50.18
       Model |  436.987469     4  109.246867           Prob > F      =  0.0000
    Residual |  637.848101   293  2.17695598           R-squared     =  0.4066
-------------+------------------------------           Adj R-squared =  0.3985
       Total |  1074.83557   297  3.61897498           Root MSE      =  1.4755

------------------------------------------------------------------------------
     trstprl |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     stflife |  -.0944838   .0400954    -2.36   0.019               -.1156208
     stfhlth |    .093558   .0415772     2.25   0.025                .1134544
      stfgov |   .4374456   .0605192     7.23   0.000                 .444568
      stfdem |   .1679884   .0465637     3.61   0.000                .2221915
       _cons |   .7043779   .2359976     2.98   0.003                       .
------------------------------------------------------------------------------

.
```

Matome, kad modelis vis buvo siaurinamas, šalinant regresorius. Gautasis modelis niekuo nesiskiria nuo ankstesnio (taip būna ne visada). Ramybės dėlei patikriname multikolinearumą. Įvykdome:

```
estat vif
```

```
    Variable |       VIF       1/VIF
-------------+----------------------
      stfdem |      1.87    0.533968
      stfgov |      1.87    0.535419
     stfhlth |      1.26    0.796739
     stflife |      1.19    0.841321
-------------+----------------------
    Mean VIF |      1.55
```

Visų regresorių VIF < 4. Taigi, formaliai žiūrint, jokio kintamųjų multikolinearumo nėra.

Dabar įsitikinsime, kad modelis nėra teisingas. Pažiūrėkime kintamųjų koreliacijas. Įvykdome

```
corr  trstprl  stflife stfeco stfgov stfdem stfedu stfhlth
```

```
             |  trstprl  stflife   stfeco   stfgov   stfdem   stfedu  stfhlth
-------------+---------------------------------------------------------------
     trstprl |   1.0000
     stflife |   0.1254   1.0000
      stfeco |   0.5148   0.3677   1.0000
      stfgov |   0.6035   0.2887   0.7702   1.0000
      stfdem |   0.5098   0.3715   0.6197   0.6522   1.0000
      stfedu |   0.2644   0.3139   0.3586   0.3942   0.4416   1.0000
     stfhlth |   0.3481   0.2651   0.4331   0.4176   0.3583   0.6456   1.0000
```

Matome, kad *trstprl* su visais kintamaisiai koreliuoja *teigiamai*. Taigi, regresijos modelyje ženklas prie *stflife* yra neteisingas (ir politiko išvada, kad geriau, kai respondentai nelaimingi). Mes susidūrėme su tipine multikolinearumo problema. Visi kintamieji stipriai koreliuoja, nors VIF jiems

ir nėra labai dideli (apie tai žr. ankstesnį skyrelį). Įdomu, kad pašainus iš modelio kintamąjį *stflife*, jis lieka visai tinkamas: $R^2 = 0{,}395$ (o buvo $R^2 = 0{,}407$).

```
. reg trstprl  stfhlth stfgov stfdem, beta

      Source |       SS       df       MS              Number of obs =     298
-------------+------------------------------           F(  3,   294) =   64.07
       Model |  424.898867     3  141.632956           Prob > F      =  0.0000
    Residual |  649.936703   294  2.21066906           R-squared     =  0.3953
-------------+------------------------------           Adj R-squared =  0.3891
       Total |  1074.83557   297  3.61897498           Root MSE      =  1.4868

------------------------------------------------------------------------------
     trstprl |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     stfhlth |   .0798142   .0414836     1.92   0.055                 .0967878
      stfgov |   .4335071   .0609627     7.11   0.000                 .4405653
      stfdem |   .1419986   .0455877     3.11   0.002                 .1878158
       _cons |   .3362553   .1782571     1.89   0.060                        .
------------------------------------------------------------------------------
```

Dabar pažiūrime, kaip atrodo modelis, kai *trstprl* regresuojame pagal *stfedu*:

```
. reg trstprl  stfeco

      Source |       SS       df       MS              Number of obs =     298
-------------+------------------------------           F(  1,   296) =  106.76
       Model |  284.900997     1  284.900997           Prob > F      =  0.0000
    Residual |  789.934573   296  2.66869788           R-squared     =  0.2651
-------------+------------------------------           Adj R-squared =  0.2626
       Total |  1074.83557   297  3.61897498           Root MSE      =  1.6336

------------------------------------------------------------------------------
     trstprl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      stfeco |   .5110578   .0494621    10.33   0.000     .4137158    .6083998
       _cons |    .889902   .1448872     6.14   0.000     .6047625    1.175042
```

Nors šis modelis ir blogiau tinka, nei ankstesnieji, bet visai jo atmesti negalima. Jeigu politikas tyrimą būtų pradėjęs nuo šio modelio, tai darytų išvadą, kad pasitikėjimas parlamentu auga, kai gerėja ekonominė situacija.

Kokią išvadą galimą padaryti iš pateiktųjų pavyzdžių? Kokią regresiją bepasirinktume, gautą modelį reikia įvertinti kritiškai, būtina pažiūrėti koreliacijas, ir pamąstyti apie modelio suderinamumą su sveiku protu (žinoma teorija, ankstesniais tyrimais ir pan.)

> Atlikus žingsninę regresiją, būtina patikrinti modelio logiškumą. Neapsiriboti tik formaliais statistikos rodikliais.

Uždaviniai

1. Failas **ZTLT**. Tyrimas: Lietuvos tolerancijos profiliai, 2003 m. Klausimų grupės:

- *r01_01 – r01_06* respondento požiūris (skalėje 1 – teigiamas, 2 – neutralus, 3 – neigiamas, 4 – sunku pasakyti) į juodaodžius, musulmonus, žydus, čigonus, migrantus, pabėgėlius).
- *r02_01 – r02_04* ranginėje skalėje įvertinta, koks procentas nurodytos grupės (kitos rasės, kitos tautybės, kitos religijos, kitos seksualinės orientacijos) gyventojų nekeltų Lietuvoje problemų.
- *r04_01 – r04_09* Lietuvoje turėtų būti toleruojama (skalėje 1 – visiškai sutinku, 2 – sutinku, 3 – nesutinku, 4 – visiškai nesutinku) kitos tautybės, kitos rasės, kitos krikščioniškos konfesijos, nekrikščioniškos religijos, suteikiamas prieglobstis politiniams pabėgėliams, suteikiamas prieglobstis ekonominiams pabėgėliams, visi Lietuvos gyventojai turi turėti lygias galimybes įsidarbinti ir socialines garantijas).
- *r06_01 – r06_09* ar galima pateisinti darbdavį, jeigu jis nepriima į darbą (skalėje 1 – niekada negalima, 2 – kartais galima, 3 – visada galima, 4 – sunku pasakyti) dėl to, kad kitos rasės, kito tikėjimo, neįgalus, vyras, moteris, kitos tautybės, homoseksualus, senas, jaunas.
- *s01* – lytis.
- *s02* – amžius.

Sukurti naujus kintamuosius:

- *a1 = r01_01 + ...+ r01_06*, tiems respondentams, kurie nepasirinko nė į vieną klausimą atsakymo 4. Kintamasis *a1* atspindi respondento požiūrį į svetimšalius. Kuo mažesnė *a1* reikšmė, tuo požiūris tolerantiškesnis.
- *a2 = r02_01 + ...+ r01_04*. Kintamasis *a2* atspindi respondento pakantumą „kitokiems“. Kuo didesnė *a2* reikšmė, tuo pakantumas didesnis.
- *a4 = r04_01 + ...+ r04_09*. Kintamasis *a4* atspindi respondento nuomonę apie valstybės, kaip institucijos, geidautiną tolerantiškumo laipsnį. Kuo mažesnė *a4* reikšmė, tuo valstybė tolerantiškesnė.
- *a6 = r06_01 + ...+ r06_09*, tiems respondentams, kurie nepasirinko nė į vieną klausimą atsakymo 4. Kintamasis *a6* atspindi respondento nuostatas apie tolerantiškumą darbo santykiuose. Kuo mažesnė *a6* reikšmė, tuo tolerantiškumas didesnis.

Reikia sudaryti regresijos modelį *a4 = f(a1, a2, a6, s02).* Atsakyti į klausimus:

1) Ar visi kintamieji modelyje reikalingi?
2) Ar teisingi koeficientų ženklai?
3) Kurie kintamieji modelyje svarbesni?
4) Ar tenkinamos regresijos prielaidos?

5) Ar modelis vyrams panašus į modelį moterims? Ar daug skiriasi modeliai, sudaryti atskirai vyrams ir moterims nuo modelio su pseudokintamuoju *s01?*
6) Kokius regresijos modelius gausime, taikydami žingsninę regresiją?

2. Failas **ESS4LT1.sav**. Socialinis Lietuvos respondentų tyrimas 2009 m. spalis – 2010 m. Reikia atlikti Lietuvos 20 – 40 metų vyrams, tą patį tyrimą, kuris skyrelyje buvo atliktas su Estijos gyventojais. Primename, ką matuoja kintamieji:

- *stfgov* (kiek respondentas yra patenkintas savo šalies vyriau-sybe),
- *stfdem* (kiek respondentas yra patenkintas demokratijos būsena šalyje),
- *trstprl* (pasitikėjimas šalies parlamentu),
- *trstlgl* (pasitikėjimas teisine sistema),
- *trstplt* (pasitikėjimas politikais),
- *trstprt* (pasitikėjimas partijomis),
- *trstun* (pasitikėjimas Jungtinėmis Tautomis),
- *happy* (laimingumas).

Kintamieji *stfgov* ir *stfdem* matuojami 10 balų skale nuo 0 (itin nepatenkintas) iki 10 (itin patenkintas). Kintamieji *trstprl, trstlgl, trstplt, trstprt, trstun* matuojami 10 balų skale nuo 0 (itin nepasitikiu) iki 9 (itin pasitikiu). Kintamasis *happy* matuojamas 10 balų skale nuo 0 (itin nelaimingas) iki 9 (itin laimingas).

Modeliuosime kintamųjų *stfgov* ir *stfdem* vidurkį (pavadinkime jį *satisfaction*). Šis kintamasis atspindi respondentų pasitikėjimą politiniu šalies klimatu. Manome, kad *satisfaction* gali priklausyti nuo likusiųjų kintamųjų: *satisfaction = f(trstprl, trstlgl, trstplt, trstprt, trstun, happy)*.

Reikiamus respondentus atsirinkti, atsižvelgus į *agea* (amžių) ir *gndr* (lytį, 1-vyr., 2-mot.), Reikia:

1) Atlikti pradinę modelio analizę (kartu ištirti, ar tenkinamos tiesinės regresijos prielaidos).
2) Patikrinti, kaip atrodo modelis, kai kintamieji *trstlgl, trstplt, trstprt, trstun* keičiami jų reikšmių vidurkiu *trust_all.*
3) Ar naujajame modelyje reikalingas kintamasis *happy*?
4) Patikrinti, kokį modelį gausime, taikydami žingsninę regresiją?
5) Ar toks pats modelis gausis, jeigu tirsime vyresnius nei 60 metų respondentus?

3. Failas **LAMS**: Universitetų absolventų integracija darbo rinkoje, 2009 m. lapkritis – 2010 m. sausis. Kintamieji:

- *K2* – aukštoji mokykla
- *K10_1* – bakalauro studijos forma (1 – dieninės),
- *K11_1* – studijų vidurkio vertinimas (1 – dažniausiai vienas žemiausių kurse,..., 5 – dažniausiai vienas aukščiausių kurse),
- *K17* – bendro darbo stažo laikas,
- *K32* – vidutinės pajamos per mėnesį (rangais nuo 1 – iki 1000, ..., 9 – virš 10000)
- *K33_1* – pasiekiant dabartinę padėtį darbo rinkoje išsilavinimas (1 - visai nereikšmingas, ..., 5 – labai reikšmingas),
- *K33_2* – pasiekiant dabartinę padėtį darbo rinkoje studijų metu įgytos žinios (1 - visai nereikšmingos, ..., 5 – labai reikšmingos),
- *K33_3* – pasiekiant dabartinę padėtį darbo rinkoje darbo patirtis (1 - visai nereikšminga, ..., 5 – labai reikšminga),
- *K33_4* – pasiekiant dabartinę padėtį darbo rinkoje asmeninės savybės (1 - visai nereikšmingos, ..., 5 – labai reikšmingos),
- *K33_5* – pasiekiant dabartinę padėtį darbo rinkoje socialiniai ryšiai (1 - visai nereikšmingi, ..., 5 – labai reikšmingi),
- *K37_1* – pasitenkinimas esamu darbu (1 – labai nepat. , ..., 5 – labai pat.).

Atsirinkus dienines bakalauro studijas baigusiuosius, taikant žingsninę regresiją modelį

K37_1 = f(K11_1, K17, K32, K33_1, K33_2, K33_3, K33_4, K33_5)

ištirti: a) veterinarijos akademijai, b) LŽŪA absolventams.

Ar visos prielaidos patenkintos? Ar galima taikyti tiesinės regresijos modelį?

3. TIESINĖS REGRESIJOS ALTERNATYVOS

Kai netenkinamos modelio prielaidos, klasikinė tiesinė regresija nebetinka. Tada visai tikėtinas toks inkvizitoriaus (oponento, dėstytojo) pokalbis su tardomuoju (doktorantu, studentu):

To *amen*[9] tardomasis visai netrokšta. Išvada: reikia kažkokio tiesinės regresijos pakaitalo. Šiame skyrelyje aptarsime kelias alternatyvas. Kada kokią alternatyvą naudoti? Dažniausiai, kai

- pažeista homoskedastiškumo prielaida – taikoma stabilizuotų liekamųjų paklaidų regresija,
- duomenyse yra didelių išskirčių – taikoma atsparioji regresija,
- pažeista normalumo prielaida – taikoma medianos regresija,
- kintamųjų priklausomybės netiesinės – taikoma netiesinė regresija.

Nereikia šio sąrašo laikyti neginčijamu ir vienintelių teisingu[10]. Visų pirma, beveik niekada nebūna pažeista tik viena prielaida. Be to, mes išvardijome ne visus galimus tiesinės regresijos prielaidų pažeidimus. Ką daryti, kai yra regresorių multikolinearumas jau buvo aptarta ankstesniame skyriuje. Tuo atveju reikia tiesiog pakeisti tiesinės regresijos kintamuosius. Jeigu duomenys autokoreliuoti, tiesinės regresijos alternatyva yra laikinių sekų (laiko eilučių) modeliai. Jų šiuose konspektuose nenagrinėjame.

Tiesinė regresija labai gerai ištyrinėta, daugybė taikymų rodo jos praktinį naudingumą. Alternatyvos nėra taip gerai išanalizuotos, jų tikimą duomenims patikrinti daug sunkiau. Todėl nereikia alternatyviųjų regresijų laikyti panacėja nuo visų bėdų. Tai šiaudas (kartais gana storas), kurio stveriamasi, kai tiesinė regresija netinka. Čia panašiai, kaip su aviečių arbata ir karštį mušančiais vaistais. Nėra vaistų, tai bent arbatos išgersime. Gal ir padės.

Alternatyvų taikymo rezultatas ir yra tik toks: suradome kažkokius modelio parametrų įverčius, kurie *galbūt* yra prasmingi. Daugybės visokių testų taikyti jau neišeis, aprašymai sutrumpės radikaliai, bet bent jau jausimės per daug nenusišnekėję. Tardomasis inkvizitoriu galės atsakyti: *Buvo nepatenkinta ..(įrašyti tinkamą).. tiesinės regresijos prielaida. Todėl parametrų įverčius apskaičiavome, taikydami ...(įrašyti tinkamą)... regresiją. Va, koks aš gudrus (neįrašyti).*

---

[9] Bakalauriniam, magistriniam, disertacijai.
[10] Kaip ir visų šių konspektų. Čia ne šv. Varanopso apreiškimas.

Nereikia dėl menkiausio prielaidos pažeidimo atsisakyti tiesinės regresijos. Ji gana atspari tokiems pažeidimams, o analizė bus daug išsamesnė.

Idealiu atveju, alternatyvios regresijos duos labai panašius parametrų įverčius ir bus pateikiamos, kaip pagrindinio tyrimo *priedas*.

3.1. Tiesinei regresijai netinkanys duomenys

Šis skyrelis yra pagalbinis. Norėsime pademonstruoti regresijos alternatyvų veikimą konkretiems duomenims. Visų pirma, metodus taikysime geriems, t.y. visas tiesinės regresijos prielaidas tenkinantiems duomenims. Tam naudosime ankstesnio skyrelio medžiaga. Apibrėšime kitas, jau ne visai tiesinei regresijai tinkamas, duomenų aibes.

3.1.1. Heteroskedastiški ir ne visai normalūs duomenys

Tyrimui pasirinksime failo **LAMS** (Universitetų absolventų integracija darbo rinkoje, 2009 m. lapkritis – 2010 m. sausis) dalį. Tirsime ISM ir veterinarijos akademijos studentus (*K2* = 15 ir *K2* = 6). Kiti kintamieji:

- *K32* – vidutinės pajamos per mėnesį (rangais nuo 1 – iki 1000, ..., 9 – virš 10000)
- *K33_4* – pasiekiant dabartinę padėtį darbo rinkoje asmeninės savybės (1 - visai nereikšmingos, ..., 5 – labai reikšmingos),
- *K37_1, ...K37_4* – pasitenkinimas įvairiais darbo aspektais (1 – labai nepat. , ..., 5 – labai pat.).

Sukuriame *K37suminis = K37_1+K37_2+K37_3+K37_4.* Šis kintamasis atspindi respondento pasitenkinimą darbu. Naudodami STATA programą, nepamirštame pašalinti reikšmes 99 (jomis SPSS buvo koduoti neatsakymai).

Nagrinėsime modelį *K37suminis = f(K32, K33_4).* koreliacijas. Įvykdome

```
reg  K37suminis K32 K33_4
```

Source	SS	df	MS
Model	280.445041	2	140.222521
Residual	657.579152	121	5.43453845
Total	938.024194	123	7.62621296

Number of obs =	124
F(2, 121) =	25.80
Prob > F =	0.0000
R-squared =	0.2990
Adj R-squared =	0.2874
Root MSE =	2.3312

K37suminis	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
K32	.8030617	.156602	5.13	0.000	.4930267	1.113097
K33_4	1.242025	.26922	4.61	0.000	.7090329	1.775017
_cons	7.412178	1.265357	5.86	0.000	4.907069	9.917286

Determinacijos koeficientas $R^2 = 0{,}299 > 0{,}20$, nors ir nėra labai didelis, bet pakankamas. ANOVA rodo gerą modelio tikimą (*p = 0,00..).* Pagal t kriterijų abudu regresoriai statistiškai reikšmingi. Tikrindami modelio prielaidas, paeiliui įvykdome komandas:

```
predict r1, rstandard
swilk r1
estat hettest
estat vif
corr  K32 K33_4 K37suminis
```

```
dfbeta  K32 K33_4
summarize  DFK32 DFK33_4
predict Cook , cooksd
summarize Cook
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r1	124	0.98870	1.118	0.250	0.40141

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of K37suminis

         chi2(1)      =     3.29
         Prob > chi2  =   0.0697
```

Variable	VIF	1/VIF
K32	1.01	0.993862
K33_4	1.01	0.993862
Mean VIF	1.01	

	K32	K33_4	K37sum~s
K32	1.0000		
K33_4	0.0783	1.0000	
K37suminis	0.4191	0.3829	1.0000

```
. summarize  DFK32 DFK33_4

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       DFK32 |        124   -.0003293    .0938039  -.5807394   .2671409
     DFK33_4 |        124    .0000477    .0869116   -.372424   .3053622

. summarize Cook

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        Cook |        124    .0080659    .0153015   9.72e-07   .1276695
```

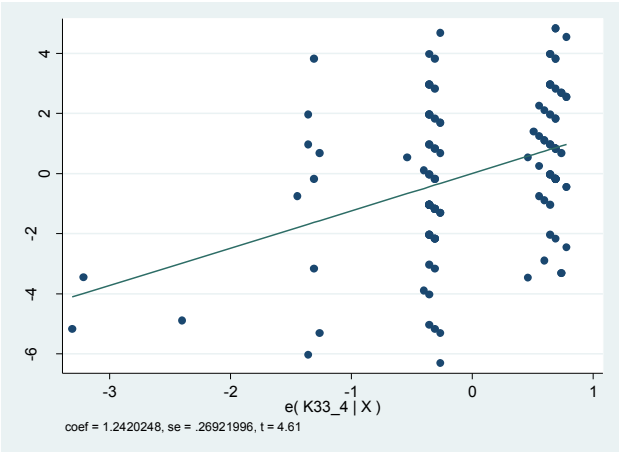
Formaliai žiūrint, visos tiesinės regresijos prielaidos patenkintos: Šapiro – Vilko testas neatmeta liekamųjų paklaidų normalumo (p = 0,401), Breušo – Pagano testas rodo, kad patenkinta liekanų homoskedastiškumo prielaida (p = 0,0697). Multikolinearumo nėra – lygties ženklai atitinka koreliacijas, o abudu VIF < 4. Esminių išskirčių irgi nėra – abiejų regresorių DFB < 1, o Kuko matas < 1.

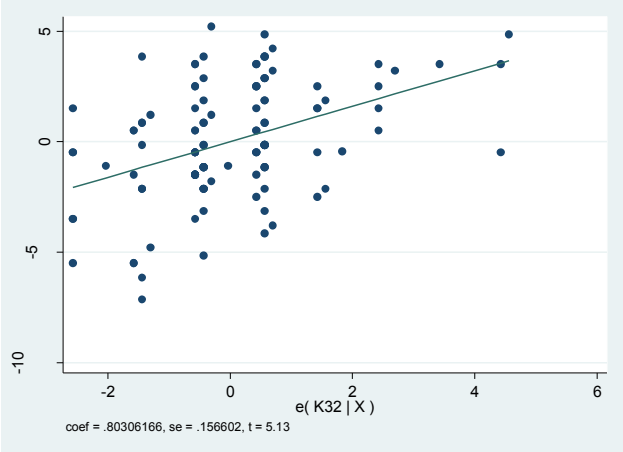
Vis dėlto stebėjimų nebuvo labai daug (n = 124). Todėl Šapiro – Vilko ir Breušo – Pagano testų p reikšmės galėjo būti nepagrįstai didelės. Be to, neatmesta prielaida, dar nėra jos įrodymas.

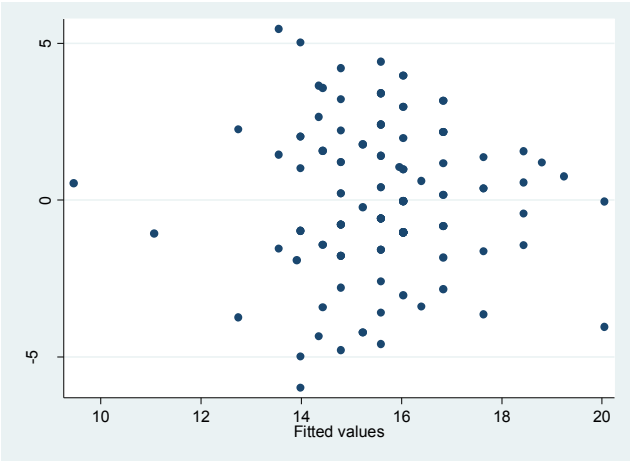
Neatmestos nulinės hipotezės *nereiškia,* kad *įrodėme* liekamųjų paklaidų homoskedastiškumą ir normalumą.

Mums *nepavyko* įrodyti priešingai. Todėl pasibraižome grafikus. Įvykdome komandas:

```
avplot  K33_4
avplot K32
rvfplot
```



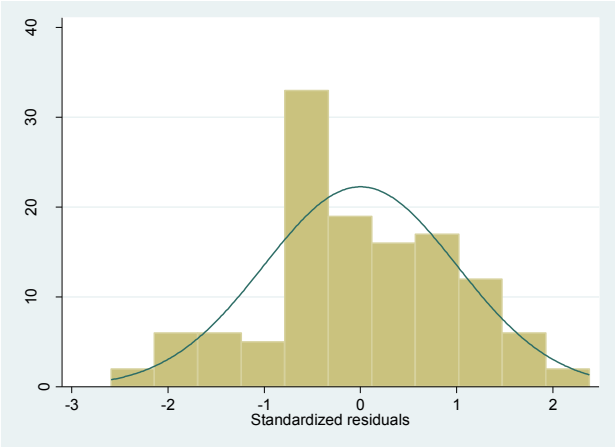


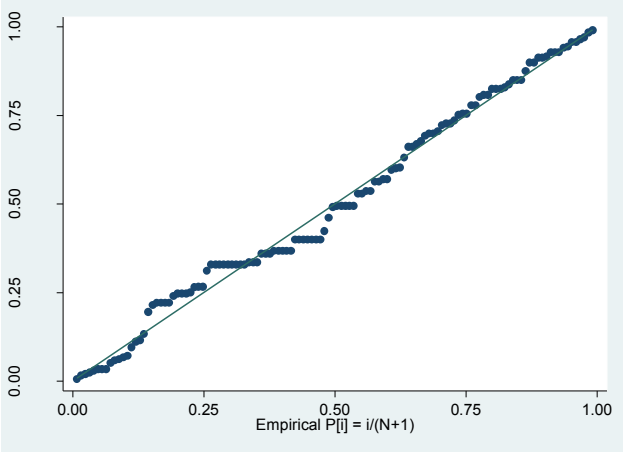


Matome, kad ir abu regresoriai ir liekamosios paklaidos yra gana heteroskedastiški. Visi grafikai primena „vėduokles‘ – tai tipinis heteroskedastiškumo požymis (homoskedastiškiems duomenims grafikai primena „juosteles“). Kita vertus, nematyti, kad duomenyse būtų išskirčių.

Vizualiai patikrinsime normalumą. Įvykdome komandas:

```
histogram r1, frequency normal
pnorm r1
```





Matome, kad ir normalumo prielaida tenkinama ne visiškai.

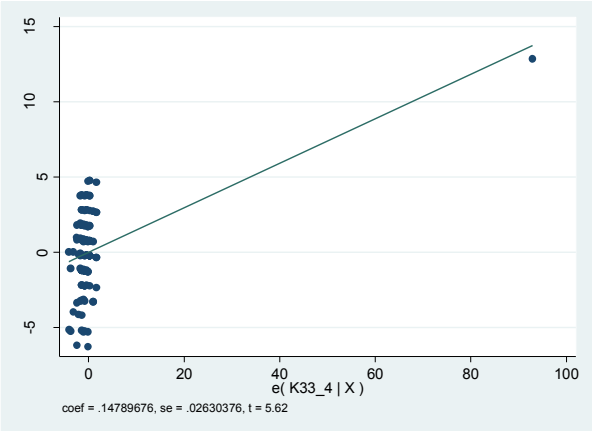
3.1.2. Duomenys su išskirtimis

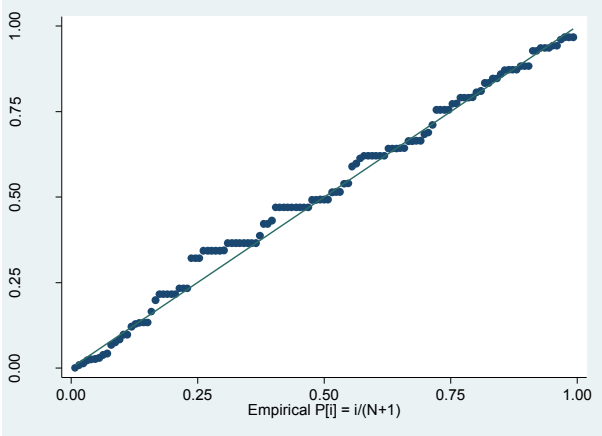
Tirdami tiesinės regresijos alternatyvas, nagrinėsime ne tik ankstesnio skyrelio pavyzdį, bet ir tą patį failą, papildytą nauja duomenų eilute su sąmoningai paliktomis klaidingomis išsiskiriančiomis reikšmėmis: *K33_4* = 99, *K37_ suminis* = 30 ir neišsiskiriančia *K32* = 5. Atlikę jiems analizę, analogišką aprašytajai ankstesniame skyrelyje, gauname

Source	SS	df	MS
Model	395.94264	2	197.97132
Residual	748.10536	122	6.13201115
Total	1144.048	124	9.22619355

Number of obs =	125
F(2, 122) =	32.28
Prob > F =	0.0000
R-squared =	0.3461
Adj R-squared =	0.3354
Root MSE =	2.4763

K37suminis	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
K32	.8470274	.1659539	5.10	0.000	.5185051	1.17555
K33_4	.1478968	.0263038	5.62	0.000	.0958258	.1999677
_cons	11.99029	.6220137	19.28	0.000	10.75896	13.22163





Brėžiniuose akivaizdžiai matyti *K33_4* išskirtis ir duomenų nenormalumas. Be to, šiems duomenims Šapiro – Vilko testas rodo, kad normalumo prielaida nepatenkinta statistiškai reikšmingai ($p = 0{,}029 < 0{,}050$).

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r2	125	0.97664	2.327	1.896	0.02897

3.1.3. Duomenų charakteristikų santrauka

Pateiksime tiesinės regresijos pagrindinius rezultatus visoms trims duomenų aibėms:

1) *A duomenimis* pavadiname Estijos duomenis, po visų jų transformacijų, aptartų skyriuje 2.3. Šie duomenys tenkina visas tiesinės regresijos prielaidas, $R^2 = 0{,}557$. Regresijos lygtis su koeficientų įverčių 95 % pasikliautiniais intervalais ir t testų *p* reikšmėmis atrodo taip:

$$satisfaction = \underset{\substack{[-0{,}076;\ 0{,}841] \\ (\ 0{,}102)}}{0{,}382} + \underset{\substack{[\ 0{,}791;\ 0{,}994] \\ (0{,}00)}}{0{,}892}\ trust\_all.$$

2) *B duomenys* – tai skyrelyje 3.1.1 aptartieji duomenys. Jiems būdingas gana rimtas heteroskedastiškumas ir švelnus normalumo prielaidos pažeidimas. Išskirčių nėra. R^2 = 0,299. Regresijos lygtis su koeficientų įverčių 95 % pasikliautiniais intervalais ir t testų *p* reikšmėmis atrodo taip:

K37suminis = 7,41 +	0,80 *K32* +	1,24 *K33_4.*	
	[4,907; 9,917].	[0,493; 1,113]	[0,709; 1,775]
	(0,00)	(0,00)	(0,00)

3) *C duomenys* – tai skyrelyje 3.1.2 aptartieji duomenys. Jiems būdingos didelės išskirtys, statistiškai reikšmingas normalumo prielaidos pažeidimas ir gana ryškus heteroskedastiškumas. R^2 = 0, 346. Regresijos lygtis:

K37suminis = 11,99 +	0,84 *K32* +	0,147 *K33_4.*	
	[10,758; 13,221].	[0,518; 1,175]	[0,096; 0,199]
	(0,00)	(0,00)	(0,00)

Palyginę B ir C duomenų modelius, matome, kad viena didelė išskirtis radikaliai padidino konstantą ir pakeitė daugiklį prie *K33_4.*

3.2. Stabilizuotų liekamųjų paklaidų regresija

3.2.1. Modelio paskirtis

Stabilizuotų liekamųjų paklaidų regresija yra skirta atvejams, kai nestipriai pažeista homoskedastiškumo prielaida. Tai ta pati tiesinė regresija, tik skaičiuojant paklaidas, naudojamasi specialiu algoritmu, užtikrinančiu skaičiavimų nejautrumą heteroskedastiškumui. Kadangi tokių algoritmų yra ne vienas, tai ir stabilizuotų liekanų regresijų yra keletas. Visas jas vienija tai, kad pokyčiai vyksta tik vertinant liekamąsias paklaidas. Statistikoje nejautrios prielaidų pažeidimams procedūros vadinamos *robastiškomis (atspariosiomis)*. Stabilizuotų likamųjų paklaidų regresija – tai regresija, kai paklaidų įverčiams taikomi robastiniai skaičiavimo metodai. Mes nevadiname šio modelio atspariąja (robastine) regresija, nes pastarąjį terminą naudosime, kai modelyje robastiškai bus vertinamos ne tik paklaidos, bet ir modelio koeficientai, žr. kitą skyrelį.

Stabilizuotų liekamųjų paklaidų regresijose:

- Modelio koeficientai išlieka tokie patys, kaip tiesinėje regresijoje.
- Kitaip apskaičiuojami pasikliautinieji intervalai.
- Kitaip įvertinamos standartinės koeficientų paklaidos – dėl to keičiasi t testų rezultatai.

Dažniausiai naudojamos regresijos, kai paklaidos stabilizuojamos, tranformuojant duomenų matricą. Tradiciškai tokios transformacijos vadinamos HC1, HC2 ir HC3 tipo korekcijomis. Kai kurie autoriai (Long ir Ervin) atlikę gausybę pavyzdžių simuliacijų teigia, kad nedidelėms imtims (iki 250 stebėjimų) geriausia naudoti HC3 korekciją. Didelėms imtims geriau tinka kitos korekcijos.

Nedidelėms imtims ($n \leq 250$) rekomenduojama taikyti HC3 korekciją.

Kitas populiarus būdas gauti stabilesnius liekamųjų paklaidų įverčius – naudoti saviranką (angl. bootstrap). Savirankos idėją galima nusakyti taip: į imtį pažvelkime, kaip į mini populiaciją. Daug kartų iš jos atrinkę mini imtis, turėtume gauti objektyvesnius paklaidų įverčius. Savirankos metodas reikalauja daug labai intensyvių skaičiavimų, nes pats procesas yra iteracinis. Liekamųjų paklaidų įverčių lyginami po kiekvieno žingsnio. Jeigu jie nesistabilizuoja, tai atsakymo galime tiesiog negauti. Nuo visų tų milijononių skaičiavimų gali ir kompiuteris „pasimirti“. Todėl regresiją su saviranka, rekomenduotume taikyti tik nedidelėms imtims. Nereikia nustebti, kad pritaikius tą pačią saviranką tiems patiems duomenims, gauname truputį kitokius rezultatus. Juk tos naujosios mini imtys yra atsitiktinės. Mes tik nurodome jų skaičių (vadinamąjį replikacijų skaičių).

Mažoms imtims galima taikyti regresiją su savirankos būdu įvertintomis liekanomis.

3.2.2. Liekamųjų paklaidų HC korekcijos su STATA

IBM SPSS 19 programoje stabilizuotų liekamųjų paklaidų regresijos nėra. STATA 10 programoje HC3 ir HC1 tipo korekcijos gaunamos, įvykdžius komandas

```
reg priklausomas regresoriai, robust
```

```
reg priklausomas regresoriai, hc3
```

Visų pirma, patikriname, ar korekcijos pakeičia įverčius, kai tiesinės regresijos prielaidos tenkinamos. Pasirenkame A duomenis. Įvykdome

```
reg satisfaction trust_all, robust
```

satisfaction	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
trust_all	.8923218	.0471038	18.94	0.000	.7995301	.9851135
_cons	.38224	.2306284	1.66	0.099	-.072084	.836564

Kaip jau minėta, patys koeficientų įverčiai nesiskiria. Pasikeičia koeficientų standartinės paklaidos (angl, standard error), todėl keičiasi ir t testų reikšmės. Naujosios t testų *p* reikšmės (0,00 ir 0,099) labai nedaug skiriasi nuo buvusių (0,00 ir 0,102, žr. 3.1.3 skyrelį). Nedaug pasikeitė ir koeficientų pasikliautinieji intervalai. Pavyzdžiui, regresoriaus *trust_all* koeficiento 95% pasikliautinasis intervalas dabar yra [0,799; 0,985], o buvo [0,791; 0,994]. Taigi, kaip ir reikėjo tikėtis, jeigu duomenys gerai tinka tiesinei regresijai, tai ir korekcijos nelabai ką pakeičia (pagadina). Analogiškas išvadas padarytume, ištyrę modelį su HC3 pataisa.

Pritaikę HC3 pataisą B duomenims (t.y. heteroskedastiškiems duomenims), gauname

```
reg satisfaction trust_all, hc3
```

K37suminis	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
K32	.8030617	.161637	4.97	0.000	.4830586	1.123065
K33_4	1.242025	.2577623	4.82	0.000	.7317163	1.752333
_cons	7.412178	1.258441	5.89	0.000	4.920762	9.903593

Palyginę su 3.3.1 skyreliu, įsitikiname, kad nei *p* reikšmės, nei pasikliautiniai intervalai daug nepasikeitė. Taigi, kai heteroskedastiškumas nebuvo labai didelis ir nebuvo išskirčių, tiesinė regresija pakankamai gerai tiko (korekcijos iš esmės nieko nepakeitė).

Pritaikę HC3 pataisą C duomenims (duomenims su išskirtimi), gauname

K37suminis	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
K32	.8470274	.1695127	5.00	0.000	.5114601	1.182595
K33_4	.1478968	1.094139	0.14	0.893	-2.01806	2.313854
_cons	11.99029	4.623618	2.59	0.011	2.837381	21.14321

Palyginę su 3.3.1 skyreliu, įsitikiname, kad regresorius *K33_4* tapo statistiškai nereikšmingas (p = 0,893). Labai padidėjo pasikliautinasis konstantos intervalas: [2,837; 21,143], o buvo [10,758; 13,221]. Nors išskirčių HC3 korekcija ir „nenukenksmino“, modelis į jas šiek tiek sureagavo.

Apibendrinsime empirines šio skyrelio išvadas. Stabilizuotų liekamųjų paklaidų regresija:

- taikytina, kai duomenys yra heteroskedastiški ir nėra didelių išskirčių,
- gali būti taikoma ir, kai liekamosios paklaidos šiek tiek skiriasi nuo normaliųjų,
- netinka jeigu duomenys yra didelių išskirčių. Nors į išskirtis ir reaguojama, bet modelis „neišgydomas“.

3.2.3. Savirankos regresija su STATA

Pritaikysime 1000 replikacijų saviranką. Tirsime heteroskedastiškus B duomenis. Įvykdome

```
reg  K37suminis  K32 K33_4, vce(bootstrap, reps(1000))
```

K37suminis	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
K32	.8030617	.1579608	5.08	0.000	.4934642	1.112659
K33_4	1.242025	.2735865	4.54	0.000	.7058052	1.778244
_cons	7.412178	1.340443	5.53	0.000	4.784957	10.0394

Palyginę su 3.1.3 skyreliu, įsitikiname, kad visi pasikliautinieji intervalai praktiškai nepasikeitė. Pritaikę tą pačią saviranką C tipo duomenims su išskirtimis, gauname

K37suminis	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
K32	.8470274	.1600612	5.29	0.000	.5333133	1.160742
K33_4	.1478968	.5629399	0.26	0.793	-.9554452	1.251239
_cons	11.99029	2.45251	4.89	0.000	7.183463	16.79712

Palyginę su 3.3.1 skyreliu, galime padaryti tas pačias išvadas, kaip ir modeliui su korekcija HC3: regresorius *K33_4* tapo statistiškai nereikšmingas (p = 0,793). Kiek padidėjo pasikliautinasis konstantos intervalas: [7,183; 16,797], o buvo [10,758; 13,221]. Taigi, nors išskirčių saviranka ir „nenukenksmino“, į jas buvo sureaguota. Vis dėlto tokį modelį laikyti geru negalima.

Konstruojant pasikliautinius intervalus, pagal nutylėjimą tariama, kad duomenys normalūs Kadangi normalumo prielaida buvo pažeista, pažiūrime, kaip pasikeičia pasikliautinieji intervalai, kai jie skaičiuojami atsižvelgus į savirankos rezultatus. Įvykdome

```
estat bootstrap, all
```

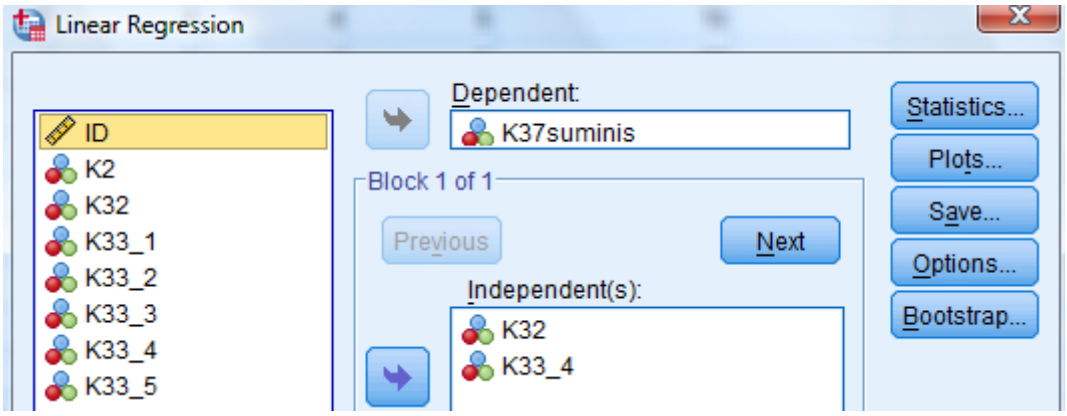
```
Linear regression                               Number of obs      =       125
                                                Replications       =      1000

------------------------------------------------------------------------------
             |   Observed               Bootstrap
  K37suminis |      Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
         K32 |  .84702744  -.0071563   .16006116    .5333133   1.160742   (N)
             |                                      .5291948   1.170506   (P)
             |                                      .5457077   1.200455  (BC)
       K33_4 |  .14789676   .4180353    .5629399   -.9554452   1.251239   (N)
             |                                      .1368578   1.595829   (P)
             |                                      .1358709   1.505192  (BC)
       _cons |  11.990293  -1.794822   2.4525098    7.183463   16.79712   (N)
             |                                      5.452297   13.05797   (P)
             |                                      7.203874   13.76336  (BC)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
```

Matome, kad išskirtys turėjo įtakos ir pasikliautiniams intervalams. Jeigu nuspręstume pateikti pasikliautinuosius intervalus, tai aprašymuose šiuo atveju reikėtų remtis (BC) pasikliautiniais intervalais.

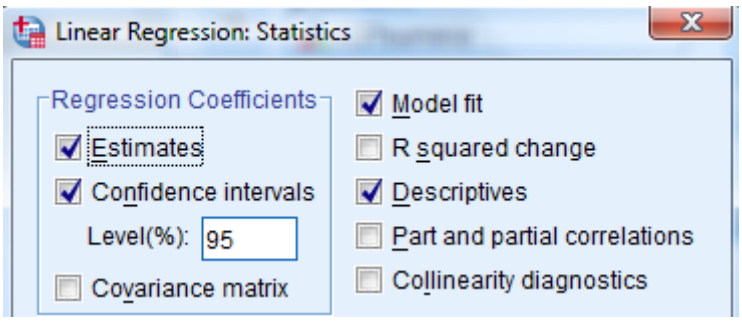
Gavome tas pačias išvadas, kaip ir ankstesniame skyrelyje: stabilizuotų liekamųjų paklaidų regresija su savirankos opcija tinka (bet nėra prasminga) A, tinka ir naudotina B duomenims, nėra naudinga ir netaikytina C duomenims.

3.2.4. Savirankos regresija su SPSS

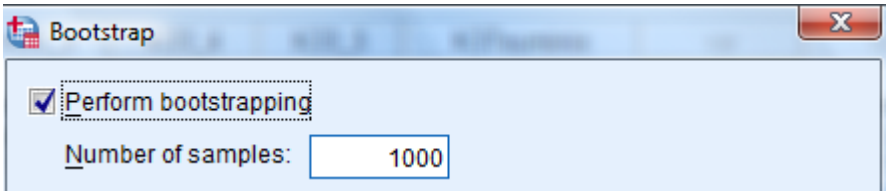
Tirsime B duomenis. Renkamės **Analyze** → **Regression** → **Linear.** Perkeliame *K37suminis* į laukelį **Dependent**, regresorius *K32, K33_4* į laukelį **Independent(s)**.



Pasirenkame **Statistics** ir pažymime **Confidence intervals**, **Descriptives.**



Spaudžiame **Continue** ir **Bootstrap**. Pažymime **Perform bootstrapping**. Šiame meniu (**Number of samples**) galima nurodyti ir kitą replikacijų skaičių.



Spaudžiame **Continue** ir **OK.** Lentelėje **Descriptive Statistics** pateiktos *populiacijos kintamųjų* (o ne jų koeficientų regresijos modelyje) vidutinės reikšmės ir jų 95 % pasikliautinieji intervalai. Turint vieną imtį, kintamojo vidurkiui pasikliautinį intervalą surasti neįmanoma. Tai kas pasikeitė? Taikant saviranką, mini imčių buvo 1000. Žinoma, labai pasitikėti šiais įverčiais nereikia.

Descriptive Statistics

			Bootstrap[a]			
					95% Confidence Interval	
		Statistic	Bias	Std. Error	Lower	Upper
K37suminis	Mean	15.59	.01	.25	15.09	16.06
	Std. Deviation	2.762	-.012	.167	2.404	3.076
	N	124	0	0	124	124
K32 Asmeninės vidutinės pajamos per mėnesį	Mean	3.48	.00	.12	3.25	3.73
	Std. Deviation	1.346	-.013	.120	1.091	1.570
	N	124	0	0	124	124
K33_4 Asmeninių savybių įtaka pasiekiant esamą padėtį darbo rinkoje	Mean	4.33	.00	.07	4.19	4.47
	Std. Deviation	.783	-.008	.083	.618	.937
	N	124	0	0	124	124

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Kitoje lentelėje pateikiami kintamųjų koreliacijų įverčiai ir tų įverčių 95 % pasikliautinieji intervalai. Visos lentelės į šiuos konspektus netalpinsime, pateiksime tik jos fragmentus kintamojo *K37suminis* ir regresoriaus *K32* koreliacijai.

Correlations

Pearson Correlation		K37suminis	K37suminis	1.000
			K32	.419
Sig. (1-tailed)		K37suminis	K37suminis	.
			K32	.000
95% Confidence Interval	Lower	K37suminis	K37suminis	1.000
			K32	.263
	Upper	K37suminis	K37suminis	1.000
			K32	.553

Tiesinės regresijos koeficientai ir jų pasikliautinieji intervalai yra lentelėje **Coefficients**.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	7.412	1.265		5.858	.000	4.907	9.917
	K32	.803	.157	.392	5.128	.000	.493	1.113
	K33_4	1.242	.269	.352	4.613	.000	.709	1.775

a. Dependent Variable: K37suminis

Savirankos metodu įvertinti modelio koeficientų 95 % pasikliautinieji intervalai ir t testų *p* reikšmės yra lentelėje **Bootstrap for Coefficients**. Matome, kad ir pasikliautinieji intervalai ir *p* reikšmės nedaug skiriasi nuo apskaičiuotųjų, naudojant tiesinę regresiją.

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
						95% Confidence Interval	
			Bias	Std. Error	Sig. (2-tailed)	Lower	Upper
1	(Constant)	7.412	-.014	1.248	.001	5.067	9.811
	K32	.803	.009	.157	.001	.495	1.117
	K33_4	1.242	-.002	.259	.001	.710	1.721

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

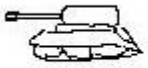
Skyrelio pabaigai pateiksime šiokias tokias rekomendacijas:

- Jeigu tiesinėje regresijoje susidūrėme su heteroskedastiškumu, bet stabilizuotų liekamųjų paklaidų regresija nustatė tuos pačius statistiškai reikšmingus regresorius ir labai panašius jų pasikliautinuosius intervalus, tai heteroskedastiškumas nebuvo labai kenksmingas. Aprašyme viską ir nurodome – ir heteroskedastiškumą paminime, ir stabilizuotų liekamųjų paklaidų regresiją su jos rezultatais.
- Nėra labai svarbu, kokį stabilizavimo metodą taikyti. Pavyzdžiui, kai kurie žmonės itin mėgsta saviranką. Kiti – nemėgsta.
- Jeigu pasikliautinieji intervalai, lyginant su tiesinės regresijos modeliu, labai pakito arba, jeigu yra didelių išskirčių, tai vertėtų išbandyti atspariąją regresiją, žr. kitą skyrelį.

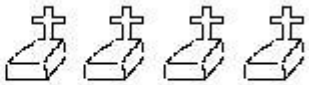
3.3. Atsparioji regresija

3.3.1. Modelio paskirtis

Eiliniam žmogui (ne statistikui) terminas *atsparioji regresija* primena Lietuvos ekonomikos būklę opozicijos kalbose – vyksta regresas ir sustabdyti jį niekaip neišeina. Susidūrusiam su statistika, atrodo, kaip tik priešingai – tai kažkokia regresijų regresija. Tvirta, kaip tankas.



Atsparioji regresija (viduje)



Neatspariosios regresijos (viduje)

Iš tikrųjų, situacija kiek kitokia. Visų pirma, atsparioji (robastinė, angl. robust) regresija atspari tik išskirtims. Stebėjimams suteikiami svoriai, į kuriuos atsižvelgiama, skaičiuojant koeficientų įverčius. Jeigu vienas stebėjimas labai išsiskiria iš kitų, tai jo svoris būna artimas nuliui. Kitaip tariant, išskirtys nedaro didelės įtakos koeficientų įverčiams.

Matematiškai žiūrint, viskas gražu. O idėjiškai? Gerai, jeigu tokia išskirtis buvo atsitiktinė, o jeigu – ne? Įsivaizduokime ekonomistą, sudariusį Lietuvos ekonomikos vystymosi modelį ir samprotaujantį: „Viskas modelyje gerai, tik per 2009 m krizę rodikliai labai suprastėjo. Tai aš tų metų duomenis išmečiau.“ Visi didžiai piktintųsi tokiu ekonomistu ir aiškintų, kad jo modeliu tikėti negalima. O dabar, tarkime, kad ekonomistas pareiškia: „Modelyje buvo išskirčių. Todėl modeliuodami, taikėme atspariąją regresiją“. Klausytojai žavėtųsi ekonomisto erudicija (Vienok, specialistas!) . O modelis, tai praktiškai tas pats.

Atsparioji regresija – tai optimistinis požiūris į duomenis. Ignoruokime išskirtis, ir gal jos gyvenime nepasitaikys. Kartais toks požiūris pasiteisina.

Įtikinti klausytojus sprendimo protingumu yra tyrėjo uždavinys. Jis pats turi tikėti, kad išskirtys duomenyse yra grynas atsitiktinumas. Pavyzdžiui, tyrėjas turi informaciją apie modelius Estijos, Latvijos, Lenkijos duomenims ir yra tikras, kad analogiškas modelis (gal su kiek kitais koeficientų įverčiais) turi būti teisingas ir Lietuvai.

> Atsparioji regresija nukenksmina išskirtis duomenyse.
>
> Tinka tik tada, kai išskirtys atsitiktinės – nežinome kodėl jos atsirado.
>
> Nėra jokios garantijos, kad taip gautas modelis tiksliau aprašo realius reiškinius.

Taikydami atspariąją regresiją, gausime ANOVA statistikos reikšmę ir jos *p* reikšmę, patikrinsime t kriterijumi koeficientų statistinį reikšmingumą, rasime keoficientų įverčius ir jų 95 % pasikliautinuosius intervalus. Šie statistiniai rodikliai interpretuojami visiškai taip pat, kaip it tiesinėje regresijoje. Atspariojoje regresijoje nėra determinacijos koeficiento R^2 analogo. Tai bene svarbiausia kliūtis visuotinam jos išplitimui. Kol kas atsparioji regresija nėra realizuota SPSS programoje.

3.3.2. Atsparioji regresija su STATA

Visų pirma, įsitikinsime, kad duomenims be išskirčių, atsparioji regresija duoda praktiškai tuos pačius koeficientų įverčius, kaip ir tiesinė regresija. Tirsime A duomenis. Įvykdome

```
rreg satisfaction trust_all
```

```
Robust regression                                      Number of obs =     241
                                                       F(  1,   239) =  303.38
                                                       Prob > F      =  0.0000

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   trust_all |   .9176449   .0526847    17.42   0.000     .8138592    1.021431
       _cons |   .2405901   .2381669     1.01   0.313    -.2285843    .7097646
------------------------------------------------------------------------------
```

Palyginę su 3.3.1 skyrelyje pateiktu tiesinės regresijos modeliu, matome, kad regresoriaus *trust_all* koeficiento įvertis (0.917) nedaug skiriasi[11] nuo tiesinės regresijos įverčio (0,892). Nedaug keitėsi ir atitinkamas pasikliautinis intervalas. Konstantos įvertis skiriasi kiek daugiau, bet abiejuose

[11] Žinoma, smagu būtų palyginti, ar šie koeficientai skiriasi statistiškai reikšmingai. Vis dėlto, atspariąjai regresijai to padaryti negalima. O štai dviejų tiesinės regresijos modelių koeficientus palyginti galima, nors tai ir sukelia papildomų problemų, žr. http://www.ats.ucla.edu/stat/stata/faq/compreg2.htm

modeliuose konstantos buvo statistiškai nereikšmingos. Taigi, kaip ir reikia tikėtis, jeigu duomenys tinka tiesinei regresijai, tai atsparioji regresija nieko nepagadina.

Analogiškai ištyrę heteroskedastiškus B duomenis, irgi gauname labai panašius įverčius, kaip ir tiesinėje regresijoje. Matome, kad atsparioji regresija gerai susitvarko („pagydo“ modelį), kai duomenyse yra nedidelis heteroskedastiškumas.

K37suminis	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
K32	.7990511	.1646414	4.85	0.000	.4731	1.125002
K33_4	1.263801	.2830408	4.47	0.000	.7034469	1.824155
_cons	7.352299	1.330316	5.53	0.000	4.718587	9.986012

Dabar ištirsime C duomenis su didele išskirtimi. Sudarę atspariosios regresijos modelį, gauname

K37suminis	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
K32	.7990511	.1646414	4.85	0.000	.4731	1.125002
K33_4	1.263801	.2830408	4.47	0.000	.7034469	1.824155
_cons	7.352299	1.330316	5.53	0.000	4.718587	9.986012

Matome, kad dirbtinai įvestoji išskirtis buvo nukenksminta – modelio parametrų įverčiai liko tokie patys, kaip ir B atveju. Taigi šiuo atveju, atsparioji regresija tinka daug geriau, nei klasikinė tiesinė regresija.

Kada taikoma atsparioji regresija, o kada stabilizuotų liekamųjų paklaidų regresija? Jeigu tik galime, tai taikome liekamųjų paklaidų regresiją, žr. ankstesnio skyrelio pabaigą. Pavyzdžiui, B duomenims. Atsparioji regresija taikoma tik iš didelės bėdos. Tipiškai ji taikoma, jeigu:

- tiesinės regresijos modelyje nustatėme išskirtį (kažkuriam stebėjimui buvo Kuko matas didelis),
- nesugalvojame, kodėl išskirtį būtų galima iš duomenų pašalinti,
- esame šventai įsitikinę, kad išskirtis duomenyse – tai grynas atsitiktinumas,
- manome, kad tuo pavyks įtikinti klausytojus (recenzentus, dėstytojus, redkolegiją),
- duomenys mums atrodo pakankamai normalūs.

Tolesni veiksmai priklauso nuo to, ar gautojo modelio koeficientų įverčiai panašūs į gautuosius tiesinės regresijos modelyje. Jeigu taip – tai, lengviau atsipūtę, toliau nagrinėjame tiesinę regresiją, tik nurodydami, kad išskirtis nebuvo labai kenksminga ir atsparioji regresija parodė, kad modelis beveik nepasikeitė. Jeigu ne – tai pateikiame atspariosios regresijos modelio keoficientus su

apgailestavimu, kad nebeliko determinacijos koeficiento reikšmės ir iš viso nebeliko bent kiek geresnio modelio tikimo įrodymo.

Jeigu duomenyse yra išskirčių, o mes švento įsitikinimo jų atsitiktinumu neturime, tai tiesog jokio tinkamo regresijos modelio sudaryti nepavyks. Yra duomenų, kuriems jokie regresijos modeliai netinka.

3.4. Medianos regresija

3.4.1. Modelio paskirtis

Medianos regresijos idėja – modeliuoti ne *vidutinę* reikšmę, o *medianos* reikšmę. Todėl nereikia tikėtis, kad parametrų įverčiai bus labai panašūs į gautuosius, taikant tiesinę regresiją.

Priminsime, kad tiesinės regresijos modelis siejamas su reikšmių vidurkiu. Net ir prognozuojama *vidutinė* priklausomo kintamoje reikšmė. Tik tas ne visada pabrėžiama. Tarkime, kad regresijos modelį naudojame gydymo trukmės prognozei. Įstatę konkretaus ligonio tyrimų rezultatus į modelį, mes iš tikro prognozuojame, kiek *vidutiniškai* laiko gyja ligoniai su tokiais tyrimų rezultatais. Tokią interpretaciją (pvz., vidutiniškai gydymas užtrunka 3 savaites) ligonis be vargo supranta.

Medianos regresijos interpretacija – žymiai bjauresnė. Išklausęs litaniją – jeigu visus galimus pagijimo laikus išrikiuosime nuo trumpiausio iki ilgiausio, tai viduryje bus 3 savaitės, – ligonis nuspręs, kad „kažko čia tas gydytojas išsisukinėja“. Laikas testamentui, nuoširdžiai asmeninei dovanai, kito gydytojo paieškoms, ar pan.

Kada taikoma medianos regresija? Dažniausiai ji rekomenduojama, kai nepatenkinta normalumo prielaida. Heteroskedastiškumas jai didelės įtakos neturi.

Mediana nejautri išskirtims. Atrodytų, kad ir medianos regresija turėtų būti nejautri išskirtims. Bet taip nėra. Mat regresijos koeficientai optimizuojami, parenkant juos absoliutiniu didumu kuo arčiau medianos, o ne tiesiog ignoruojant labai dideles reikšmes.

Medianos regresija naudotina, kai negalioja normalumo prielaida.
Gautieji modelio parametrų įverčiai nesutampa su gautais, taikant tiesinę regresiją.
Interpretacija nesutampa su tiesinės regresijos interpretacija.
Jeigu duomenyse yra didelių išskirčių, medianos regresija netinka.

Statistinių rodiklių medianos regresijoje yra labai nedaug. Pseudo R^2 tėra gana netikęs determinacijos koeficiento analogas. Visas tyrimas susiveda į parametrų įverčius ir jų pasikliautinius intervalus. Norint, galima standartinių paklaidų skaičiavimui pritaikyti saviranką.

Medianos regresija tėra atskiras (populiariausias) *kvantilių regresijos* atvejis. Galima sudaryti modelį respondentams, kuriems įgyjamos 10% didžiausių priklausomo kintamojo reikšmių ir t.t.

Kvantilių regresija dar vadinama mažiausio absoliutinio didumo (angl. LAV), minimalaus absoliutaus nuokrypio (angl. MAD) arba L1 – normos modeliu.

Kol kas medianos (kvantilių) regresija nėra realizuota SPSS programoje. Todėl aptariame tik STATA paketą.

3.4.2. Medianos regresija su STATA

Visų pirma, išbandysime, kaip medianos regresija atrodo duomenims, kurie tenkina tiesinės regresijos prielaidas. Jiems taikyti medianos regresiją tikrai galima.Tiriame A duomenis. Įvykdome:

```
qreg satisfaction trust_all
```

```
Median regression                                      Number of obs =       241
  Raw sum of deviations    444.5 (about 4)
  Min sum of deviations 286.3718                       Pseudo R2     =    0.3557

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   trust_all |   .9615385   .0522533    18.40   0.000     .8586026    1.064474
       _cons |   2.95e-08   .2331381     0.00   1.000     -.459268     .459268
------------------------------------------------------------------------------
```

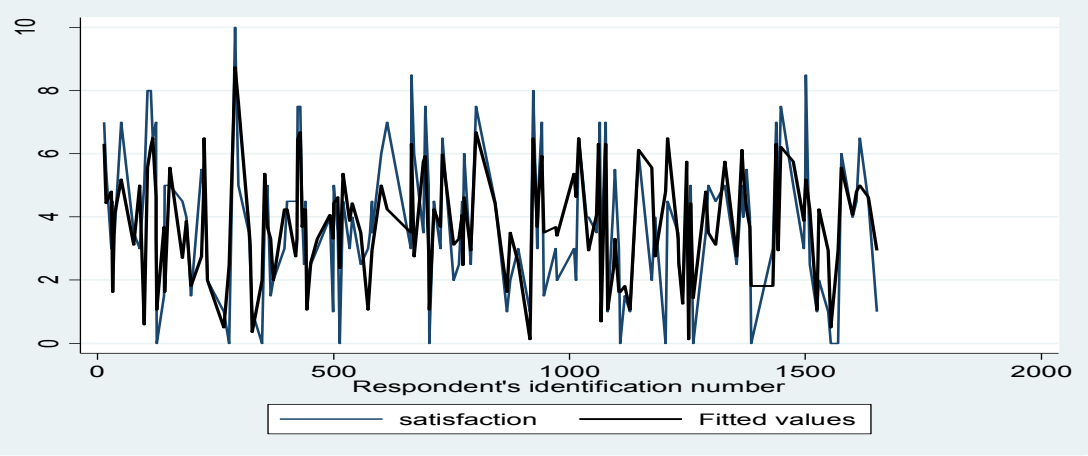
Palyginę su 3.3.1 skyrelyje pateiktu tiesinės regresijos modeliu, matome, kad regresoriaus *trust_all* koeficiento įvertis (0.961) nedaug skiriasi nuo tiesinės regresijos įverčio (0,892). Nedaug keitėsi ir atitinkamas pasikliautinis intervalas. Konstantos įvertis skiriasi radikaliai (2.95e-08, tai užrašas, kuris reiškia, kad 2,95 reikia padalinti iš 10^8. Gausime praktiškai 0, todėl ir atitinkama *p* reikšmė lygi 1). Žodžiu, rekomendacija pašalinti iš modelio konstantą, šįkart labai primygtina. Pseudo R^2 = 0,355 yra priimtino didumo, nors gerokai mažesnis, nei buvo (0,557).

Modelio gerumą visada verta patikrinti ir grafiškai. Sukuriame prognozuojamų *satisfaction* reikšmių stulpelį duomenyse (pavadiname jį *PRED*) įvykdydami komandą:

```
predict PRED
```

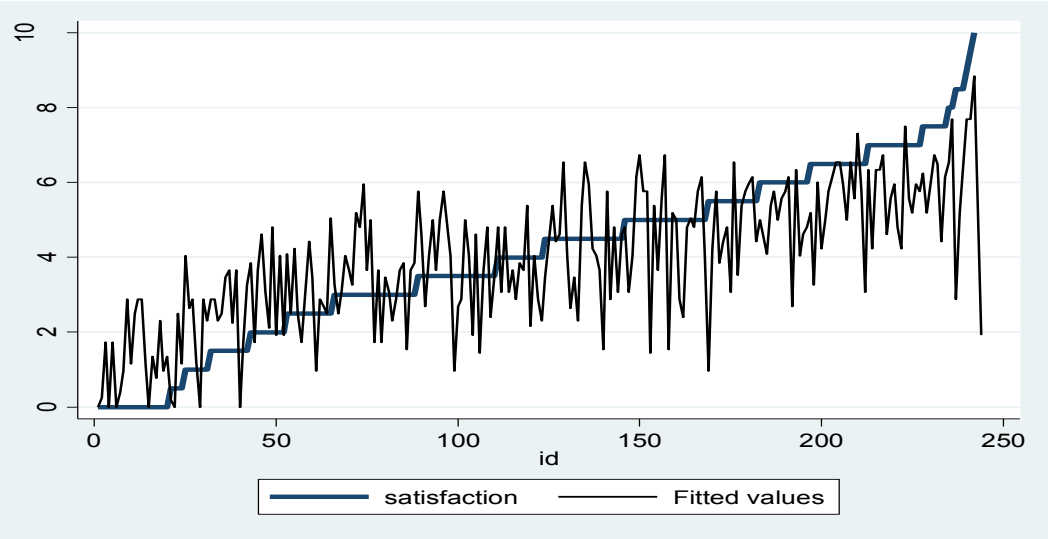
Braižydami grafiką, pasinaudosime duomenyse esančiu respondento identifikaciniu numeriu *idno.* Įvykdome:

```
graph twoway line  satisfaction PRED  idno
```



Prognozuojamų ir tikrųjų kintamojo *satisfaction* reikšmių grafikai skiriasi nedaug. Matome, kad ir medianos regresijos modelis gerai aprašo duomenų elgesį. Žinoma, galima ir kitaip pateikti duomenų tikimo grafiką. Pavyzdžiui, išdėliosime visas reikšmes pagal *satisfaction* augimą. Įvykdome komandas:

```
sort satisfaction
gen id = _n
graph twoway line  satisfaction PRED  id
```



Sumodeliuokime 75% kvantilio elgesį. Apytikslė tokio modelio interpretacija skambėtų taip: „modelis, aprašantis respondentus, kurie pagal pasitenkinimą politika yra tarp ketvirtadalio labiausiai patenkintų“ . Ši interpretacija yra tik apytikslė (bet lengviau suprantama), nei tikslesnis pasakymas, kad modelis sudarytas respondentams, už kurios tiksliai 75% respondentų yra mažiau patenkinti, o 25 % labiau patenkinti. Įvykdome

```
qreg satisfaction trust_all, q(0.75)
```

```
.75 Quantile regression                                   Number of obs =       241
  Raw sum of deviations    344.5 (about 5.5)
  Min sum of deviations 232.6399                          Pseudo R2     =    0.3247

------------------------------------------------------------------------------
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   trust_all |   .8928571   .0720372    12.39   0.000     .7509483    1.034766
       _cons |   1.392857   .3213599     4.33   0.000     .7597977    2.025917
------------------------------------------------------------------------------
```

Pasikeitė ne tik konstantos įvertis (daugiklis prie *trust_all* liko panašus į tiesinės regresijos modelio daugiklį), bet ir pseudo R^2. Analogiškai galima suskaičiuoti ir kitų kvantilių regresijas. Kada yra prasmė skaičiuoti kvantilio regresiją? Kai norime iš visos respondentų grupės pasirinkti dalį, o duomenis panaudoti visus. Pavyzdžiui, sumodeliuoti priklausomybę labiausiai patenkintiems politika respondentams. Tada ir interpretuosime, kad pasitenkinimas politika nuo *trust_all* respondentams, kurie yra tarp labiausiai politika patenkintųjų, priklauso štai taip:

satisfaction = 1,39 + 0,89 *trust_all.*

Analogiškai galima sumodeliuoti priklausomybę 25% labiausiai nepatenkintų politika respondentų. STATA yra numatyta galimybė pateikti kelių kvantilių regresijos modelius iškart. Įvykdome

```
sqreg satisfaction trust_all, q(0.25, 0.5, 0.75)
```

```
Simultaneous quantile regression                          Number of obs =       241
  bootstrap(20) SEs                                       .25 Pseudo R2 =    0.3695
                                                          .50 Pseudo R2 =    0.3557
                                                          .75 Pseudo R2 =    0.3247

------------------------------------------------------------------------------
             |               Bootstrap
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q25          |
   trust_all |      .9375   .0769749    12.18   0.000      .785864    1.089136
       _cons |      -.875   .2759415    -3.17   0.002    -1.418588   -.3314119
-------------+----------------------------------------------------------------
q50          |
   trust_all |   .9615385   .0490163    19.62   0.000     .8649793    1.058098
       _cons |   2.95e-08   .1892589     0.00   1.000    -.3728284    .3728285
-------------+----------------------------------------------------------------
q75          |
   trust_all |   .8928571   .0776942    11.49   0.000     .7398042     1.04591
       _cons |   1.392857   .4082513     3.41   0.001     .5886269    2.197087
------------------------------------------------------------------------------
```

Liekamosios paklaidos, vykdant šią komandą, buvo vertinamos savirankos metodu. Matome, kad koeficientai prie *trust_all* daug nesiskiria, o jų pasikliautinieji intervalai labai persidengia. Tai

leidžia įtarti, kad modeliuose reikšmingai skiriasi tik konstantos. Parodysime, kad 75% ir 25% kvantilių regresijų koeficientai prie *trust_all* statistiškai reikšmingai nesiskiria. Įvykdome:

```
test[q25]trust_all = [q75]trust_all
```

```
( 1)  [q25]trust_all - [q75]trust_all = 0

      F(  1,   239) =    0.17
           Prob > F =    0.6817
```

Matome, kad $p = 0{,}68 > 0{,}05$. Analogiškai įsitikiname, kad statistiškai reikšmingai nesiskiria ir medianos regresijos modelio koeficientas prie *trust_all.* Konstantas palyginame, įvykdydami:

```
test[q25]_cons = [q75]_cons
```

```
( 1)  [q25]_cons - [q75]_cons = 0

      F(  1,   239) =   30.69
           Prob > F =    0.0000
```

Matome, kad $p = 0{,}00 < 0{,}05$. Taigi modelių konstantos skiraisi statistiškai reikšmingai.

Tyrimo prasme šie rezultatai nėra labai įdomūs. Kur kas įdomiau būtų, jeigu būtų pavykę nustatyti, kad labiau politika patenkintiems *trust_all* įtaka yra silpnesnė, nei politika nepatenkintiems. Taigi, tirtiems duomenims 75% arba 25% kvantilių regresijų išskyrimas ir aprašymas nėra prasmingas.

Be modelio konkrečiam kvantiliui, galima sumodeliuoti ir skirtumą tarp pirmojo ir trečiojo kvartilių (t.y. skirtumą tarp 0,75 ir 0,25 kvantilių). Toks modelis gali turėti prasmę, jeigu norime išryškinti modelių skirtumus tarp palankiai ir nelabai palankiai vertinančius politinę padėtį. Įvykdome

```
iqreg  satisfaction trust_all, q(0.25, 0.75)
```

```
.75-.25 Interquantile regression                    Number of obs =       241
  bootstrap(20) SEs                                 .75 Pseudo R2 =    0.3247
                                                    .25 Pseudo R2 =    0.3695

------------------------------------------------------------------------------
             |               Bootstrap
satisfaction |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   trust_all |  -.0446429   .0872086    -0.51   0.609    -.2164385    .1271528
       _cons |   2.267857    .418498     5.42   0.000     1.443441    3.092273
------------------------------------------------------------------------------
```

Labai politika patenkintų ir mažai patenkintų požiūrio į politiką skirtumas nuo *trust_all* priklauso štai taip:

Skirtumas satisfaction = 2,27 - 0,044 *trust_all.*

Regresorius *trust_all* tapo statistiškai nereikšmingu. Tai dar kartą patvirtina jau anksčiau gautą išvadą, kad priklausomybė nuo *trust_all* tokia pati ir labiau politika patenkintiems ir menkiau patenkintiems respondentams. Kvartilių skirtumo regresiją nagrinėti reikia tik tada, jeigu manome, kad atotrūkis tarp požiūrių kinta. Šiuo atveju tai buvo neaktualu. Apskritai socialiniuose tyrimuose kvartilių skirtumo regresija naudojama retai.

Pritaikysime medianos regresiją B duomenims. Primename, kad jiems neišpildyta homoskedastiškumo prielaida. Įvykdome

```
qreg  K37suminis K32 K33_4
```

```
Median regression                                      Number of obs =       124
  Raw sum of deviations      271 (about 16)
  Min sum of deviations    226.5                       Pseudo R2     =    0.1642

------------------------------------------------------------------------------
  K37suminis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         K32 |          1   .2029565     4.93   0.000      .598194    1.401806
       K33_4 |        1.5   .3295227     4.55   0.000     .8476229    2.152377
       _cons |        5.5   1.560052     3.53   0.001     2.411466    8.588534
------------------------------------------------------------------------------
```

Matome, kad koeficientų įverčiai gana žymiai skiriasi nuo tiesinės regresijos modelio (pvz. konstantos įvertis 5,5 daug mažesnis už 7,412 iš tiesinės regresijos).

Minėjome, kad medianos regresijoje irgi galima taikyti savirankos opciją. Galima įvykdyti 1000 replikacijų saviranką taip:

```
bsqreg  K37suminis K32 K33_4, reps(1000)
```

Vis dėlto, toks metodas skaičiuoti pasikliautiniams intervalams nėra labai geras. Nors standartinės paklaidos ir surandamos savirankos metodu, pasikliautiniai intervalai apskaičiuojami pagal formules, tinkančias *normaliems* duomenims. O pati medianos regresija skirta duomenims, kurie *nėra normalūs.* Todėl geriau ir pačius intervalus konstruoti, atsižvelgiant į saviranką. Įvykdome:

```
bootstrap "qreg  K37suminis K32 K33_4" _b[K32] _b[K33_4]  _b[_cons], reps(1000)
```

Nesunku suprasti, kad „_b[regresorius]“ nurodo, kuriam regresoriui bus konstruojamas pasikliautinis intervalas. Savirankos replikacijų skaičius nurodomas reps(....). Nereikia pasirinkti labai didelio skaičiaus, nes skaičiavimai užims daug laiko. Manoma, kad saviranka su 1000 replikacijų yra visiškai pakankama.

```
command:      qreg K37suminis K32 K33_4
statistics:   _bs_1      = _b[K32]
              _bs_2      = _b[K33_4]
              _bs_3      = _b[_cons]

Bootstrap statistics                              Number of obs    =       124
                                                  Replications     =      1000

------------------------------------------------------------------------------
Variable       |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
---------------+--------------------------------------------------------------
       _bs_1   |  1000         1 -.0610312  .1993424   .6088221   1.391178   (N)
               |                                       .3333333        1.2   (P)
               |                                              1        1.4  (BC)
       _bs_2   |  1000       1.5 -.1910651  .4267538   .6625633   2.337437   (N)
               |                                             .5          2   (P)
               |                                              1        2.6  (BC)
       _bs_3   |  1000       5.5  .9099349  2.074094   1.429919   9.570081   (N)
               |                                              3       10.6   (P)
               |                                              3     8.1875  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

Gavome trimis būdais įvertintus koeficientų pasikliautinuosius intervalus. Didelių skirtumų tarp rezultatų vis dėlto nėra. Dažniausiai rekomenduojama cituoti (BC) metodu gautuosius pasikliautinius intervalus. BC (bias corrected) metodas, pateikia įverčius, kai atsižvelgiama į galimą jų paslinktumą.

Reziumuodami galime konstatuoti, kad heteroskedastiškiems B duomenims medianos regresija yra tinkama alternatyva tiesinei regresijai. Primename, kad B duomenims galima naudoti ir stabilizuotų liekamųjų paklaidų regresiją. Kurį regresijos modelį pasirinkti priklauso nuo tik nuo tyrėjo. Labiau įprasta taikyti tiesinę regresiją ir jos modifikacijas.

Skyrelio pabaigai ištirsime, kaip medianos regresijos modelis tinka C duomenims:

```
Median regression                                   Number of obs =       125
  Raw sum of deviations      285 (about 16)
  Min sum of deviations 238.9255                    Pseudo R2     =    0.1617

------------------------------------------------------------------------------
  K37suminis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         K32 |          1   .1948748     5.13   0.000     .6142258    1.385774
       K33_4 |   .1382979   .0053317    25.94   0.000     .1277433    .1488524
       _cons |   11.30851   .7223108    15.66   0.000     9.878624     12.7384
------------------------------------------------------------------------------
```

Gauname, kad praktiškai visi koeficientai yra panašūs į tiesinės regresijos koeficientus, o ne į B duomenų medianos regresijos koeficientus. C duomenys nuo B duomenų skyrėsi ryškia išskirtimi. Medianos regresija jos „nenukenksmino“.

Medianos regresija nėra tinkama duomenims, kuriuose yra didelių išskirčių.

Jeigu duomenys nėra normalūs ir turi dideles išskirtis, jokio gero regresijos modelio sudaryti nepavyks.

3.5. Netiesinė regresija

3.5.1. Modelio paskirtis

Tiesinė regresija vadinama tiesine būtent todėl, kad aprašo tik tiesines priklausomybes – regresoriai nekeliami laipsniu ir nedauginami tarpusavyje. Gyvenime ne visos kintamųjų priklausomybės yra tiesinės. Štai tada ir taikoma netiesinė regresija. Vis dėlto, netiesinėje regresijoje išvados daromos labai negriežtai (apie statistinį reikšmingumą ir visokius kriterijus faktiškai galima tik pasvajoti), todėl, jeigu tik galima, jos vengiama.

Kai kuriuos modelius tinkamomis kintamųjų transformacijomis galima suvesti į tiesinius. Tada taikoma tiesinė regresija.

Pateiksime keletą pavyzdžių, kada tokios transformacijos įmanomos:

a) Polinominė regresija: $Y = a + b\,X + cX^2 + dX^3$. Tada tiesiog sukuriame naujus kintamuosius $Z = X^2, W = X^3$ ir tiriame modelį $Y = a + b\,X + cZ + dW$.
b) Kobo – Daglo produkcijos funkcija $Q = AK^aL^b$. Čia K (kapitalas) ir L (darbo jėga) yra regresoriai, o *a,b* ir *A* (technologijos daugiklis) nežinomi parametrai, kuriuos reikia įvertinti. Kadangi *Q,A,K* ir *L* yra teigiami dydžiai, tai tiesinį modelį gauname logaritmuodami abi lygybės puses: $\ln Q = \ln A + a \ln K + b \ln L$. Pažymėję $Y = \ln Q$, $C = \ln A,\ X = \ln K$, $Z = \ln L$, gauname tiesinės regresijos modelį $Y = C + a\,X + b\,Z$.
c) Paprasčiausias Verhulst – Pearl populiacijos augimo modelis: $P(t) = P_0e^{kt}$. Nežinomas modelio parametras – *k*. Logaritmuojame abi lygybės puses ir įvedame naujus pažymėjimus: $\ln P\ = \ln P_0\ +\ kt, Y = \ln P\ - \ln P_0$. Gauname modelį $Y = k\,t$.

Ne visada tokios transformacijos įmanomos. Pateiksime keleto tipinių netiesinių modelių pavyzdžius iš SPSS 19 aprašo:

Pavadinimas	**Išraiška**
Asimptotinė regresija	b1 + b2 * exp(b3 * x)
Asimptotinė regresija	b1 – (b2 * (b3 ** x))
Tankumo	(b1 + b2 * x) ** (–1 / b3)
Gauso	b1 * (1 – b3 * exp(–b2 * x ** 2))
Gompertz	b1 * exp(–b2 * exp(–b3 * x))
Grąžų	b1 + b2 * exp(–b3 * x)
Richards	b1 / ((1 + b3 * exp(–b2 * x)) ** (1 / b4))
Veibulo	b1 – b2 * exp(–b3 * x ** b4)
Verhulst	b1 / (1 + b3 * exp(–b2 * x))
Derliaus tankio	(b1 + b2 * x + b3 * x ** 2) ** (–1)

Žinoma, galima susikurti ir visiškai naują nematytą modelį, pavyzdžiui, varanopsų augimo sąlyginai inteligentiškoje aplinkoje modelį (VASIA) .

Netiesinėje regresijoje modelio formą pasirenka pats tyrėjas.

Visos nežinomų parametrų reikšmės vertinamos iteraciniu būdu – jas po truputį keičiant (tą daro programa) ir meldžiantis, kad pasisektų (tą daro tyrėjas). Pradines nežinomų parametrų reikšmes turi (gali) pasirinkti pats tyrėjas. Nuo sėkmingo pasirinkimo labai priklauso galutinio modelio tinkamumas. Kaip žinoti, kokias reikšmes rinktis? Manoma, kad modeliuotojas turi nujausti, kokie tie parametrai turi būti. Kitaip, ko tą modelį rinktis? Maždaug taip: pažiūrėjome į varanopsų svorio augimo kreivę ir matome – kaip iš akies luptas Gompertz modelis su parametrais $1, \sqrt{2}, \pi$. Nuo jų ir pradedame.

Dar viena nemaloni naujiena – visi įverčiai jautrūs išskirtims. Geriau, kai duomenyse jų nėra.

Tikrinant, ar modelis tinka, pasižiūrima į R^2 reikšmę. Jeigu modelis būtų tiesinis, tai R^2 atitiktų determinacijos koeficientą. Dabar tai viso labo determinacijos koeficiento analogas. Gerai duomenims tinkantiems modeliams R^2 nedaug skiariasi nuo vieneto.

Jeigu parametrų koreliacijos didelės, tai *galbūt* modelyje yra per daug parametrų.

Visada rekomenduojama nusibraižyti realių ir prognozuojamų duomenų grafikus ir vizualiai palyginti, ar jie labai skiriasi. Tiesa, jeigu duomenų pakankamai daug, toks grafikas primena kintamo dažnio srovę ir jame sunku ką nors įžiūrėti (bet atrodo įspūdingai).

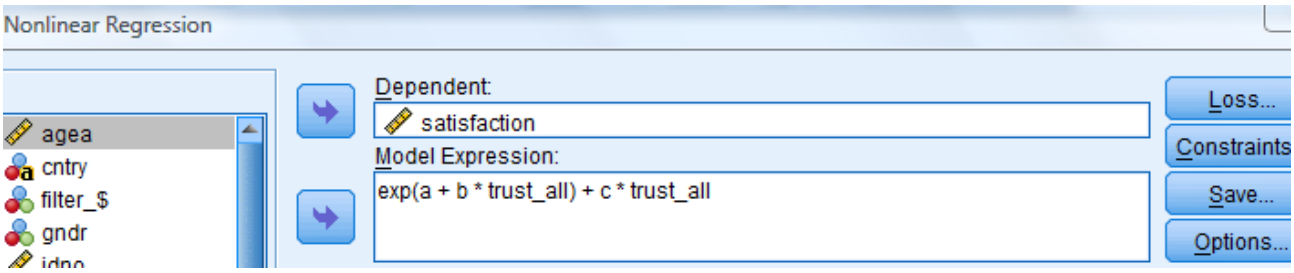
3.5.2. Netiesinė regresija su SPSS

Pasirinksime jau tiesinėje regresijoje nagrinėtą respondentų pasitikėjimo politiniu šalies klimatu *satisfaction* priklausomybės nuo pasitikėjimo svarbiausiomis valstybės institucijomis *trust_all* tyrimą (A duomenys). Tirsime 35 – 40 m Estijos vyrus (agea >= 35 & agea <= 40 & gndr = 1). Pasirenkame tokį modelį su trimis nežinomais parametrais *a, b, c*:

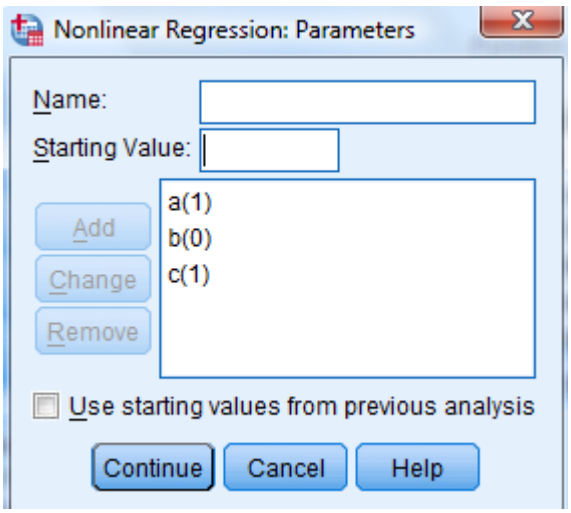
$$satisfaction = e^{a+b\ trust\_all} + c\ trust\_all.$$

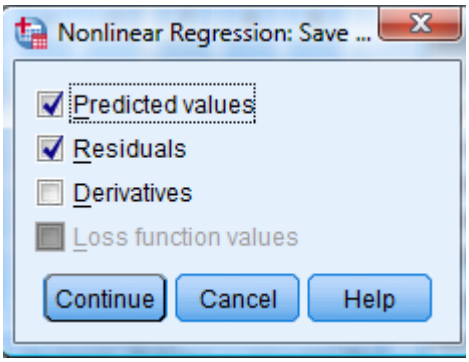
Kodėl pasirinkome tokią modelio išraišką? Mes jau žinome, kad duomenims tinka tiesinė regresija. Pasirinkome modelį, kuris tik truputį bendresnis. Jeigu $b = 0$, tai faktiškai gauname tiesinę regresiją. Šią informaciją išnaudojame, pasirinkdami pradines reikšmes $a = 1$, $b = 0$, $c = 1$.

Renkamės **Analyze → Regression → Nonlinear.** Atsidariusiame lange kintamąjį *satisfaction* įkeliame į laukelį **Dependent**. Langelyje **Model Expression** įrašome modelio išraišką su visomis nežinomomis konstantomis: exp(a+b**trust_all*) +c**trust_all*.



Paspaudę **Parameters** surašome nežinomus parametrus ir jų pradines reikšmes. Renkamės **Continue**. Pasirenkame **Save**. Pažymime **Predicted values** ir **Residuals**. Renkamės **Continue .**





Jeigu norėtume apriboti galimas parametrų reikšmes (pvz., c > 0), tai būtų galima padaryti, pasirinkus **Constraints.**

Rezultatų išklotinė prasideda nuo iteracijų istorijos. Joje matyti, kaip vis kartojant skaičiavimus keitėsi parametrų reikšmės. Procesas nutraukiamas, kai parametrų reikšmės stabilizuojasi. Čia galima sulaukti informacijos apie netikusius parametrų pasirinkimus (neigiamo skaičiaus logaritmavimą ir pan.) . Lentelėje **Parameter Estimates** pateikiami parametrų įverčiai.

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.237	.538	-.837	1.311
b	.199	.030	.139	.260
c	.279	.392	-.504	1.063

Matome, kad b = 0,199 nors ir netoli nulio, bet jam nelygus. Jeigu nuspręstume, kad modelis tinkamas, šiuos įverčius įstatytume į modelio išraišką. Kitoje lentelėje surašyti parametrų koreliacijų įverčiai. Matome, kad c ir a stipriai koreliuoja. Galbūt, reikėtų ištirti modelį be vieno iš šių parametrų.

Correlations of Parameter Estimates

	a	b	c
A	1.000	-.160	-.944
B	-.160	1.000	-.158
C	-.944	-.158	1.000

Kita lentelė **ANOVA** su ANOVA tiesiogiai sietųsi, jeigu tirtume tiesinę priklausomybę. SPSS nepateikia jokių p reikšmių, nes netiesiniame modelyje nelabai šiuo kriterijumi galima tikėti, ypač

nedidelėms imtims. Dažniausiai visus užganėdina R^2 suradimas. Mūsų atveju R^2 = 0,571, pakankamai didelis.

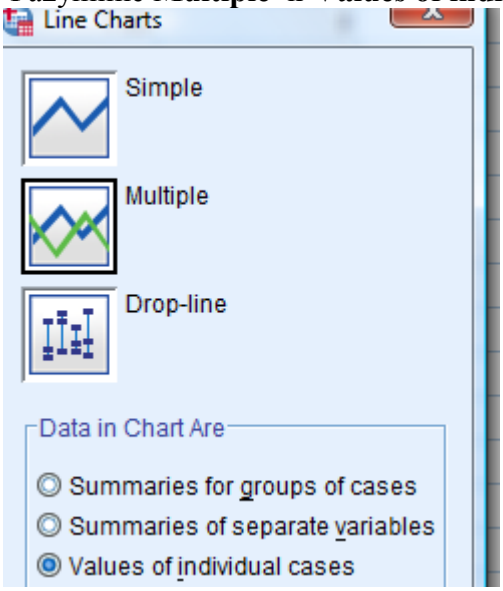
ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	1548.629	3	516.210
Residual	155.121	67	2.315
Uncorrected Total	1703.750	70	
Corrected Total	361.718	69	

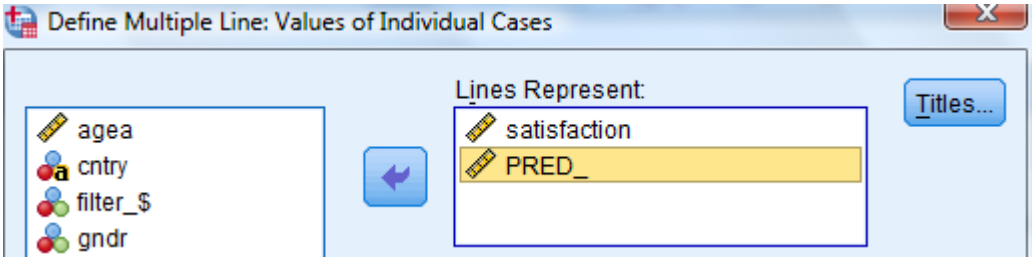
Dependent variable: satisfaction

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .571.

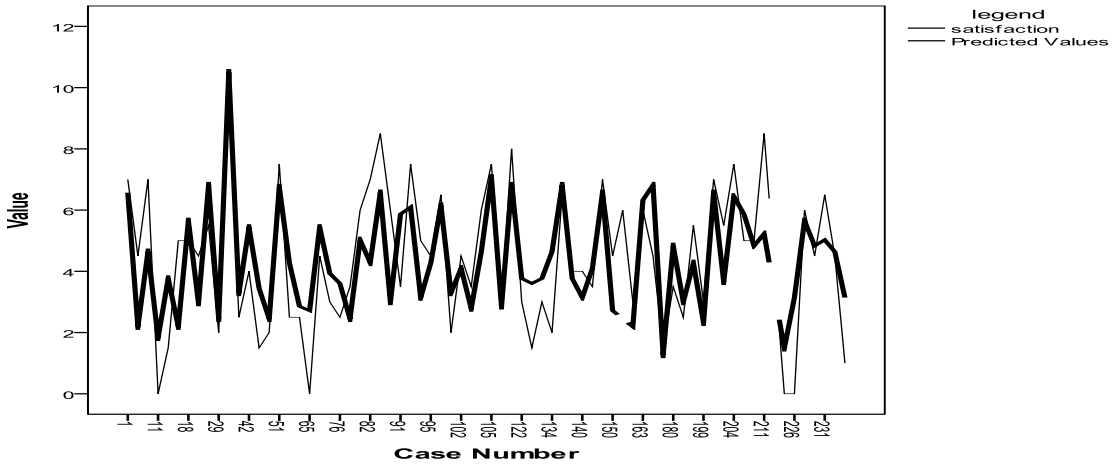
Vis dėlto, aklai pasitikėti *vieninteliu* R^2 negalima. Pabandysime vizualiai palyginti tikras *satisfaction* reikšmes su gautomis, naudojant modelį. Spaudžiame **Graphs → Legacy Dialogs → Line**. Pažymime **Multiple** ir **Values of individual cases.** Spaudžiame **Define.**



Atsidariusiame lange perkeliame *satisfaction* ir *PRED_* į laukelį **Lines Represent**.



Spaudžiame **OK**. Gauname tikrųjų ir progozuotų reikšmių grafikus. Kuo geresnis prognozavimas, tuo labiau šie grafikai sutampa. Paredagavome grafiką taip, kad kreivės būtų skirtingo storio. Matome, kad modelis gana gerai tinka.



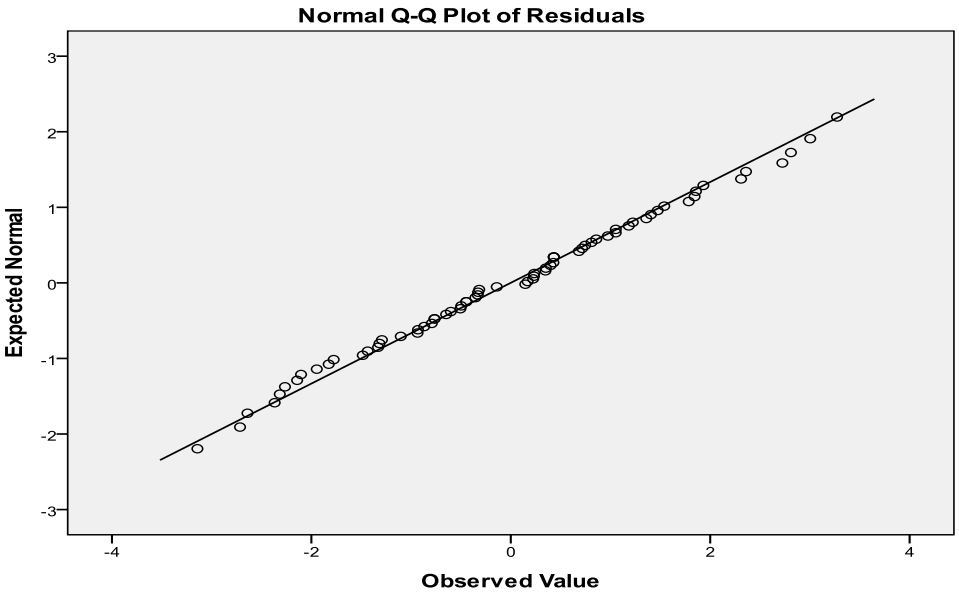
Labai norint įsitikinti liekamųjų paklaidų normalumu (tada labiau galima pasitikėti R^2), galima pritaikyti Šapiro – Vilko testą. Jį aprašėme tiesinės regresijos skyriuje. Surinkę **Analyze → Descriptive Statistics → Explore → Plots**, pažymėję **Normality plots with tests**, į **Dependent List** įkeliame *RESID* ir spaudžiame **OK.** Gautasis Šapiro – Vilko testas rodo, kad liekanas galima laikyti normaliomis ($p = 0{,}792 > 0{,}05$). Tą patį leidžia teigti ir Q-Q grafikas. Tai kvantilių grafikas, kuris vertinamas analogiškai P-P grafikui: kuo taškai arčiau tiesės, tuo liekamosios paklaidos normalesnės.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RESID Residuals	.055	70	.200*	.989	70	.792

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.



Gavome, kad siūlomas modelis pakankamai gerai tinka duomenims. Ar tai reiškia, kad jau galima užmiršti analogišką tiesinės regresijos modelį? Jokiu būdu. Nerašyta taisyklė skelbia, jeigu yra keli daugiau mažiau vienodai gerai tinkantys modeliai, tai naudotinas *paprastesnis.* Reikėtų užmesti akį į tiesinės regresijos prognozių grafiką, ir tik tada apsispręsti.

Jeigu yra keli, vienodai tinkantys modeliai, tai pranašesniu laikomas paprastesnis modelis.

3.5.3. Netiesinė regresija su STATA

Pasirinksime A duomenis ir pakartosime tyrimą, kurį jau aprašėme ankstesniame skyrelyje. Tirsime 35 – 40 m Estijos vyrus (keep if agea >= 35 & agea <= 40 & gndr == 1). Viena keisčiausių STATA kūrėjų užgaidų – praleistas reikšmes (tuščią langelį duomenyse) traktuoti, kaip begalybę. Tas sukelia didžiulių nepatogumų skaičiuojant, jeigu duomenyse tokių „tuštumų“ yra. Apsisaugome nuo keistųjų STATA kūrėjų fantazijų, pasirinkę galimas baigtines kintamųjų reikšmes: Įvykdome:

```
keep if  satisfaction < 30 & trust_all < 30.
```

Tiriamas modelis:

$$satisfaction = e^{a+b\ trust\_all} + c\ trust\_all.$$

Siekiame įvertinti nežinomus parametrus *a, b, c*. Surenkame (visi parametrai **būtinai** rašomi riestiniuose skliaustuose: {a} ir pan.) ir įvykdome:

```
nl ( satisfaction = exp({a}+{b}* trust_all ) + {c}*  trust_all), initial (a 1 b 0 c 1)
```

Source	SS	df	MS
Model	1548.62869	3	516.209564
Residual	155.121307	67	2.31524339
Total	1703.75	70	24.3392857

Number of obs =	70
R-squared =	0.9090
Adj R-squared =	0.9049
Root MSE =	1.521592
Res. dev. =	254.3512

satisfaction	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
/a	.2370719	.5379013	0.44	0.661	-.8365835	1.310727
/b	.1994025	.0304736	6.54	0.000	.1385769	.260228
/c	.2791581	.392484	0.71	0.479	-.5042433	1.06256

Matome, kad R^2 = 0,909, t.y. tiesiog fantastiškas, jeigu tai būtų tiesinė regresija. Beje, ir R^2 ir kitos charakteristikos skiriasi nuo gautųjų SPSS programa. Taip įvyksta todėl, kad naudojami

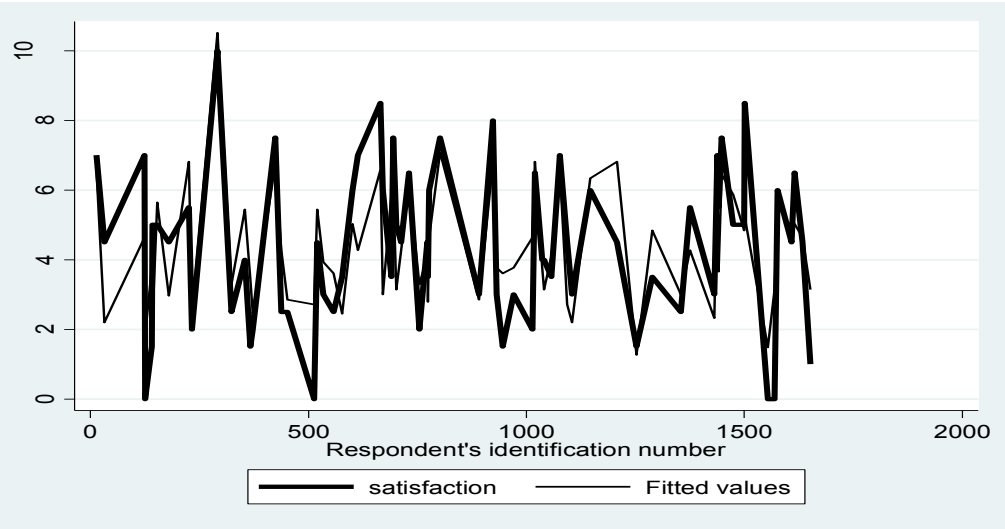
skirtingi algoritmai. SPSS remiasi Levenberg – Marquardt algoritmu, o STATA naudoja patį paprasčiausią mažiausiųjų kvadratų metodą, tarytum tirtume tiesinės regresijos modelį. Todėl SPSS rezultatai (nors ir ne tokie puikūs) yra patikimesni. STATA lentelę su parametrų galutiniais įverčiais (jie praktiškai nesiskiria nuo gautųjų SPSS programa), jų pasikliautiniais intervalais ir *p* reikšmėmis galima naudoti tik, kaip *rekomendacijas.* Pavyzdžiui, atsižvelgti, kad modelio koeficientas *a* buvo statistiškai nereikšmingas (p = 0,661), ir todėl prasminga patyrinėti modelį be koeficiento *a*.

Modelio gerumą visada verta patikrinti ir grafiškai. Sukuriame prognozuojamų *satisfaction* reikšmių stulpelį duomenyse (pavadiname jį *PRED*) įvykdydami komandą:

```
predict PRED
```

Braižydami grafiką, pasinaudosime duomenyse esančiu respondento identifikaciniu numeriu *idno.* Įvykdome:

```
graph twoway line  satisfaction PRED  idno
```



Matome, kad naujasis modelis gana neblogai tinka duomenims. Tačiau jis sudėtingesnis, nei tiesinės regresijos modelis ir turi žymiai mažiau modelio tinkamumą aprašančių charakteristikų.

Uždaviniai

1. Failas **ESS4LT1**.sav. Socialinis Lietuvos respondentų tyrimas 2009 m. spalis – 2010 m. Naudojant kintamuosius *agea* (amžius) ir *gndr* (lytis), atrinkti nejaunesnes nei 65 m amžiaus moteris. Kintamieji

- *trstprl* (pasitikėjimas šalies parlamentu),

- *trstlgl* (pasitikėjimas teisine sistema),
- *trstplt* (pasitikėjimas politikais)

matuojami 10 balų skale nuo 0 (itin nepasitikiu) iki 9 (itin pasitikiu). Reikia:

a) Ištirti tiesinės regresijos modelį *trstprl =f(trstlgl, trstplt).*

b) Kurios tiesinės regresijos prielaidos nepatenkintos?

c) Kokius modelius gautume, taikydami šio skyrelio tiesinės regresijos alternatyvas?

2. Pateiksime tyrimo „Darbo valandų skaičius ir atlyginimas, 2008 m birželis“ (http://www.lidata.eu/data/quant/LiDA_SEB_0111_STUDY_01) truputį paredaguotą fragmentą. Patektame fragmente 42 lenkų tautybės respondentai atsakė į klausimą, kiek valandų per savaitę jie sutiktų dirbti, jeigu gautų 1000 LT atlyginimą (*K1*) ir 2000 LT atlyginimą (*K2*). Sudaryti regresiją, nusakančia *K2* priklausomybę nuo *K1*.

a) Ar patenkintos tiesinės regresijos prielaidos?

a) Kuri tiesinės regresijos alternatyva geriausiai tinka duomenims?

b) Kaip atrodo atsakymas į b), pašalinus iš duomenų išskirtį?

ID	*K1*	*K2*	*ID*	*1*	*K2*	*ID*	*K1*	*K2*
1	20	26	15	0	50	29	20	40
2	10	25	16	35	35	30	40	40
3	40	40	17	30	30	31	40	40
4	25	34	18	25	35	32	40	40
5	2	20	19	40	40	33	20	30
6	18	22	20	20	30	34	15	25
7	20	30	21	0	15	35	30	30
8	30	40	22	30	30	36	30	40
9	40	40	23	5	1	37	10	20
10	30	30	24	20	30	38	5	7
11	20	25	25	20	25	39	30	40
12	20	40	26	30	45	40	30	35
13	24	40	27	15	20	41	20	30
14	20	35	28	20	40	42	10	15

3. Norint paskatinti kaimo turizmo vystymąsi, vienoje besivystančioje šalyje buvo pradėta supirkinėti vikriojo driežo uodegas. Lentelėje pateikta kiek kiekvienais metais po naujojo potvarkio tokių uodegų buvo supirkta .

Metai	Kiekis
1	40
2	113
3	285
4	548
5	836
6	1333
7	2149
8	3233
9	3537
10	4349
11	4566
12	4771
13	4970
14	5025
15	5227

Ištirti supirkto uodegų kiekio priklausomybę nuo metų.

- Ar tinka tiesinė regresija?
- Ar tinka tiesinė regresija, panaudojus logaritminę *Kiekio* transformaciją?
- Patikrinti, ar tinka netiesinės regresijos modelis:

$$Kiekis = \frac{a}{b + e^{ct}},$$

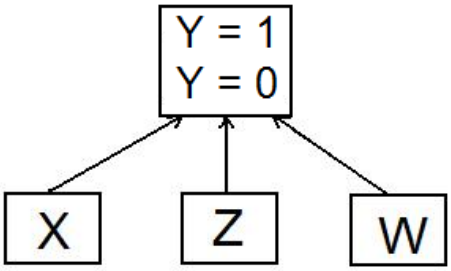
Pradinės parametrų reikšmės $a = 1$, $b = 1$, $c = 0$.

4. DVINARĖ LOGISTINĖ REGRESINĖ ANALIZĖ

Tipinės situacijos, kai taikoma dvinarė logistinė regresinė analizė yra šios:

- Norime įvertinti onkologinės ligos remisijos po operacijos tikimybę, atsižvelgdami į paciento lytį, amžių, išoperuoto auglio didumą ir kraujo parametrus.
- Tiriame, kaip požiūris į narystę NATO (pritaria narystei – nepritaria) priklauso nuo respondento amžiaus, tautybės, partiškumo ir metinių pajamų.
- Norime atsakyti į klausimą, kaip kiekviena papildoma su repetitoriumi praleista valanda pakeičia tikimybių išlaikyti egzaminą / neišlaikyti egzamino santykį.

Modeliuojamas *dvireikšmis* kintamasis Y priklauso nuo X, Z, W. Modelio schema:



Kintamasis *Y* vadinamas *priklausomu* (arba *regresuojamu*) *kintamuoju*, kintamieji *X, Z, W* vadinami *nepriklausomais* (*aiškinamaisiais*) kintamaisiais arba *regresoriais*. Tradiciškai manoma, kad *Y* įgyjamos reikšmės yra **kodai 0 ir 1**. Ką šie kodai reiškia, priklauso nuo konkretaus tyrimo. Gali būti, kad 1 žymės, jog pacientas susirgo diabetu; o gali ir tai, kad respondentas yra vegetaras. Kai kurios statistinės programos (pvz., SPSS) automatiškai perkoduoja kiekvieną logistinėje analizėje naudojamą dvireikšmį priklausomą kintamąjį, mažesnę reikšmę keisdamos į nulį, o didesnę – į vienetą.

4.1. Dvinarės logistinės regresijos modelis

4.1.1. Modelio lygtis

Matematinis modelis sudaromas ne pačiam priklausomam kintamajam, o jo tikimybių santykio logaritmui (vadinamajai *logit funkcijai*):

$$\ln \frac{\mathrm{P}(Y=1)}{\mathrm{P}(Y=0)} = C + b_1 X + b_2 Z + b_3 W.$$

Ekvivalentus modelio užrašas yra

$$\mathrm{P}(Y=1)=\frac{\mathrm{e}^{z}}{1+\mathrm{e}^{z}}=\frac{1}{1+\mathrm{e}^{-z}}, \qquad \mathrm{P}(Y=0)=1-\mathrm{P}(Y=1).$$

Čia

$$z = C + b_1X + b_2Z + b_3W.$$

Koeficientų C, b_1, b_2, b_3 reikšmės nėra žinomos. Jų įverčiai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$ gaunami, panaudojus imties duomenis.

Kokybinei analizei pakanka atsižvelgti į koeficientų reikšmes. Jeigu koeficientas prie kažkurio kintamojo teigiamas, tai šiam kintamajam didėjant, tikimybė $Y = 1$ (o kokią situaciją ta modelio lygybė $Y = 1$ atitinka, turime žinoti iš sąlygos) didėja. Jeigu koeficientas neigiamas, tai atitinkamam kintamajam didėjant, tikimybė Y įgyti vienetą mažėja (didėja tikimybė, kad Y įgis 0).

Jeigu $b_1 > 0$, tai X didėjant, didėja ir tikimybė P(Y= 1).

Jeigu $b_1 < 0$, tai X didėjant, didėja ir tikimybė P(Y= 0).

Kadangi regresorių matavimo skalės gali labai skirtis, tai bendriausiu atveju negalima teigti, kad svarbiausias yra tas kintamasis, kurio koeficientas absoliutiniu didumu didžiausias. Pavyzdžiui, toks lyginimas yra nekorektiškas, jeigu X yra pragyventi metai, o Z – kūno masės indeksas. Koeficientų didumus galima lyginti, jeigu visi regresoriai matuojami ta pačia skale (pavyzdžiui, regresoriai yra skirtingų egzaminų pažymiai). Standartizuotų beta koeficientų analogų, leidžiančių palyginti regresorių svarbą, logistinėje regresijoje nėra. Šiek tiek pagelbsti galimybių santykiai.

4.1.2. Galimybių santykis

Puiku būtų pasakyti, kaip stipriai nuo regresorių priklauso tikimybė P(Y = 1). Deja, tą padaryti galima tik konkrečioms regresorių reikšmėms, o norisi kažkokios apibendrintos charakteristikos. Tokia charakteristika yra galimybių santykis.

Galimybe[12] (angl. odds) vadinamas tikimybių santykis P(Y=1) / P(Y= 0). Jis parodo, kiek kartų viena Y reikšmė tikėtinesnė už kitą. Informacijos atžvilgiu galimybė yra tolygi tikimybei. Jeigu pacientui susirgti diabetu tikimybė yra 0,75, tai tą pačią informaciją galime pasakyti taip: tikimybė susirgti diabetu yra tris kartus didesnė, nei nesusirgti, t.y. galimybė lygi 3:1. Kaip ir tikimybę, galimybę galima apskaičiuoti tik konkretiems duomenims. Pavyzdžiui, surasti, koks bus santykis P(Y=1) / P(Y= 0), kai X = 2, W = 3, o Z = 9.

[12] Lietuviškas terminas *galimybė* nėra geras. Tiesiog, kai reikėjo, nieko geresnio su Gediminu Murausku nesumąstėme. Be to, niekas iš mūsų kolegų ir studentų nepasiūlė tinkamesnio termino, nors visų prašėme ir laukėme porą metų (protingiausias buvo *šansas*). Tai (vis dar) puiki proga įsiamžinti lietuviškoje terminijoje.

Galimybių santykis (angl. odds ratio) parodo, kaip pasikeis tikimybių santykis (galimybė), kai atitinkamas regresorius padidės vienetu, fiksavus visų kitų regresorių reikšmes. Kai regresorius X padidėja vienetu, ankstesnis tikimybių santykis pasikeičia $e^{\hat{b}_1}$ kartų. Šis dydis ir vadinamas galimybių santykiu. Nesunku įsitikinti, kad galimybių santykis didesnis už vienetą tik tada, kai atitinkamas koeficientas b teigiamas. Todėl, galima sakyti, kad didesnis už vienetą galimybių santykis rodo *kiek kartų,* lyginant su ankstesne galimybe, tampa labiau tikėtina, kad Y =1, o ne Y = 0. Analogiškai, mažesnis už vienetą galimybių santykis kiekybiškai charakterizuoja *kiek kartų,* lyginant su ankstesne galimybe, tampa mažiau tikėtina, kad Y =1, o ne Y = 0.

Skaičius $e^{\hat{b}_1}$ yra galimybių santykis (odds ratio):

$$\left(\frac{\mathrm{P}(Y=1)}{\mathrm{P}(Y=0)}\right)_{naujas} = e^{\hat{b}_1}\left(\frac{\mathrm{P}(Y=1)}{\mathrm{P}(Y=0)}\right)_{senas}$$

Pavyzdys. Tarkime, kad tirdami diabetą sukeliančius veiksnius (susirgs/nesusirgs) , nustatėme, jog galimybių santykis kintamajam X (kiek savaičių pacientas laikėsi kiniškų žolelių dietos) yra 0,5. Tai reiškia, kad X padidėjus vienetu, tikimybių santykis P(susirgs)/P(nesusirgs) sumažėja dvigubai. Pavyzdžiui, jeigu konkrečiam pacientui pradinis tikimybių P(susirgs)/P(nesusirgs) santykis buvo 3:1, tai papildoma dietos savaitė jam sumažina šį santykį iki 1,5:1.

Galimybių santykis priklauso tik nuo modelio koeficientų, todėl tai ir yra ta kintamojo svarbą parodanti kiekybinė charakteristika, kuri dažnai pateikiama logistinės regresijos išvadų aprašymuose. Statistinės programos dažniausiai kiekvienam kintamajam pateikia ne tik patį galimybių santykį, bet ir šio santykio 95% pasikliautinąjį intervalą. Jeigu visų regresorių skalės vienodos, tai galimybių santykius galima taikyti regresorių svarbos nustatymui. Tada galima tarti, kad svarbesnis tas regresorius, kurio pokytis *daugiau kartų pakeičia* galimybių santykį (padidina jį arba sumažina). Nesunku suvokti, kad kintamasis, kurio pokytis galimybę patrigubina yra svarbesnis už tą, kuris galimybių santykį padvigubina. Analogiškai išmąstome, kad svarbesnis tas kintamasis, kurio pokytis galimybę sumažina triskart (galimybių santykis 1/3), nei tas, kurio pokytis galimybę sumažina dukart (galimybių santykis 1/2). Kai regresorių skalės visiškai skirtingos (pvz., vienas matuojamas tonomis, o kitas metrais), tai jokie galimybių santykiai nepadės nustatyti, kuris kintamasis modelyje svarbesnis. Ir niekas nepadės.

4.1.3. Prognozavimas

Prognozė, kam bus lygi tikimybė P(Y = 1), atliekama tik konkrečioms *X, Z, W* reikšmėms. Prognozę lengviausia atlikti dviem etapais. Pradžioje į formulę

$$z = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W$$

įstatome tas konkrečias *X*, *Z* ir *W* reikšmes, kurioms reikia prognozės. Tada tikimybių įverčiai apskaičiuojamas pagal formules:

$$\widehat{\mathrm{P}}(Y = 1) = \frac{\mathrm{e}^z}{1 + \mathrm{e}^z}, \quad \widehat{\mathrm{P}}(Y = 0) = 1 - \widehat{\mathrm{P}}(Y = 1) = \frac{1}{1 + \mathrm{e}^z}.$$

Čia e = 2,71828182845904523536028747135 27……Praktiniams skaičiavimams visiškai pakanka apytikslės reikšmės e $\approx$ 2,7183.

Pavyzdys. Tarkime, kad atlikę logistinę regresiją, gavome tokius koeficientų įverčius $\hat{C} = 2$, $\hat{b}_1 = -4$, $\hat{b}_2 = 3$, $\hat{b}_3 = 1$. Užrašome logit funkcijos modelį:

$$z = 2 - 4X + 3Z + W.$$

Tarkime, kad turime respondentą, kuriam *X* = 0, *Z* = 8, *W* = 2. Šiam respondentui norime įvertinti tikimybę P(Y = 1). Pradžioje įstatome *X, Z* ir *W* reikšmes į logit funkciją ir gauname *z* = 2+0+24+2 = 28. Tada pakeliame e gautuoju laipsniu:

$$\mathrm{e}^{28} = 1446257064291{,}475.$$

Beliko apskaičiuoti tikimybės įvertį:

$$\widehat{\mathrm{P}}(Y = 1) = \frac{1446257064291{,}475}{1 + 1446257064291{,}475} = 0{,}99999\ldots$$

Taigi, šiam respondentui labai tikėtina, kad *Y* = 1.

Kartais reikia tik nustatyti, kuriai iš dviejų kategorijų turėtų priklausyti respondentas. Pavyzdžiui, atsižvelgę į antsvorį ir dietą, prognozuojame, kad Pūkuotukas susirgs diabetu, o Paršelis nesusirgs. Toks grubus priskyrimas vienai iš kategorijų (Y = 1 arba Y = 0) vadinamas *klasifikavimu*. Jis atliekamas taip: konkrečioms regresorių reikšmėms taip, kaip aprašyta aukščiau, surandamas tikimybės P(Y = 1) įvertis $\widehat{\mathrm{P}}(Y = 1)$. Prognozuojame tą reikšmę, kuri labiau tikėtina. Jeigu $\widehat{\mathrm{P}}(Y = 1) > 0{,}5$, tai prognozuojame Y = 1. Jeigu $\widehat{\mathrm{P}}(Y = 1) < 0{,}5$, tai prognozuojame Y = 0. Jeigu $\widehat{\mathrm{P}}(Y = 1) = 0{,}5$, tai rekomenduojama sprendimą priimti atsitiktinai, pavyzdžiui, metant monetą. Tiesą sakant, klasifikavimui tikimybių įverčiai nebūtini. Pakanka surasti *z*. Primename, kad tikslus *z* pavadinimas yra *tikimybių santykio logit funkcija*.

Prognozė, atsižvelgiant į tikimybių santykio logit funkcijos reikšmę *z*.

- Jeigu z > 0, tai prognozuojama, kad *Y* = 1.
- Jeigu z < 0, tai prognozuojama, kad *Y* = 0.
- Jeigu z = 0, tai rekomenduojama sprendimą priimti, metant monetą.

4.1.4. Reikalavimai duomenims

Duomenys turi tenkinti tokius reikalavimus:

a) *Priklausomas kintamasis Y yra dvireikšmis*. Formaliai kalbant, visi kiti kintamieji gali būti arba intervaliniai, arba dvireikšmiai (įgyti reikšmes 0 arba 1). Vis dėlto, kai modelyje daug kategorinių kintamųjų, korektiškiau kalbėti apie apibendrintąjį tiesinį modelį, o ne apie logistinę regresiją. Tada ir truputį kitokia analizė taikoma, žr. 4.2.11 skyrelį. Todėl geriausia, kai dauguma regresorių yra intervaliniai, o dvireikšmių kintamųjų nedaug.

a) *Duomenyse negali vyrauti viena iš Y reikšmių*. Dažniausiai reikalaujama, kad tarp stebimų *Y* reikšmių vienetų (nulių) būtų ne mažiau penktadalio. Tokio reikalavimo logika paprasta – kad galėtume nustatyti, kas būdinga kiekvienai kategorijai, turime turėti informacijos apie abiejų kategorijų respondentus. Negalima išsiaiškinti, kas būdinga diabetu sergantiesiems, tiriant tik sveikus respondentus**.**

b) Jeigu modelyje yra daug kategorinių kintamųjų, tai *kiekvienam tų kategorijų deriniui turi būti pakankamai daug respondentų*. Minimalus reikalavimas – bent 5 respondentai. Jeigu atsižvelgiame į lytį ir žvairumą, tai negerai, jei imtyje nėra žvairų moterų. Ir neužteks to, kad ir moterų yra ne mažiau 20 % ir žvairuojančių tiriamųjų yra ne mažiau 20 %.

c) *Regresoriai neturi stipriai koreliuoti*. Dėl multikolinearumo gali atsirasti neteisingi kintamųjų ryšiai modelyje. Pavyzdžiui, gali būti taip, kad logistinės regresijos modelį papildžiui naujais regresoriais, regresoriaus *X* daugiklis pakeis ženklą. Vis dėlto, multikolinearumo problema dvinarėje logistinėje regresijoje tiriama retai.

4.1.5. Modelio tinkamumas

Modelio tikimą duomenims parodo tokie rodikliai:

1) *Klasifikacinė lentelė.* Tai bene svarbiausia modelio tikimo charakteristika. Jeigu jau sudarėme logistinės regresijos modelį duomenims, kuriems žinomos *Y* reikšmės, tai natūralu pasitikrinti, ar gerai veikia klasifikavimas. Konkretiems stebėjimams prognozuojama *Y*

reikšmė, ir žiūrima, ar spėjimas sutapo su tikrąja *Y* reikšme. Kuo daugiau sutapimų, tuo modelis geresnis. Kiekvienos kategorijos stebėjimų turi būtu suklasifikuota daugiau, nei gautume atsitiktinai spėdami. Jeigu abiejų *Y* reikšmių imtyje yra panašus kiekis, tai tinkami modeliai kiekvienos iš *Y* kategorijų teisingai klasifikuos ne mažiau, kaip po 50 procentų.

2) *Didžiausio tikėtinumo chi kvadrato statistika*. Parodo ar modelyje yra bent vienas reikalingas regresorius. Jeigu statistikos p reikšmė $\geq$ 0,05, tai regresijos modelio tinkamumas labai abejotinas (gavome, kad visi regresoriai nereikalingi). Jeigu p < 0,05, tai gavome patvirtinimą, jog modelis nėra beviltiškas (o gal ir visai geras – reikia tirti toliau).
3) *Hosmerio – Lemešou chi kvadrato statistika.* Hosmerio – Lemešou kriterijus yra žymiai mažiau populiari alternatyva aukščiau minėtam didžiausio tikėtinumo chi kvadrato kriterijui. Modelis gerai tinka duomenims, kai Hosmerio – Lemešou chi kvadrato statistikos p reikšmė $\geq$ 0,05. Hosmerio – Lemešou kriterijų rekomenduojama taikyti tik nedidelėms imtims (iki 200 stebėjimų taikyti galima drąsiai, virš 500 stebėjimų – jau labai atsargiai).
4) *Voldo testai regresoriams*. Padeda nuspręsti ar kintamasis šalintinas iš modelio. Jeigu regresoriui Voldo testo p reikšmė < 0,05, tai sakome, kad *regresorius yra statistiškai reikšmingas* ir dažniausiai jį modelyje paliekame. Jeigu $p \geq$ 0,05, tai regresorius yra statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais (žr. kitą skyrelį). Dažniausiai modelio konstanta *C* paliekama net ir tada, kai ji statistiškai nereikšminga. Voldo kriterijus nelaikomas labai patikimu kintamojo svarbos rodikliu – pasitaiko, kad itin svarbų regresorių jis rodo esant statistiškai nereikšmingą.
5) *Determinacijos (pseudo) koeficientai* (angl. R square, pseudo - R^2). Rodo bendrąjį modelio tikimą duomenims. Įgyja reikšmes iš intervalo [0, 1]. Kuo koeficiento reikšmė didesnė, tuo modelis geriau tinka duomenims. Ne itin gerai, kai R^2 yra labai mažas. Vis dėlto, priešingai tiesinei regresijai logistinėje regresijoje R^2 vaidina tik labai menką pagalbinį vaidmenį. Jeigu R^2 yra mažas, bet pagal visus kitus rodiklius logistinė regresija tinka, tai vis tiek tariama, kad modelis tinkamas. Determinacijos pseudokoeficientų yra net keli. Galima naudoti bet kurį. Tiesa, privalu nurodyti tikslų determinacijos koeficiento pavadinimą.
6) *Kuko matas.* Parodo ar duomenų imtyje yra išskirčių. Išskirčiai Kuko matas viršija 1. Skaičiuojamas kiekvienam respondentui. Modelis su išskirtimis – nepatikimas.
7) *DFB statistika.* Tai alternatyva Kuko matui, padedanti surasti tas duomenų išskirtis, kurios daro įtaką koeficientų *b* įverčiams. Išskirčiai *DFB > 1. DFB* statistika skaičiuojama kiekvienam regresoriui atskirai. Modelis su išskirtimis – nepatikimas.
8) *Prognozuotų tikimybių grafikas.* Tai tikrai nėra labai svarbus tikimo rodiklis (ir itin negražus). Vis dėlto, kai kada toks grafikas pridedamas. Jame matyti, ar teisingai

klasifikuotų stebėjimų $Y = 1$ modelio tikimybės buvo didelės, o teisingai klasifikuotų stebėjimų $Y = 0$ tikimybės buvo mažos. Taip turi būti geriems modeliams.

Yra ir kitų rodiklių, pavyzdžiui, standartizuotosios liekamosios paklaidos, į kuriuos galima atsižvelgti, tikrinant liekamųjų paklaidų normalumą.

4.1.6. Dvinarės logistinės regresinės analizės etapai

Išsamios modelio analizės etapai:

1) Sudarome dvinarės logistinės regresijos modelį ir peržiūrime klasifikacinę lentelę. Tikriname, ar kiekvienai Y kategorijai teisingai klasifikuota bent 50% atvejų. Geriau, kai šie procentai žymiai didesni.
2) Patikriname, ar chi kvadrato statistikos p reikšmė $< 0,05$. Jeigu ne – modelis netinkamas. (Nedidelėms imtims alternatyviai galima patikrinti, ar Hosmerio - Lemšou chi kvadrato statistikos p reikšmė $\geq 0,05$. Jeigu ne – modelis netinkamas.)
3) Patikriname, ar visi regresoriai statistiškai reikšmingi (visos Voldo testo p reikšmės $< 0,05$). Jeigu ne – modelis taisytinas. Konstantai p reikšmės nežiūrime.
4) Patikriname, ar modelyje nėra išskirčių (nėra stebėjimų, kuriems Kuko matas > 1). Alternatyviai galima patikrinti, ar nėra stebėjimų, kuriems $DFB > 1$.
5) Žiūrime, ar koeficientų ženklai atitinka spėjamą regresorių įtaką (neįtariame multikolinearumo).
6) Jeigu viskas gerai – modelį aprašome. Jeigu ne – tobuliname.

Labai gerai duomenims tinkančiame modelyje:

- Chi kvadrato $p < 0,05$.
- Visiems regresoriams Voldo kriterijaus $p < 0,05$.
- Koeficientų ženklai neatrodo nelogiški.
- Teisingai klasifikuojama ne mažiau 50% atvejų, kai $Y = 1$ ir kai $Y = 0$.
- Nedideliai imčiai Hosmerio – Lemešou statistikos $p \geq 0,05$.
- Visų duomenų Kuko matai ≤ 1 ir visi DFB ≤ 1 .
- Pasirinktasis determinacijos koeficientas $\geq 0,20$.

4.1.7. Modelio tobulinimas

Modelio tobulinimas reiškia regresorių šalinimą arba jų transformaciją. Transformacija (logaritmavimas ir pan.) retai kada padeda, bet pabandyti visada galima.

Pašalinus bent vieną kintamąjį, keičiasi visos modelio charakteristikos: visos *p* reikšmės, klasifikacinė lentelė, determinacijos koeficientai.

Modelis tobulinamas, jeigu yra problemų:

a) *Yra statistiškai nereikšmingas regresorius* (jam Voldo kriterijaus $p \geq 0{,}05$). Regresorius pašalinamas iš modelio ir logistinė regresinė analizė pakartojama be jo. Viskas gerai, jeigu klasifikacinėje lentelėje teisingai klasifikuotų atvejų sumažėjo nežymiai (kartais teisingų klasifikavimų net padaugėja). Ką reiškia nežymiai, sprendžia pats tyrėjas. Jeigu teisingai klasifikuotų atvejų sumažėja radikaliai, tai regresorių į modelį gražiname. Aprašydami tada konstatuosime, kad „regresorius buvo statistiškai nereikšmingas. Vis dėlto jis paliktas modelyje, nes be jo labai sumažėja teisingų klasifikavimų procentas“.

b) *Yra multikolinearumo problema.* Multikolinearumas reiškia, kad modelyje yra stipriai koreliuotų regresorių, kurie iškreipia tikrąsias kintamųjų priklausomybes (modelyje atsiranda neįtikėtini koeficientų ženklai). Tada pasitikrinama, ar intervaliniai regresoriai stipriai koreliuoja. Jeigu taip – tai dalį regresorių pašaliname. Šalinami tie regresoriai, kurių standartinės paklaidos SE koeficientų lentelėje prie Voldo kriterijaus reikšmių yra didžiausios.

Kai kurie autoriai pataria palikti tik tuos regresorius, kurių SE neviršija 5.

Kartais (jeigu toks veiksmas prasmingas) stipriai koreliuojančius regresorius keičiame jų vidurkiu.

c) *Yra išskirčių.* Jeigu visi DFB ≤ 1 , o daliai stebėjimų Kuko matai nedaug viršija 1 (arba atvirkščiai), tai modelis laikomas pakankamai geru. Žinoma, aprašymuose tada ir figūruoja tik DFB (Kuko matas). Apskritai nederėtų to 1 suabsoliutinti. Nieko jau taip blogo neįvyks, jeigu ir DFB ir Kuko matas nedaug viršys 1. Kartais išskirtį galima pašalinti (pavyzdžiui, įtarus vedimo klaidą ar gryną atsitiktinumą). Jeigu modelyje yra gyvybiškai svarbi piktybinė išskirtis, tai (deja, deja) modelis yra netinkamas. Teks verstis be regresinio modelio.

d) *Pasirinktasis determinacijos pseudokoeficientas < 0,20.* Galima pabandyti pasirinkti kitą koeficientą (pavyzdžiui, Nagelkerkės, o ne Kokso-Snelo). Be to, logistinėje regresijoje nėra labai svarbu, jeigu determinacijos koeficientas mažas. Jeigu R^2 = 0,15, o visos kitos charakteristikos rodo gerą modelio tikimą, tai tariama, kad modelis yra tinkamas.

4.1.8. Pastabos apie logistinės regresijos taikymą

- Kaip taisyklė, taikant logistinę regresinę analizę, apsiribojama tik vienu – chi kvadrato arba Hosmerio - Lemešou – kriterijumi. Hosmerio – Lemešou kriterijus taikytinas, kai imtis nedidelė, nes didelėms imtims *p* reikšmė gali būti maža vien dėl imties didumo.
- Aprašymuose neretai nurodomas tik bendrasis teisingų klasifikavimų procentas. Tai nėra blogai. Tik reikia nepamiršti, kad modelis turi padėti gerai atpažinti abi *Y* reikšmes. Priešingu atveju, galima padaryti neteisingą išvadą. Pavyzdžiui, galima pamanyti, kad modelis latentiniam kairiarankiškumui nustatyti gerai tinka, nes bendras teisingų klasifikacijų procentas yra 60%. Tuo tarpu, atidžiau pažvelgus, ims ir paaiškės, kad taikomas modelis 100% atpažįsta dešiniarankius ir 0% – kairiarankius, t.y., klasifikavimo taisyklė labai paprasta (ir kvaila) – nesvarbu kokios ten tos kintamųjų reikšmės, sakysime, kad respondentas dešiniarankis. Tiesiog dešiniarankių imtyje buvo 60%.
- Šiame konspekte išsamiai aptariama tik tiesioginė dvinarė logistinė regresija, kai visus modelio tobulinimus atlieka pats tyrėjas. Tyrimuose dar taikomi keliaetapiai (žingsniniai) logistinės regresijos modeliai, kai modelyje automatiškai paliekami patys statistiškai reikšmingiausi iš pradinių kintamųjų.
- Priešingai, nei tiesinėje regresijoje, jeigu daugiklis prie *X* yra $\hat{b}_1 = 3$, tai **negalima** sakyti, kad *X* padidėjus vienetu, *Y* padidės 3.
- Gali būti ir vienas regresorius. Juo turi būti intervalinis kintamasis.
- Gana dažnai pasitaikanti klaida išraišką $z = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W$ pavadinti tikimybe. Todėl dar kartą primename: z nėra tikimybė, z yra tikimybių santykio logaritmas.
- Kai *Y* įgyja reikšmes 1 ir 2, jos modelyje **perkoduojamos:** reikšmė Y=1 traktuojama, kaip modelio 0, o Y=2 traktuojama, kaip modelio 1. Todėl, sudarę modelį modelio vieneto tikimybei, turime prisiminti, kad šis vienetas atitinka pradinį kodą Y=2. Tai sukelia papildomas problemas, interpretuojant rezultatus. Todėl labai primygtinai rekomenduojame, skaitinį kintamąjį iš pat pradžių perkoduoti, suteikiant reikšmes 0 ir 1.

4.1.9. Standartiniai regresijos modelių aprašymai

Visada aprašomas galutinis, po visų tobulinimų gautasis, modelis. Privaloma aprašyti modelio chi kvadrato statistikos reikšmingumą ir aptarti regresorių statistinius reikšmingumus. Pateiksime keletą galimų aprašymų fragmentų.

- *Logistinės regresijos modelis tiko, $\chi^2 = 13,27$; $p < 0,01$. Visi kintamieji buvo statistiškai reikšmingi. Teisingai klasifikuota 78,3% imties respondentų. Nagelgerkės determinacijos koeficientas lygus 0,72. Regresoriui X galimybių santykis 1,3; (95% pasikliautinis intervalas [0,99;1,61]),...*
- *Iš modelio buvo pašalintas statistiškai nereikšmingas rgeresorius X. Galutinis modelis buvo statistiškai reikšmingas, $\chi^2 = 10,02$; $p = 0, 023$. Taikant modelį imties duomenims, teisingai klasifikuoti 73% diabetu sergantys pacientai ir 88% sveiki respondentai. Kokso ir Snelo determinacijos koeficientas $R^2 = 0,57$. Didėjant W, diabeto rizika išauga. Didėjant Z, diabeto rizika sumažėja.....*
- *Modelyje paliktas statistiškai nereikšmingas regresorius Z, nes jį pašalinus, labai sumažėja teisingai klasifikuotų atvejų procentas.....*

Skaitytojui labai padeda ir pateikti rezultatų išklotinių fragmentai.

4.2. Dvinarė logistinė regresinė analizė su SPSS

4.2.1. Duomenys

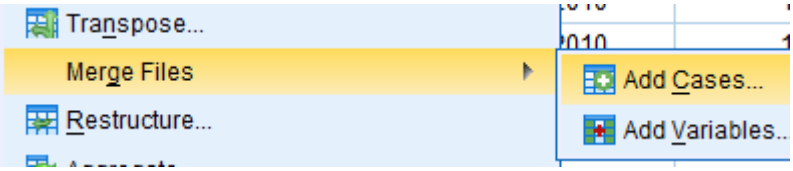
Kaip atlikti dvinarę logistinę regresinę analizę, parodysime, tirdami 2008 metų Europos Sąjungos socialinio tyrimo (European Social Survey) Estijos ir Portugalijos duomenis **ESS4EE** ir **ESS4PT**. Tirsime 20 – 30 metų vyrus. Tyrime naudosime tokius kintamuosius:

- *cntry* – šalies kodas (EE – Estija, PT – Portugalija),
- *stfedu* – respondento požiūris į savo šalies švietimo sistemos būklę (reikšmės nuo 0 – itin bloga, iki 10 – itin gera),
- *happy* – laimingumas (matuojamas 10 balų skale nuo 0 – itin nelaimingas, iki 10 – itin laimingas),
- *freehms* – gėjai ir lesbietės gali gyventi, kaip nori (matuojamas skale nuo 1 – pilnai sutinku, iki 5 – visiškai nesutinku),

- *imsclbn* – nuo kada imigrantai gali pilnai naudotis šalies socialinėmis lengvatomis (matuojamas nominalia skale: 1 – iškart po atvykimo, 2 – po vienerių metų, nepriklausomai nuo to, ar dirbo, 3 – po vienerių darbo ir mokesčių mokėjimo metų, 4 – tapus piliečiu, 5 – niekada negalės naudotis visomis lengvatomis).

Duomenų atrankai naudosime kintamuosius: *agea* – amžių ir *gndr* – lytį (1 – vyr., 2 - mot.),

Abu duomenų failus sujungsime į vieną. Atsidarome ESS4EE ir renkamės **Data → Merge Files→ Add Cases** opciją.

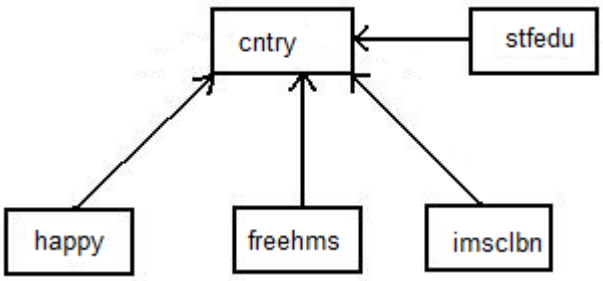


Atsidariusiame lange renkamės **Browse**, nurodome failą ESS4PT ir spaudžiame **Continue**. Atsidariusiame lange spaudžiame **OK**.

Su komanda **Select Cases** atrenkame reikiamus respondentus. Tai daroma taip: meniu **Select Cases** pažymime **If condition is satisfied** ir spaudžiame klavišą **If**. Naujajame meniu suformuluojame sąlygą *agea >= 20 & agea <= 30 & gndr = 1*.

Tada spaudžiame **Continue** ir **OK.**

Tirsime, ar 20 – 30 metų respondento laimingumo laipsnis ir požiūris į seksualines mažumas, švietimo sistemos būklę bei imigrantų socialines lengvatas gerai charakterizuoja, kurios šalies gyventojas jis yra. Grafiškai tiriamą modelį galima pavaizduoti taip:



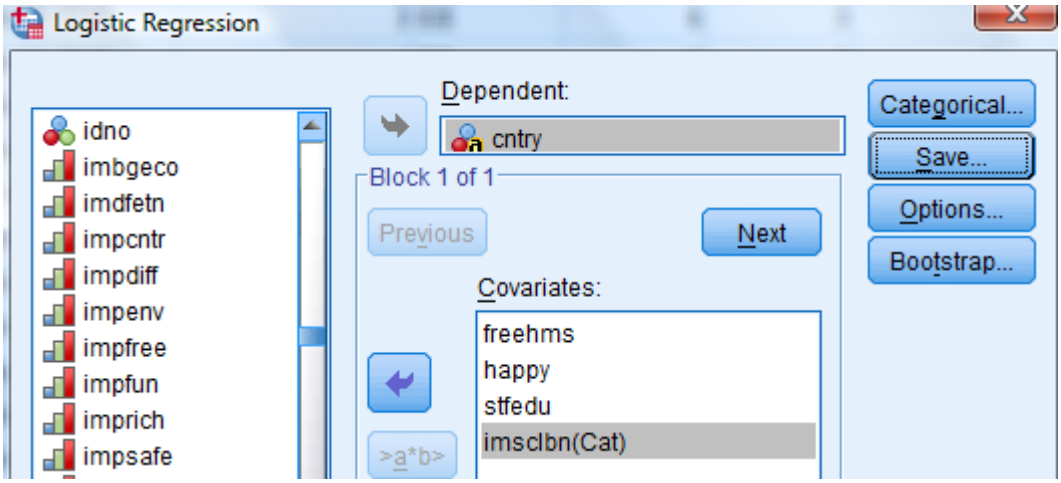
Kitais žodžiais tariant, tikimės, kad parinktieji regresoriai labia gerai apibūdina respondento kilmės šalį. Sąlyginai modelį galima užrašyti taip:

cntry = f(stfedu, happy, freehms, imsclbn).

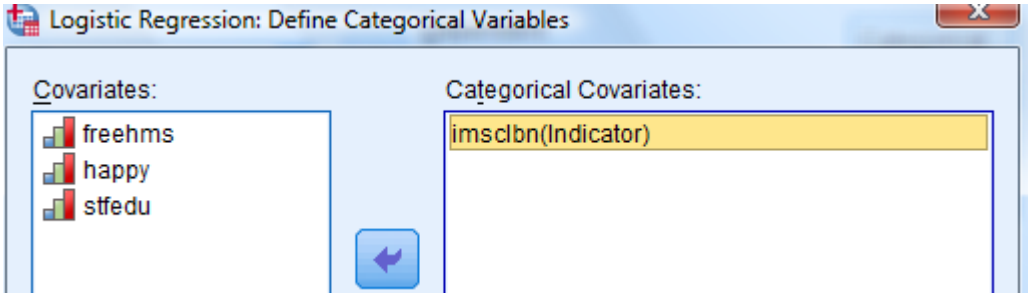
Mes atlikome pradinę modelio specifikaciją. Dabar patikrinsime, ar parinktasis modelis tinka duomenims.

4.2.2. SPSS parinktys pradiniam regresijos modeliui

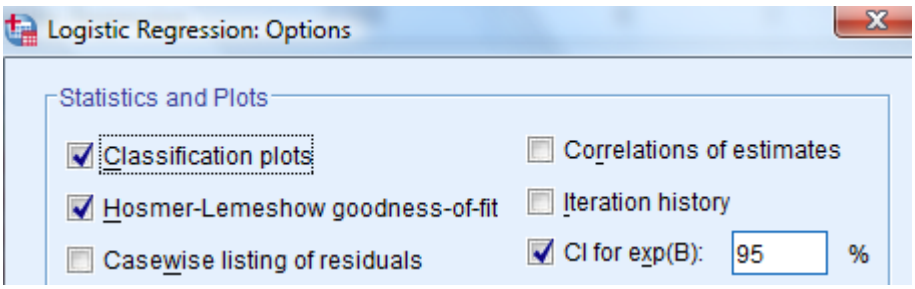
Atsidarome langelį **Analyze** ir renkamės **Regression → Binary Logistic.** Atsidariusiame meniu perkeliame kintamąjį *cntry* į laukelį **Dependent**. Kintamuosius *freehms*, *happy, imsclbn* ir *stfedu* įkeliame į laukelį **Covariates**.



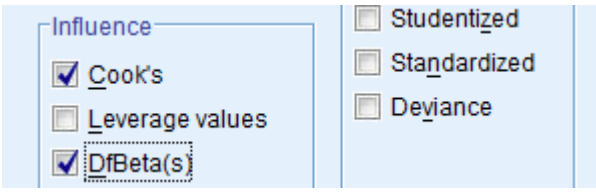
Regresorius *imsclbn* yra kategorinis. Tiesa, kategorijų yra pakankamai daug ir jos beveik ranginės. Labai didelės nuodėmės nebūtų ir traktuoti šį regresorių, kaip intervalinį kintamąjį. Vis dėlto, norėdami parodyti logistinės regresijos taikymo niuansus, tarsime, kad *imsclbn* yra nominalus kintamasis. Todėl renkamės opciją **Categorical** ir perkeliame *imsclbn* į dešinį lauką.



Options meniu pažymime **CI for exp(B).** Kadangi nemanome, kad imtis bus labai didelė, tai pažymime ir **Hosmer-Lemeshow goodness-of-fit.** Pastarojo kriterijaus galima ir nenaudoti, ypač jeigu imtis didelė. Be to, pažymime **Classification Plots.**



Grįžę į prieš tai buvusį meniu, pasirenkame opciją **Save**. Pažymime **Cook's** ir **DfBeta(s)**. Šiame meniu galima pažymėti ir *Probabilities* arba *Group membership*. Tada duomenyse bus sukurtas naujas stulpelis su tikimybės P(Y = 1) įverčiu kiekvienam respondentui. Pasirenkame **Continue** ir **OK**.



Pažymėjus *Standardized* arba *Studentized*, duomenyse atsirastų nauji stulpeliai su normalizuotomis liekamosiomis paklaidomis. Jas galima naudoti, jei jau labai kiltų noras patikrinti liekamųjų paklaidų normalumą.

4.2.3. Rezultatai

Lentelėje **Dependent variable Encoding** pateikiama informacija, kad *cntry* = EE (t.y. šalis Estija) modelyje atitiks $Y = 0$, o *cntry* = PT (t.y., Portugalija) modelyje atitiks $Y = 1$.

Dependent Variable Encoding

Original Value	Internal Value
EE Estonia	0
PT Portugal	1

Nurodėme, kad kintamasis *imsclbn* yra kategorinis. Kintamasis įgyja 5 reikšmes. Todėl modelyje jis keičiamas 4 pseudokintamaisias *imsclbn(1), imsclbn(1), imsclbn(1), imsclbn(1),* galinčiais įgyti reikšmes 0 ir 1. Lentelėje **Categorical Variables Codings** pateikiama informacija, kokia pesudokintamųjų reikšmių kombinacija atitiks įvairias kintamojo *imsclbn* kategorijas. Kai visi pseudokintamieji įgis nulines reikšmes, tai atitiks kintamojo *imsclbn* reikšmę 5. *Kai imsclbn(1)=0, imsclbn(2)=0, imsclbn(3)=1, imsclbn(4)=0*, tai atitiks kintamojo *imsclbn* reikšmę 3 ir t.t.

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)
imsclbn When should immigrants obtain rights to social benefits/services	1 Immediately on arrival	26	1.000	.000	.000	.000
	2 After a year, whether or not have worked	25	.000	1.000	.000	.000
	3 After worked and paid taxes at least a year	128	.000	.000	1.000	.000
	4 Once they have become a citizen	63	.000	.000	.000	1.000
	5 They should never get the same rights	6	.000	.000	.000	.000

Surandame **Block 0:Beginning block : Classification Table.** Tai **NĖRA** klasifikacinė lentelė, kuri rodo klasifikavimo tikslumą, naudojant logistinę regresiją.

SPSS pateikia **dvi** klasifikacines lenteles:

- Pirmoji skirta atsitiktinio spėjimo efektyvumui įvertinti.
- Antroji rodo modelio gerumą (būtent ji ir yra svarbiausia).

Pirmoje lentelėje yra informacija apie klasifikavimo tikslumą, jeigu respondento šalį spėjame aklai. Imtyje yra 125 respondentai iš Estijos ir 123 respondentai iš Portugalijos. Jeigu visą laiką sakysime, kad respondentas iš Estijos, tai turėsime 50,4 % teisingų sprendimų. Logistinės regresijos modelis bus geras tik tada, jeigu modeliuojant atitinkamas teisingų sprendimų procentas bus didesnis.

Classification Table[a,b]

Observed			Predicted: cntry Country: EE Estonia	PT Portugal	Percentage Correct
Step 0	cntry Country	EE Estonia	125	0	100.0
		PT Portugal	123	0	.0
	Overall Percentage				50.4

Lentelėje **Classification Table** (tai antroji klasifikacinė lentelė, ji yra rezultatų išklotinės gale) yra informacija apie tai, kaip gerai pavyksta atpažinti respondento šalį, taikant logistinės regresijos modelį imties duomenims. Matome, kad teisingai buvo atpažinti 77,6% Estijos respondentų ir 78,9% Portugalijos respondentų. Bendrasis teisingai klasifikuotų atvejų procentas yra 78,2%. Paaiškinsime, kaip šie procentai suskaičiuojami. Pirmojoje lentelės eilutėje yra pažymėta, kad iš 125 Estijos respondentų 97 buvo teisingai klasifikuoti, kaip Estijos atstovai, o 28 respondentai klaidingai priskirti Portugalijai. 97 iš 125 ir sudaro 77,6 %. Prisiminę (žr. aukščiau), kad aklas spėjimas duotų tik 50,4 procento teisingų Estijos respondentų klasifikavimų, darome išvadą, kad modelis pakankamai geras.

Classification Table[a]

Observed			Predicted: cntry Country: EE Estonia	Predicted: cntry Country: PT Portugal	Percentage Correct
Step 1	cntry Country	EE Estonia	97	28	77.6
		PT Portugal	26	97	78.9
	Overall Percentage				78.2

a. The cut value is .500

Chi kvadrato kriterijaus statistika ir p reikšmė pateikiamos lentelėje **Omnibus Tests of Model Coefficients**. Mes pasirinkome tiesioginę (ne žingsninę) regresiją, todėl visos trys eilutės lentelėje yra identiškos. Gerai, kai $p < 0,05$. Kadangi $p = 0,000..$, tai modelis gerai tinka. Nors neaišku, ar visi kintamieji modelyje reikalingi.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	122.047	7	.000
	Block	122.047	7	.000
	Model	122.047	7	.000

Lentelėje **Model Summary** pateikiami du determinacijos pseudokoeficientai. Kokso ir Snelo $R^2 = 0,389$, Nagelkerkės $R^2 = 0,518$. Abudu determinacijos koeficientai rodo neblogą modelio tikimą duomenims. Blogai būtų, jeigu $R^2 < 0,20$. (Tiesa, logistinėje regresijoje, jeigu viskas tinka, o determinacijos koeficientai < 0,20, tai modelis dar visai tinkamas).

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	221.738[a]	.389	.518

Lentelėje **Hosmer and Lemeshow Test** yra Hosmerio-Lemešou chi kvadrato statistika ir kriterijaus *p* reikšmė. Imtis buvo vidutinio didumo (n=248), todėl į *p* atsižvelgsime. Gerai, kai *p* > 0,05. Kadangi *p* = 0,665 > 0,05, tai darome išvadą, kad modelio duomenims tinka.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.846	8	.665

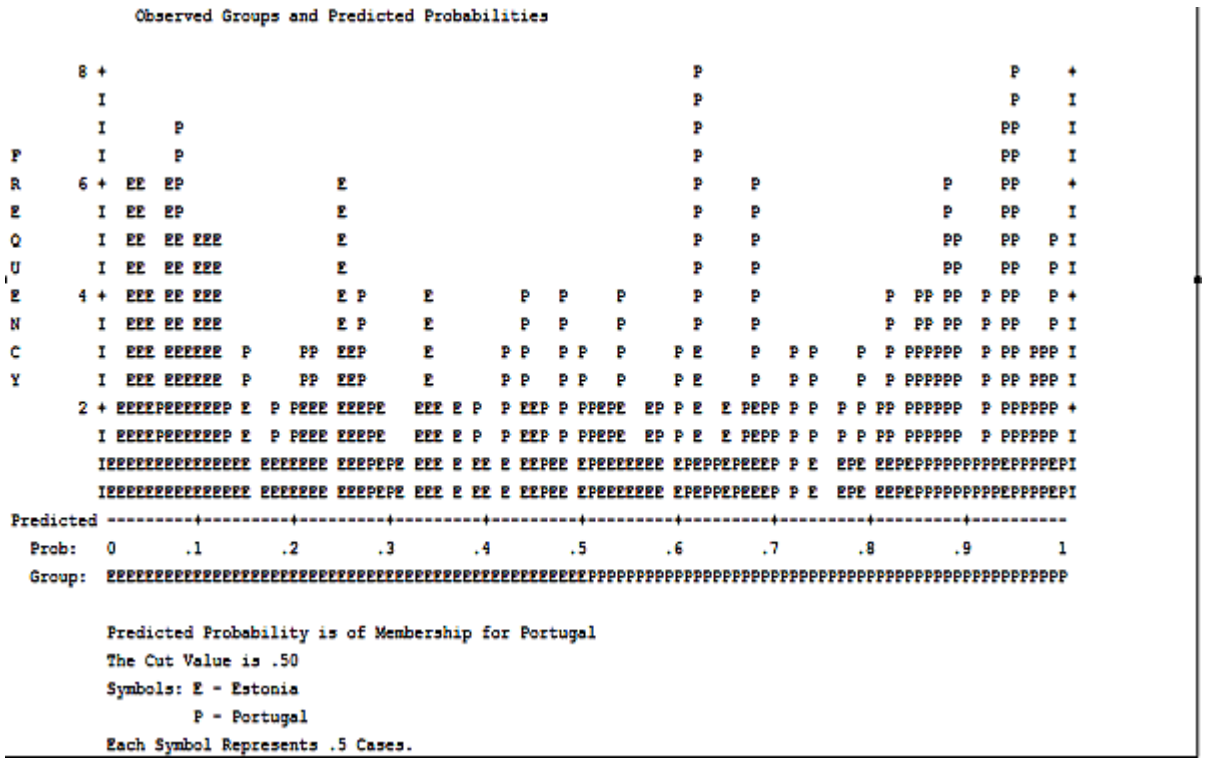
Lentelėje **Variables in the Equation** yra pačių koeficientų reikšmės, o taip pat informacija apie jų statistinį reikšmingumą ir galimybių santykius. Statistiškai reikšmingi tie kintamieji, kuriems stulpelyje **Sig.** pateiktos Voldo kriterijaus *p* reikšmės < 0, 05. Matome, kad yra nemažai statistiškai nereikšmingų kintamųjų. Todėl darome išvadą, kad modelį reikia tobulinti. Modelio užrašą, galimybių santykius ir praktinį prognozavimą aptarsime nagrinėdami galutinį modelį.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	freehms	-.730	.160	20.938	1	.000	.482	.352	.659
	happy	.314	.099	10.025	1	.002	1.369	1.127	1.663
	stfedu	-.564	.092	37.329	1	.000	.569	.475	.682
	imsclbn			15.751	4	.003			
	imsclbn(1)	1.849	1.284	2.075	1	.150	6.354	.513	78.647
	imsclbn(2)	-.039	1.197	.001	1	.974	.962	.092	10.046
	imsclbn(3)	.195	1.123	.030	1	.862	1.215	.134	10.977
	imsclbn(4)	-1.057	1.168	.819	1	.366	.347	.035	3.430
	Constant	2.599	1.422	3.341	1	.068	13.445		

a. Variable(s) entered on step 1: freehms, happy, stfedu, imsclbn.

Rezultatų išklotinė baigiama ganėtinai neišvaizdžiu prognozuotų tikimybių grafiku. Galimos tikimybės reikšmės (nuo 0 iki 1) atidėtos x-ų ašyje. Mūsų modelyje *Y* = 1 ekvivalentu teiginiui, kad respondentas yra iš Portugalijos. Todėl, jeigu P(*Y*=1)> 0,5, tai prognozuojame, kad respondentas yra iš Portugalijos. Jeigu P(*Y*=1) < 0,5, tai prognozuojame, kad respondentas yra iš Estijos. Norint tą pabrėžti, x-ų ašyje iki 0,5 prirašyta raidžių E, o nuo 0,5 – raidžių P. Kiekvienam respondentui suskaičiuojama tikimybė, kad jis yra iš Portugalijos. Atitinkamoje vietoje virš x-ų ašies atidedama raidė E, jeigu tai buvo Estijos atstovas, ir P – jeigu Portugalijos. Raidė E atsidūrusi virš x-ų ašies į kairę nuo 0,5, žymi teisingą esto klasifikavimą. Raidė E atsidūrusi už 0,5 – jau yra Estijos atstovas klaidingai priskirtas portugalams. Analogiškai yra su Portugalijos atstovais – gerai, kai P raidės yra į dešinę nuo 0,5, ir blogai – kai į kairę.



Jeigu logistinės regresijos modelis gerai tinka duomenims, tai matysime daug E raidžių į kairę nuo 0,5 ir daug P raidžių – į dešinę nuo 0,5. Taip šiame grafike ir yra. Todėl darome išvadą, kad prognozuotų tikimybių grafikas irgi rodo, kad logistinės regresijos modelis gerai tinka. Reikia atkreipti dėmesį į tai, kad kiekvienas simbolis atitinka 0,5 respondento, todėl raidžių grafike dvigubai daugiau nei respondentų.

Norint nustatyti, ar imtyje nėra išskirčių reikia patikrinti, ar duomenyse (ne rezultatų išklotinėje) atsiradusio stulpelio *COO_1* ir visų DFB stulpelių reikšmės neviršija vieneto.

COO_1	DFB0_1	DFB1_1	DFB2_1	DFB3_1	DFB4_1	DFB5_1	DFB6_1	DFB7_1
.00038	-.00246	-.00044	.00129	-.00008	-.00101	-.00115	-.00646	-.00084
.01090	.07333	-.01453	-.00353	-.00201	.00052	-.01094	.00477	-.00229

Kuko mato reikšmė yra viena vienam respondentui. Joje atsispindi informacija apie visus to respondento duomenis (*cntr, freehms, happy, stfedu, imsclbn*). Tuo tarpu DFB rodikliai skirti kiekvienam regresoriui ir konstantai atskirai: DFB0_1 – konstantai, DFB1_1 – *freehms*, DFB2_1 – *happy*, DFB3_1 – DFB6_1 skirti kintamiesiems *imsclbn(1) – imsclbn(4).* Peržiūrėti visas stulpelio reikšmes nebūtina. Panaudojame komandą **Analyze → Descriptive Statistics → Descriptives.** Sukeliame visus kintamuosius į **Variable(s)**. Pasirenkame **Options** ir paliekame pažymėtą tik parinktį **Maximum**. Renkamės **Continue** ir **OK**.

Descriptive Statistics

	N	Maximum
COO_1 Analog of Cook's influence statistics	248	.83938
DFB0_1 DFBETA for constant	248	1.03108
DFB1_1 DFBETA for freehms	248	.04821
DFB2_1 DFBETA for happy	248	.02601
DFB3_1 DFBETA for imsclbn(1)	248	.77091
DFB4_1 DFBETA for imsclbn(2)	248	.77503
DFB5_1 DFBETA for imsclbn(3)	248	.77403
DFB6_1 DFBETA for imsclbn(4)	248	.77786
DFB7_1 DFBETA for stfedu	248	.03491
Valid N (listwise)	248	

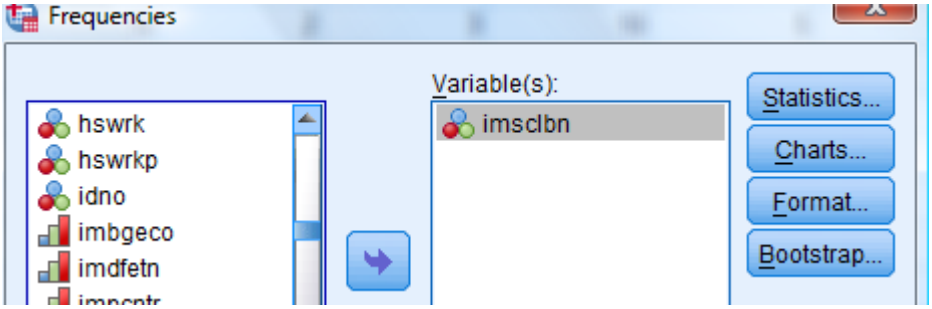
Matome, kad maksimali Kuko mato reikšmė 0,839 < 1. Visos DFB reikšmės, išskyrus DFB0_1, neviršija vieneto. O ir DFB0_1 = 1,03108, t.y. reikšmė praktiškai lygi 1. Be to, tai reikšmė skirta konstantai, t.y. mažiausiai svarbiam iš visų koeficientų. Todėl darome išvadą, kad duomenyse išskirčių nėra.

Nors dauguma rodiklių rodė gerą modelio tikimą, buvo daug statistiškai nereikšmingų kintamųjų. Todėl darome išvadą, kad modelį reikia tobulinti.

4.2.4. Modelio tobulinimas

Standartinė procedūra, kai yra statistiškai nereikšmingų regresorių – dalį jų pašalinti iš modelio. Mūsų atveju taip daryti netinka, nes statistiškai nereikšmingi yra pseudokintamieji, aprašantys **vieną** kategorinį kintamąjį. Nebus logiška, jeigu iš tyrimo pašalinsime visus respondentus, išskyrus tuos, kurie griežtai pasisako prieš imigrantų teises. Reikėtų arba iš viso iš modelio pašalinti kintamąjį *imsclbn*, arba pakeisti jį kitu kintamuoju.

Pradžioje patikrinsime, kiek respondentų rinkosi įvairias *imsclbn* reikšmes. Pasirinkę **Analyze → Descriptice Statistics → Frequences,** gauname kintamojo *imslbnc* dažnių lentelę.

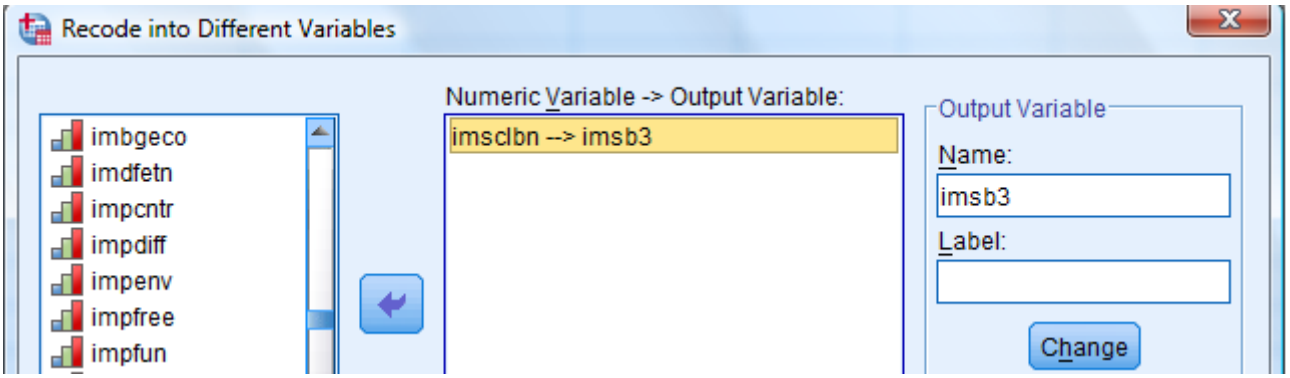


Įsitikiname, kad atsakydami į šį klausimą, tik septyni respondentai pasirinko penketą, dvidešimt septyni respondentai rinkosi vienetą. Tiek pat respondentų rinkosi dvejetą.

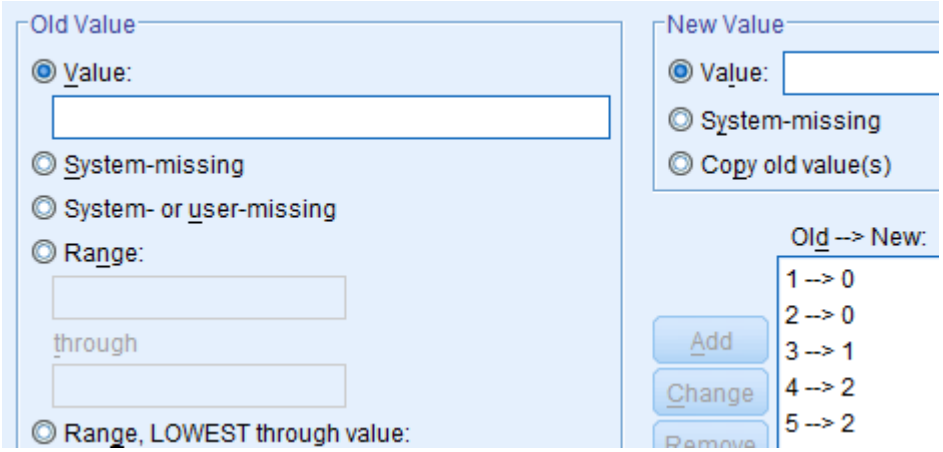
imsclbn When should immigrants obtain rights to social benefits/services

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Immediately on arrival	27	9.7	10.4	10.4
	2 After a year, whether or not have worked	27	9.7	10.4	20.8
	3 After worked and paid taxes at least a year	131	47.3	50.4	71.2
	4 Once they have become a citizen	68	24.5	26.2	97.3
	5 They should never get the same rights	7	2.5	2.7	100.0
	Total	260	93.9	100.0	
Missing	7 Refusal	3	1.1		
	8 Don't know	14	5.1		
	Total	17	6.1		
Total		277	100.0		

Sustambinsime kintamojo *imslbnc* kategorijas. Sukursime naują kintamąjį *imsb3*, kuris įgis reikšmę 0, kai *imslbnc* = 1 arba 2. Kintamasis *imsb3* įgis reikšmę 1, kai *imslbnc* = 3, ir *imsb3* = 2, kai *imslbnc* = 4 arba 5. Pasirenkame **Transform → Recode into Different Variables** ir perkeliame *imsclbn* į dešinį laukelį. Laukelyje **Name** užrašome *imsb3* ir paspaudžiame **Change.**



Pasirenkame **Old and New Value**. Naudodami Old **value – New Value – Add** parinktis nurodome, kaip turi atrodyti naujasis regresorius. Po to spaudžiame **Continue** ir **OK**.



Norėdami paaiškinti naujo kintamojo prasmę papildomai (**Variable View – Labels – Value Labels**) aprašome jį, kaip požiūrį į socialines lengvatas imigrantams (angl. attitude towards social benefits for immigrants), o reikšmes pakomentuojame taip: 0 – teigiamas požiūris (angl. positive attitude), 1 – labiau teigiamas, nei neigiamas (angl. more positive than negative), 2 – neigiamas požiūris (angl. negative attitude).

Tiriame naują modelį:

cntry = f(stfedu, happy, freehms, imsb3).

Tyrimą atliekame lygiai, kaip ankstesnio modelio atveju. Nepamirštame nurodyti, kad *imsb3* yra kategorinis kintamasis. Nepateiksime visų lentelių, tik surašysime, jose gautus rezultatus:

- Priklausomo kintamojo modelio kodai liko nepakitę: *cntr* = EE (Estija), atitinka modelio priklausomo kintamojo reikšmę 0, o *cntr* = PT (Portugalija) atitinka modelio kintamojo reikšmę 1.
- Naujasis kategorinis regresorius *imsb3* pakeičiamas dviem pseudokintamaisiais *imsb3(1)* ir *imsb3(2)*, įgyjančiais tik dvi reikšmes: 0 ir 1. Pseudokintamųjų kodų derinių reikšmės pateiktos lentelėje **Categorical Variables Codings**:

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
imsb3 Attitude towards social benefits for immigrants	.00 positive attitude	51	1.000	.000
	1.00 more positive than negative	128	.000	1.000
	2.00 negative attitude	69	.000	.000

Pavyzdžiui, *imsb3* = 1 atitiks derinys *imsb3(1)* = 0, *imsb3(2)* = 1, o *imsb3* = 2 atitiks derinys *imsb3(1)* = 0, *imsb3(2)* = 0.

- Pradinė klasifikacinė lentelė rodo, kad aklas spėjimas jog kiekvienas respondentas yra iš Estijos, duotų 50,4 % teisingų imties duomenų klasifikavimų.

- Chi kvadrato statistikos (χ^2 = 115,195) p-reikšmė p < 0,05, o Hosmerio – Lemešou χ^2 = 13,132, p = 0,107 > 0,05 rodo gerą modelio tikimą duomenims.
- Kokso ir Snelo R^2 = 0,372 , o Nagelkerkės R^2 = 0,495. Determinacijos koeficientai yra kiek mažesni, nei tiriant pradinį modelį, bet pakankamai didelės (gerokai viršija 0,20).
- Patikrinus Kuko mato ir DFB reikšmes, įsitikiname, kad visos jos neviršija vieneto. Darome išvadą, kad modelyje išskirčių nėra.
- Klasifikacinė lentelė rodo praktiškai tokį patį gerą modelio tikimą, kaip ir pradinis modelis. Bendrasis teisingai klasifikuotų imties respondentų procentas yra 77,4

Classification Table[a]

Observed			Predicted: cntry Country: EE Estonia	Predicted: cntry Country: PT Portugal	Percentage Correct
Step 1	cntry Country	EE Estonia	95	30	76.0
		PT Portugal	26	97	78.9
	Overall Percentage				77.4

- Kintamųjų koeficientų lentelėje visi kintamieji, išskyrus konstantą, statistiškai reikšmingi. Tik konstantai Voldo kriterijaus p = 0,065 > 0,050. Tradiciškai į konstantos statistinį reikšmingumą nekreipiame dėmesio. Juolab, kad ir p reikšmė viršija 0,05 visai nedaug.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	freehms	-.760	.158	23.117	1	.000	.468	.343	.637
	happy	.300	.099	9.258	1	.002	1.350	1.113	1.637
	stfedu	-.564	.091	38.819	1	.000	.569	.476	.679
	imsb3			11.371	2	.003			
	imsb3(1)	1.661	.508	10.688	1	.001	5.263	1.945	14.243
	imsb3(2)	1.137	.426	7.132	1	.008	3.117	1.353	7.178
	Constant	1.833	.992	3.415	1	.065	6.255		

a. Variable(s) entered on step 1: freehms, happy, stfedu, imsb3.

Įsitikinome, kad naujasis modelis gerai suderintas su duomenimis. Koeficientai B prie *freehms* ir *slfedu* yra neigiami, o koeficientas prie *happy* teigiamas. Primename, kad teigiamas koeficientas rodo, jog atitinkamam kintamajam didėjant, auga ir P(Y=1) tikimybė. Ir atvirkščiai,

jeigu koeficientas neigiamas, tai auga P(Y=0) tikimybė. Todėl, didėjant *freehms* reikšmei didėja *cntry* = EE (modelio kodas 0) tikimybė. Taigi, kuo respondentas priešiškiau nusistatęs seksualinių mažumų atžvilgiu (*freehms* didesnis), tuo tikėtiniau, kad jis iš Estijos. Analogiškai darome išvadą, kad kuo respondentas palankiau vertina švietimo sistemą, tuo didesnė tikimybė, kad jis iš Estijos, o kuo laimingesnis tuo tikėtiniau, kad iš Portugalijos.

Stulpelyje **EXP(B)** pateikiamos galimybių santykių reikšmės, t.y. $e^{\hat{b}_i}$. Jau rezultatų išklotinės pradžioje buvo pažymėta, kad teigiamą požiūrį į socialines lengvatas *imsb3*=1 atitinka kategorinių kintamųjų derinys *imsb3(1)*=1, *imsb3(2)*=0, o neigiamą *imsb3(1)*=0, *imsb3(2)*=0. Atitinkamas galimybių santykis pateiktas prie *imsb3(1)*=1 ir yra lygus 5,263. Todėl galime padaryti tokią išvadą: palankus požiūris į socialines lengvatas imigrantams, lyginant su nepalankiu, padidina tikimybių santykį

$$\frac{\widehat{\mathrm{P}}(cntry = \mathrm{PT})}{\widehat{\mathrm{P}}(cntry = \mathrm{EE})}$$

5,263 karto. Labiau tampa tikėtina, kad respondentas buvo iš Portugalijos. Analogiškai labiau teigiamas, nei neigiamas požiūris į socialines lengvatas (*imsb3(1)*=0, *imsb3(2)*=1) padidina tikimybių santykį 3,117 karto.

Analogiškai interpretuojami galimybių santykiai ir tolydiems kintamiesiems. Pavyzdžiui, kiekvienas papildomas balas vertinant savo laimingumą (*happy*) padidina tikimybių santykį 1,35 karto. Labiau tampa tikėtina, kad respondentas iš Portugalijos. Kiekvienas papildomas balas vertinant švietimo būklę *stfedu*, sumažina tikimybių santykį 0,569 karto. Tampa labiau tikėtina, kad respondentas iš Estijos.

Kadangi *happy* ir *stfedu* matavimo skalės daugmaž vienodos, labai supaprastindami situaciją, galime tarti, kad *stfedu* modelyje svarbesnis, nes jo pokytis galimybių santykį pakeičia dvigubai, o *happy* pokytis galimybių santykį pakeičia tik 1,35 karto.

Grafoje **95 % C.I. for EXP(B)** yra galimybių santykių 95% pasikliautinųjų intervalų apatiniai (Lower) ir viršutiniai (Upper) rėžiai. Matome, kad intervaliniams kintamiesiems rėžiai yra pakankamai maži.

Pavyzdys. Pateiksime prognozavimo pavyzdį. Tarkime, kad respondentas labiau teigiamai, nei neigiamai įvertino imigrantų galimybes naudotis socialinėmis lengvatomis (*imsb3* = 1), savo laimingumo laipsnį įvertino 5 (*happy* = 5), požiūrį į seksualines mažumas 2 (*freehms* = 2), o šalies švietimo sistemą supeikė (*stfedu* = 4). Kokia tikimybė, kad tai respondentas iš Portugalijos? Visų pirma, užrašome bendrąjį logit finkcijos pavidalą, naudodami koeficientų reikšmes iš lentelės **Variables in the Equation** stulpelio **B**:

z = - 0,76 freehms + 0,30 happy – 0,564 stfedu +1,661 imsb3(1) + 1,137 imsb3(2) +1,833.

Įstatome į šią lygtį *imsb3(1)* = 0, *imsb3(2)* = 1 (nes būtent tokia kombinacija atitinka *imsb3* = 1, žr. aukščiau), *freehms* = 2, *happy* = 5, *stfedu* = 4, ir gauname *z* = 2,95. Tikimybė, kad atsakinėjo respondentas iš Portugalijos skaičiuojame pagal formulę (primename, kad e =*2,71828...)*

$$\widehat{\mathrm{P}}(cntry = \mathrm{PT}) = \frac{\mathrm{e}^{z}}{1 + \mathrm{e}^{z}} = \frac{\mathrm{e}^{2,95}}{1 + \mathrm{e}^{2,95}} = 0,950 \ldots$$

Taigi, tikimybė, kad atsakinėjo respondentas iš Portugalijos, yra 0,95. Tikimybė, kad respondentas yra iš Estijos lygi 0,05.

4.2.5. Rezultatų aprašymas

Trumpai reziumuodami atliktą tyrimą, galime konstatuoti, kad požiūris į seksualines mažumas, požiūris į imigrantus, požiūris į šalies švietimo sistemą ir bendrasis respondento laimingumas gerai charakterizuoja respondento šalį. Belieka išsamiau aprašyti, kaip pasireiškia tas charakterizavimas.

Pradžioje aprašome, ką tyrėme, kokius kintamuosius naudojome ir kaip juos perkodavome. Tada kompaktiškai surašome visa tai, ką išsamiai aptarėme ankstesniame skyrelyje. Pateikiame vieną iš daugybės galimų tokio aprašymo variantų.

Pritaikytas dvinarės logistinės regresijos modelis gerai tiko duomenims. Chi kvadrato statistika $\chi^2 = 115,195$; $p < 0,05$, Hosmerio – Lemešou chi kvadratas 13,132; $p = 0,107 > 0,05$. Nagelkerkės $R^2 = 0,495$. Teisingai klasifikuojama 76% Estijos respondentų ir 78,9% Portugalijos respondentų. Voldo kriterijus rodo, kad visi regresoriai yra statistiškai reikšmingi. Gautas toks modelis

$$\ln\frac{\widehat{\mathrm{P}}(cntry = \mathrm{PT})}{\widehat{\mathrm{P}}(cntry = \mathrm{EE})} = -\,0,76\, freehms \;+\; 0,30\, happy - \;0,564\, stfedu$$
$$+1,661\, imsb3(1) \;+\; 1,137\, imsb3(2) \;+ 1,833$$

Čia *imsb3(1)* = 1, *imsb3(2)* = 0 atitinka *imsb3* =0; *imsb3(1)* = 0, *imsb3(2)* = 1 atitinka *imsb3* =1; *imsb3(1)* = 0, *imsb3(2)* = 0 atitinka *imsb3* =2. Didesnės kintamųjų *freehms, stfedu* reikšmės didina tikimybę, kad respondentas yra iš Estijos. Didesnė *happy* reikšmė didina tikimybę kad respondentas yra iš Portugalijos. Galimybių santykis kintamajam *freehms* yra 0,468 (95% pasikliautinis intervalas [0,343; 0,637]), galimybių santykis kintamajam *happy* yra 1,35 ([1,113; 1,637]), o galimybių santykis kintamajam *stfedu* yra 0,569 ([0,476; 0,679]). Lyginant su nepalankia nuomone apie imigrantus (*imsb3* = 2), palanki nuomonė (*imsb* = 0) padidina galimybę Portugalijos naudai 5,263 karto (95% galimybių santykio pasikliautinis intervalas [1,945; 14,243]). Lyginant su nepalankia nuomone apie imigrantus (*imsb3* = 2), labiau palanki, nei nepalanki nuomonė (*imsb* = 1) padidina galimybę Portugalijos naudai 3,117 karto ([1,353; 7,178]).

4.2.6. SPSS komandų sintaksė

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam pradiniam logistinės regresijos modeliui. Jas galima naudoti sintaksės lange, pasirenkant opciją **Run**.

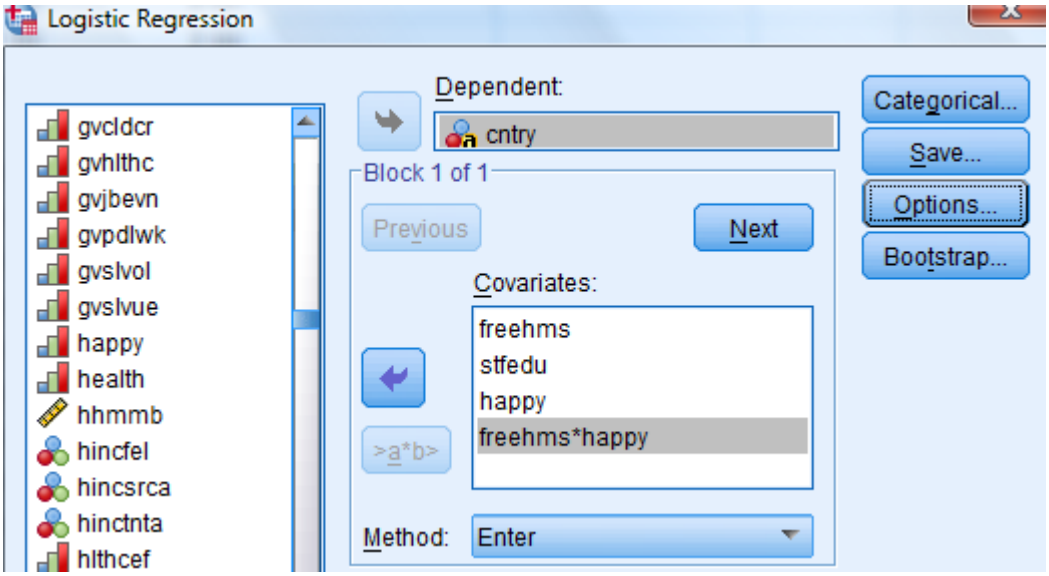
```
LOGISTIC REGRESSION VARIABLES cntry
 /METHOD=ENTER freehms happy imsclbn stfedu
 /CONTRAST (imsclbn)=Indicator
 /SAVE=COOK
 /PRINT=GOODFIT CI(95)
 /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

4.2.7. Regresorių sąveika

Regresorių tarpusavio sąveika (interakcija) vadiname situaciją, kai į modelį įtraukiama jų sandauga. Toks modelis reiškia, kad regresoriai gali vienas kitą veikti. Visos interpretacijos ir tyrimo etapai lieka nepakitę. Tereikia modelyje nurodyti norimą sąveiką. Tarkime, norime patikrinti modelį, kurį sąlygiškai užrašome taip:

$$cntry = f(stfedu,\ happy,\ freehms, happy*freehms).$$

Atsidarę **Logistic Regression** meniu po vieną perkeliame kintamuosius į **Covariates** langą, po to pažymime abudu kintamuosiu *happy* ir *freehms* ir paspaudžiame **>a*b>** .



Pasirinkę kitas opcijas taip, kaip ir aukščiau aptartuose pavyzdžiuose gauname, kad visi rodikliai, išskyrus koeficientų lentelę, rodo gerą modelio tikimą duomenims. Tuo tarpu Voldo kriterijus parodė, kad *happy* ir *freehms* sąveika yra statistiškai nereikšminga.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	freehms	-.847	.570	2.204	1	.138	.429
	stfedu	-.537	.082	43.414	1	.000	.584
	happy	.291	.216	1.824	1	.177	1.338
	freehms by happy	.013	.077	.030	1	.863	1.013
	Constant	2.746	1.663	2.725	1	.099	15.579

a. Variable(s) entered on step 1: freehms, stfedu, happy, freehms * happy .

Įdomu, kad statistiškai nereikšmingas tapo ir kintamasis *happy*. Darome išvadą, kad modelyje regresorių sąveika buvo nereikalinga.

4.2.8. Viena dominuojanti priklausomo kintamojo reikšmė

Tarkime, kad norime pagal požiūrį į seksualines mažumas (*freehms*) atskirti kasdien besimeldžiančius Estijos 20 – 40 metų respondentus *(pray = 1)* nuo to paties amžiaus niekada nesimeldžiančiųjų *(pray = 7)*. Pateikiame klasifikacinę lentelę. Matome, kad bendrasis 96,9 teisingų klasifikacijų procentas visai nereiškia gero klasifikavimo. Tiesiog imtyje buvo tik 5 kasdien besimeldžiantys Estijos gyventojai, kurių nė vienas nebuvo atpažintas. Buvo pažeistas reikalavimas, kad kiekviena priklausomo kintamojo reikšmė turi sudaryti bent penktadalį imties

Classification Table[a]

			Predicted		
			pray How often pray apart from at religious services		Percentage Correct
	Observed		1 Every day	7 Never	
Step 1	pray How often pray apart from at religious services	1 Every day	0	5	.0
		7 Never	0	158	100.0
	Overall Percentage				96.9

4.2.9. Multikolinearumas

Pateiksime vieną pavyzdį, kai dvinarės logistinės regresijos modelyje pasireiškia multikolinearumas. Tirsime Estijos 20 – 40 metų vyrų požiūrį į Europarlamentą. Naudosime 2008 metų Europos Sąjungos socialinio tyrimo Estijos duomenis. Tyrime naudosime tokius kintamuosius:

trstprl (pasitikėjimas šalies parlamentu), *trstlgl* (pasitikėjimas teisine sistema), *trstplc* (pasitikėjimas policija), *trstplt* (pasitikėjimas politikais), *trstprt* (pasitikėjimas partijomis), *trstun* (pasitikėjimas Jungtinėmis Tautomis, *trstep* (pasitikėjimas Europarlamentu).

Visi kintamieji matuojami 10 balų skale nuo 0 (itin nepasitikiu) iki 9 (itin pasitikiu). Kintamąjį *trstep* keičiame dvireikšmiu kintamuoju *zzzz* (*zzz = 0*, kai *trstep* ≤ *2* ir *zzzz = 1*, kai *trsep* ≥ *8*). Taigi *zzzz* rodo požiūrio į Europarlamentą kraštutinumus. Sudarome kintamajam *zzzz* logistinės regresijos modelį, regresoriais imdami visus likusius kintamuosius. Visi rodikliai, išskyrus koeficientų lentelę tiesiog fantastiški:

- Chi kvadrato p = 0,000, Hosmerio – Lemešou chi kvadrato p = 1, Nagelkerkės R^2=1.
- Kvalifikacinė lentelė rodo, kad modelis 100 % imties stebėjimų klasifikuoja teisingai. Beje, abiejų kategorijų stebėjimų buvo panašus skaičius (28 ir 19).

Tuo tarpu iš koeficientų lentelės **Variables in the Equation** rezultatų matyti, kad visi kintamieji yra statistiškai nereikšmingi. Be to, visos Voldo statistikos standartinės paklaidos **S.E.** yra tiesiog milžiniškos. Visuose aukščiau nagrinėtuose modeliuose jos retai kada viršydavo vienetą, o čia visos didesnės už 1900. Galų gale, neigiami koeficientų **B** ženklai irgi kelia įtarimą. Išeitų, kad kuo labiau respondentas pasitiki savo šalies parlamentu, tuo mažesnė tikimybė, kad jis pasitikės Europarlamentu. Visi šie požymiai rodo, kad modelyje yra multikolinearumo problema. Darome išvadą, kad modelis blogas. **Nepaisant visų kitų gerų rodiklių**.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	trstlgl	4.366	1917.425	.000	1	.998	78.691
	trstplc	12.254	2355.696	.000	1	.996	209789.593
	trstplt	-2.964	4088.301	.000	1	.999	.052
	trstprl	-4.368	2761.944	.000	1	.999	.013
	trstprt	11.969	5018.000	.000	1	.998	157850.256
	trstun	9.491	2069.086	.000	1	.996	13245.274
	Constant	-182.445	20226.246	.000	1	.993	.000

a. Variable(s) entered on step 1: trstlgl, trstplc, trstplt, trstprl, trstprt, trstun.

Beje, sumodeliavus *zzz* vien pagal *trstprl*, atitinkamo koeficiento ženklas tampa teigiamu. Taigi multikolinearus modelis iš tikrųjų iškreipė kintamųjų priklausomybes.

4.2.10. Žingsninė logistinė regresija

Pastabose minėjome, kad pagrindinį dėmesį sutelksime ties logistinės regresijos modelio tyrimu, kai kintamuosius šalina arba keičia pats tyrėjas. SPSS yra realizuota galimybė, kai iš pateiktų regresorių

sudaromas geriausias modelis. Tai atliekama formaliai, atsižvelgus į jų statistinį reikšmingumą, bet ne į prasmę. Yra du principiniai žingsninės regresijos metodai. Kaupiamojoje (angl. forward) žingsninėje regresijoje pradedama nuo trivialaus modelio, sudaryto iš konstantos. Šis modelis palaipsniui pildomas kitais regresoriais. Alternatyviai (angl. backward) šalinamojoje žingsninėje regresijoje pradedamai nuo modelio su visais regresoriais, palaipsniui vis mažinant regresorių skaičių. Žinoma, tai tik principiniai metodų skirtumai. Iš tikro, dalis regresorių gali būti viename žingsnyje pašalinti, o kitame grąžinti atgal. SPSS yra realizuota keletas tokių algoritmų. Dažniausiai atsižvelgiama į didžiausio tikėtinumo funkcijos reikšmes. Pademonstruosime vieną iš tokių metodų.

Pasirinksime pradinį tirtą modelį, t.y. tirsime, ar skiriasi 20 – 30 m vyrai iš Portugalijos nuo tokio paties amžiaus vyrų iš Estijos, jeigu atsižvelgsime į kintamųjų *stfedu, happy, freehms* ir *imsclbn* reikšmes. Sąlyginai modelį galima užrašyti taip:

$$cntry = f(stfedu, happy, freehms, imsclbn).$$

Visas opcijas pasirenkame taip, kaip aprašyta 3.2.2 skyrelyje. Vienintelis skirtumas tas, kad pagrindiniame meniu pasirenkame **Method: Forward LR**.

Pasirinkome metodą, kai pradinis modelis vis papildomas naujais regresoriais. Rezultatų išklotinė prasideda jau matyta informacija apie kintamųjų kodus ir klasifikavimą, kai aklai spėjama priklausomybė Estijai. Po to pateikiama informacija apie pradinį modelį, kurį sudaro tik konstanta. Tai *a priori* blogas modelis, nes jis teigia, kad joks regresorius nepadės rasti skirtumų tarp portugalų ir estų. Iš tikro įsitikiname, kad jis statistiškai nereikšmingas (Voldo statistikos $p = 0{,}899 > 0{,}05$).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.016	.127	.016	1	.899	.984

Lentelėje **Variables not in the Equation** išvardyti visi regresoriai, kartu pateikiant jų Roa statistikos (Voldo statistikos tam tikras analogas) reikšmes stulpelyje **Score** ir *p* reikšmes stulpelyje **Sig.** Visi regresoriai statistiškai reikšmingi ir juos galima bandyti įtraukti į modelį.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	freehms	34.961	1	.000
		happy	7.024	1	.008
		stfedu	46.378	1	.000
		imsclbn	39.756	4	.000
		imsclbn(1)	14.248	1	.000
		imsclbn(2)	.028	1	.866
		imsclbn(3)	7.143	1	.008
		imsclbn(4)	31.529	1	.000
	Overall Statistics		99.531	7	.000

Iš kitų lentelių matyti, kad regresoriai į modelį buvo įtraukiami keturiais etapais (keturi regresijos žingsniai). Po kiekvieno žingsnio svarbiausios modelio charakteristikos rodė, kad naujasis modelis gerai tinka duomenims. Chi kvadrato statistikos *p* reikšmės < 0,05; Hosmerio – Lemešou statistikos *p* reikšmės > 0,05. Determinacijos pseudokoeficientų reikšmės vis didėjo.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.313	6	.634
2	10.556	8	.228
3	4.928	8	.765
4	5.846	8	.665

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	293.600[a]	.183	.244
2	251.174[b]	.312	.416
3	232.530[b]	.361	.482
4	221.738[b]	.389	.518

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	50.185	1	.000
	Block	50.185	1	.000
	Model	50.185	1	.000
Step 2	Step	42.426	1	.000
	Block	92.611	2	.000
	Model	92.611	2	.000
Step 3	Step	18.644	4	.001
	Block	111.255	6	.000
	Model	111.255	6	.000
Step 4	Step	10.792	1	.001
	Block	122.047	7	.000
	Model	122.047	7	.000

Jungtinė klasifikacinė lentelė rodo, kad naujų kintamųjų įtraukimas, pradedant antruoju žingsniu, klasifikavimą nelabai gerino.

Classification Table[a]

Observed			Predicted: cntry Country: EE Estonia	Predicted: cntry Country: PT Portugal	Percentage Correct
Step 1	cntry Country	EE Estonia	92	33	73.6
		PT Portugal	40	83	67.5
	Overall Percentage				70.6
Step 2	cntry Country	EE Estonia	98	27	78.4
		PT Portugal	30	93	75.6
	Overall Percentage				77.0
Step 3	cntry Country	EE Estonia	96	29	76.8
		PT Portugal	27	96	78.0
	Overall Percentage				77.4
Step 4	cntry Country	EE Estonia	97	28	77.6
		PT Portugal	26	97	78.9
	Overall Percentage				78.2

Galų gale gauname visų žingsnių logistinės regresijos modelius. Matome, kad kiekvienas įtrauktas tolydus kintamasis buvo statistiškai reikšmingas. Kadangi kategorinio kintamojo *imsclbn* kai kurie pseudokintamieji irgi buvo statistiškai reikšmingi, tai ir jis modelyje buvo paliktas.

Rezultatų išklotinėje dar pateikiama lentelė **Model if Term Removed**, kuri rodo, kad visi regresoriai modelyje reikalingi (**Sig. of the Change** visiems regresoriams yra < 0,05).

Galutinis, tokiu būdu gautas modelis, sutampa su 4.2.2 skyrelyje nagrinėtuoju modeliu, kurį sąlyginai (tikslus užrašas užimtų daug vietos, nes reikėtų išvardyti visas *imsclbn* kategorijas) galima užrašyti taip:

cntry = f(freehms, happy,stfedu, imsclbn).

Šiame modelyje liko statistiškai nereikšmingų pseudokintamųjų. Žinome, kad toks modelis nėra geras, nes duomenyse tiesiog nėra pakankamai daug visų *imsclbn* kategorijų reikšmių. Mes modelį patobulinome, transformuodami (perkoduodami) kintamąjį *imsclbn*. Žingsninė regresija čia pagelbėti negali.

Kita vertus, mes galime panaudoti žingsninę regresiją visų modelių analizei ir apsiriboti antruoju modeliu, kuris yra visai neblogas:

cntry = 5,145 -0,899*freehms* – 0,539*stfedu*.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	stfedu	-.463	.074	39.518	1	.000	.630	.545	.727
	Constant	2.511	.431	33.974	1	.000	12.317		
Step 2[b]	freehms	-.899	.155	33.588	1	.000	.407	.300	.551
	stfedu	-.539	.084	40.802	1	.000	.584	.495	.688
	Constant	5.145	.711	52.408	1	.000	171.531		
Step 3[c]	freehms	-.781	.156	25.149	1	.000	.458	.338	.622
	stfedu	-.511	.087	34.832	1	.000	.600	.506	.711
	imsclbn			15.886	4	.003			
	imsclbn(1)	1.990	1.197	2.761	1	.097	7.312	.700	76.418
	imsclbn(2)	.260	1.116	.054	1	.816	1.296	.146	11.552
	imsclbn(3)	.511	1.037	.243	1	.622	1.667	.218	12.732
	imsclbn(4)	-.737	1.081	.465	1	.495	.479	.058	3.981
	Constant	4.378	1.233	12.608	1	.000	79.652		
Step 4[d]	freehms	-.730	.160	20.938	1	.000	.482	.352	.659
	happy	.314	.099	10.025	1	.002	1.369	1.127	1.663
	stfedu	-.564	.092	37.329	1	.000	.569	.475	.682
	imsclbn			15.751	4	.003			
	imsclbn(1)	1.849	1.284	2.075	1	.150	6.354	.513	78.647
	imsclbn(2)	-.039	1.197	.001	1	.974	.962	.092	10.046
	imsclbn(3)	.195	1.123	.030	1	.862	1.215	.134	10.977
	imsclbn(4)	-1.057	1.168	.819	1	.366	.347	.035	3.430
	Constant	2.599	1.422	3.341	1	.068	13.445		

a. Variable(s) entered on step 1: stfedu.

b. Variable(s) entered on step 2: freehms.

c. Variable(s) entered on step 3: imsclbn.

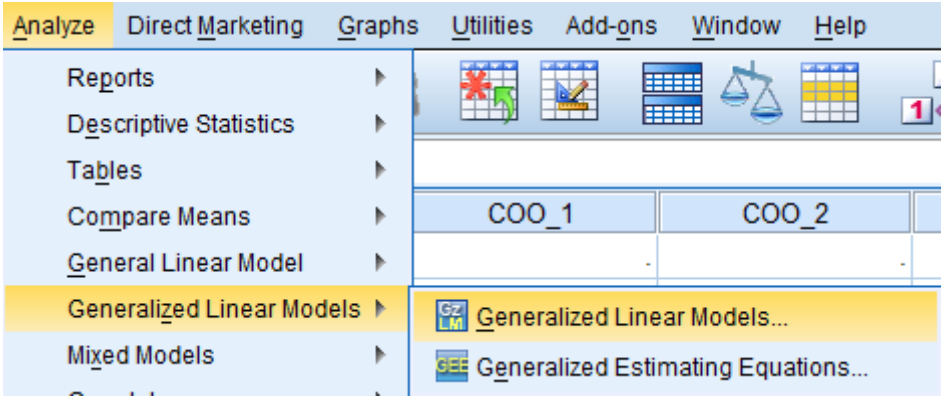
d. Variable(s) entered on step 4: happy.

4.2.11. Apibendrintas tiesinis modelis su logit jungtimi

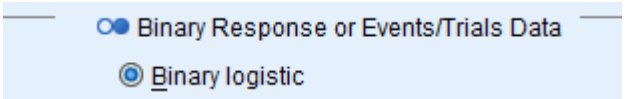
Formuluodami rekalavimus duomenims (žr., 4.1.4 skyrelį), minėjome, kad logistinės regresijos modelis naudojamas ir tada, kai dauguma regresorių kategoriniai. Tiesiog tokį modelį labiau įprasta vadinti apibendrintuoju tiesiniu modeliu su logit jungtimi, o, jį aprašant, pateikti *deviacijos ir jos laisvės laipsnių santykį*. Kiekvieną logistinės regresijos modelį galima tirti, kaip atskirą apibendrintojo tiesinio modelio atvejį, tiesiog tai nėra patogu. Parodysime tai, tirdami 4.2.4 skyrelio modelį:

cntry = f(stfedu, happy, freehms, imsb3).

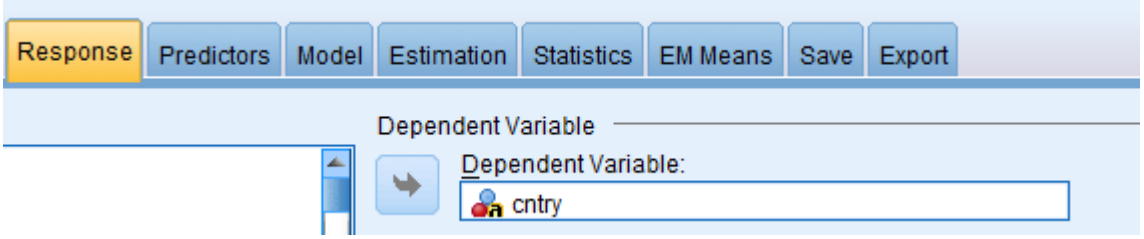
Tirsime 20 – 30 metų vyrus (atrenkame *agea* ≥ 20 & *agea* ≤ 30 & *gndr* = 1). Pasirenkame **Analyze → Generalized Linear Models → Generalized Linear Models…**



Lange **Type of Model** pažymime **Binary logistic**.

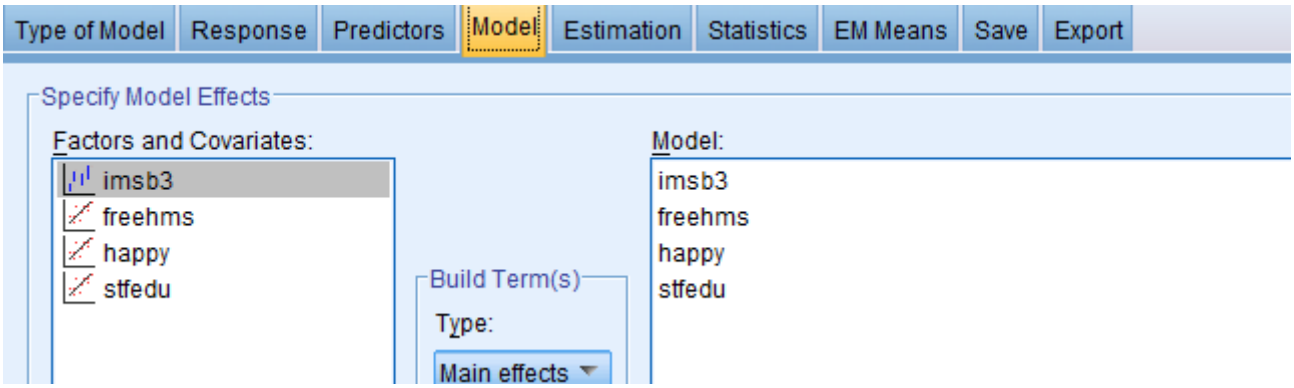


Renkamės **Response**. Į langelį **Dependent Variable** įkeliame *cntry*.



Renkamės **Predictors**. Į langelį **Factors** įkeliame *imsb3*, į langelį **Covariates** įkeliame regresorius *freehms, happy, stfedu*.

Renkamės **Model**. Perkeliame visus regresorius į lauką **Model.**



Renkamės Statistics**. Papildomai pažymime** Include exponential parameter estimates**.** Renkamės **Save** ir pažymime **Predicted category** ir **Cook's distance**. Spaudžiame **OK**.

Save	Item to Save	Variable Name
☐	Predicted value of mean of response	MeanPredicted
☐	Lower bound of confidence interval for mean of response	CIMeanPredicte
☐	Upper bound of confidence interval for mean of response	CIMeanPredicte
☑	Predicted category	PredictedValue
☐	Predicted value of linear predictor	XBPredicted
☐	Estimated standard error of predicted value of linear predictor	XBStandardErro
☑	Cook's distance	CooksDistance

Rezultatų išklotinėje nesunkiai randame dalį tos informacijos, kurią jau aprašėme 4.2.4 skyrelyje. Pavyzdžiui, modelio parametrai (kategorinis kintamasis *imsb3* keičiamas pseudokintmaisiais), jų statistinis reikšmingumas, galimybių santykiai ir galimybių santykių pasikliautinieji intervalai surašyti lentelėje **Parameter Estimates**.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-1.833	.9922	-3.778	.111	3.415	1	.065	.160	.023	1.118
[imsb3=.00]	-1.661	.5080	-2.656	-.665	10.688	1	.001	.190	.070	.514
[imsb3=1.00]	-1.137	.4256	-1.971	-.303	7.132	1	.008	.321	.139	.739
[imsb3=2.00]	0[a]	.	.	.	.	.	.	1	.	.
freehms	.760	.1581	.450	1.070	23.117	1	.000	2.139	1.569	2.916
happy	-.300	.0985	-.493	-.107	9.258	1	.002	.741	.611	.899
stfedu	.564	.0906	.387	.742	38.819	1	.000	1.759	1.472	2.100
(Scale)	1[b]									

Dependent Variable: Country

Model: (Intercept), imsb3, freehms, happy, stfedu

Didžiausio tikėtinumo chi kvadrato reikšmė ir jos *p* reikšmė (0,00) pateikiama lentelėje **Omnibus Test.** Kaip jau buvo nurodyta 4.2.4 skyrelyje visos šios charakteristikos rodo gerą modelio tikimą.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
115.195	5	.000

Rezultatų išklotinėje nėra determinacijos pseudokoeficientų (nedidelė bėda) ir klasifikacinės lentelės (rimtas trūkumas). Užtat lentelėje **Goodness of Fit** pateikiama deviacija. Deviacija parodo, kiek tiriamas modelis skiriasi nuo pilnai duomenis aprašančio (bet labai sudėtingo ir todėl naudoti netinkamo) modelio. Dar reikia atsižvelgti į imties didumą ir regresorių skaičių, kuriuos nusako vadinamieji laisvės laipsniai (angl. *degrees of freedom, df*).

Gerai duomenims tinkančio apibendrintojo tiesinio modelio deviacijos ir laisvės laipsnių santykis turi būti artimas vienetui.

Deviacijos ir laisvės laipsnių santykį surandame grafoje **Value/df**. Šis santykis lygus 0,976. Tai patvirtina gerą modelio tikimą duomenims.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	180.584	185	.976
Scaled Deviance	180.584	185	
Pearson Chi-Square	191.020	185	1.033
Scaled Pearson Chi-Square	191.020	185	
Log Likelihood[a]	-100.038		
Akaike's Information Criterion (AIC)	212.076		
Finite Sample Corrected AIC (AICC)	212.424		
Bayesian Information Criterion (BIC)	233.156		
Consistent AIC (CAIC)	239.156		

Dependent Variable: Country

Model: (Intercept), imsb3, freehms, happy, stfedu

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Rezultatų išklotinėje pateikiamas ir didžiausio tikėtinumo santykio chi kvadrato statistikos reikšmės ir *p* reikšmės kiekvienam regresoriui. Tai gera alternatyva Voldo kriterijui, pateiktajam lentelėje **Parameter Estimates**. Įsitikiname, kad visi regresoriai yra statistiškai reikšmingi, taigi svarbūs modeliui.

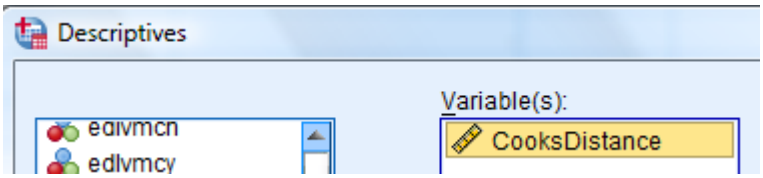
Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	9.067	1	.003
imsb3	11.371	2	.003
freehms	23.117	1	.000
happy	9.258	1	.002
stfedu	38.819	1	.000

Dependent Variable: Country

Model: (Intercept), imsb3, freehms, happy, stfedu

Mes parinkome tokias SPSS **Save** opcijas, kad duomenyse atsirado du nauji stulpeliai: *CooksDistance* su Kuko mato reikšmėmis ir *PredictedValue* su prognozuojama respondento šalimi kiekvienam respondentui. Norėdami įsitikinti, kad duomenyse nėra išskirčių, renkamės **Analyze → Descriptive Statistics → Descriptives**. Perkeliame *CooksDistance* į laukelį **Variable(s)** ir spaudžiame **OK.**

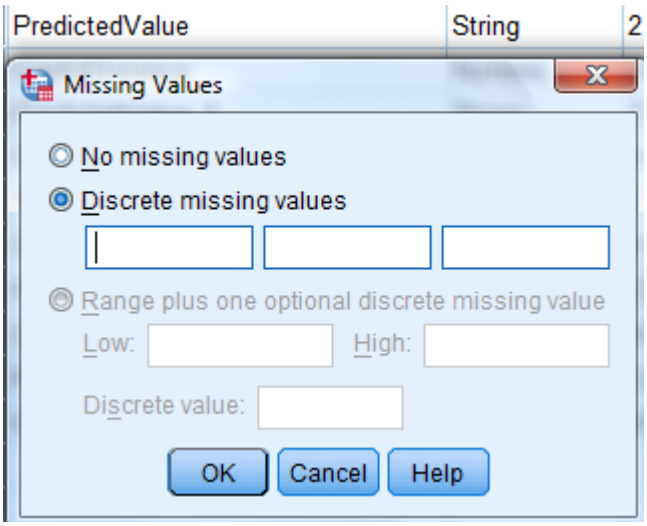


Matome, kad maksimali Kuko mato reikšmė neviršija 1, todėl darome išvadą, kad duomenyse išskirčių nėra.

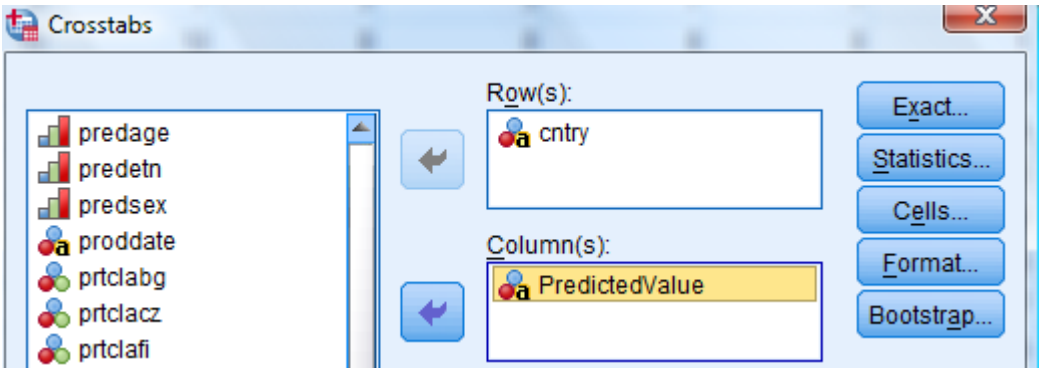
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CooksDistance_4 Cook's Distance	248	.000	.050	.00439	.008036
Valid N (listwise)	248				

Gauti klasifikacinę lentelę yra kiek sunkiau, nes SPSS į prognozuojamas reikšmes įtraukia ir tuščius langelius, t.y. tuos 29 respondentus, kurių duomenys buvo nepilni ir modeliuojant nebuvo naudojami. Todėl duomenyse kairiame apatiniame kampe paspaudžiame Variable View, kintamųjų sąrašo apačioje randame *PredictedValue*, jam atidarome langelį Missing Values ir pažymime Discrete missing values. Su pele „kliktelime“ ant pirmo langelio ir paspaudžiame tarpo klavišą (įvedame tučią lauką). Ekrane niekas nepasikeičia, tik pelės kursorius kiek atsitraukia nuo langelio krašto.



Paspaudžiame OK. Grįžtame į duomenų langą Data View. Pasirenkame Analyze → Descriptive Statistics → Crosstabs. Į langelį Row(s) perkleiame kintamąjį *cntry*, į langelį Column(s) – kintamąjį *PredictedValue*.



Renkamės **Cells** ir pažymime **Row**. Spaudžiame **Continue** ir **OK**. Gauname klasifikacinę lentelę.

cntry Country * PredictedValue Predicted Category Value Crosstabulation

			PredictedValue Predicted Category Value		
			EE Estonia	PT Portugal	Total
cntry Country	EE Estonia	Count	95	30	125
		% within cntry Country	76.0%	24.0%	100.0%
	PT Portugal	Count	26	97	123
		% within cntry Country	21.1%	78.9%	100.0%
Total		Count	121	127	248
		% within cntry Country	48.8%	51.2%	100.0%

Taikant modelį, teisingai klasifikuojama 76% Estijos respondentų ir 78,9% Portugalijos respondentų.

Skyrelio pabaigoje dar kartą pakartosime savo asmeninį požiūrį, kad SPSS programa logistinę regresiją daryti, naudojantis apibendrintuoju tiesiniu modeliu su logit jungtimi, verta tik tada, kai turime daug kategorinių regresorių. Šiaip jau daug lengviau tyrimą atlikti su **Regression → Binary Logistic**.

4.3. Dvinarė logistinė regresinė analizė su STATA

4.3.1. Duomenys

Kaip atlikti dvinarę logistinę regresinę analizę parodysime, tirdami Europos Sąjungos socialinio tyrimo (European Social Survey) Lietuvos ir Portugalijos duomenis **ESS4LT1** ir **ESS4PT**. Tirsime 20 – 30 metų vyrus. Tyrime naudosime tokius kintamuosius:

- *cntry* – šalies kodas (LT – Lietuva, PT – Portugalija),
- *stfgov* – respondento požiūris į savo šalies vyriausybę (0 – itin bloga, ..., 9 – itin gera),
- *freehms* – gėjai ir lesbietės gali gyventi, kaip nori (1 – pilnai sutinku,..., 5 – visiškai nesutinku),
- *imsclbn* – nuo kada imigrantai gali pilnai naudotis šalies socialinėmis lengvatomis (matuojamas nominalia skale: 1 – iškart po atvykimo, 2 – po vienerių metų, nepriklausomai nuo to, ar dirbo, 3 – po vienerių darbo ir mokesčių mokėjimo metų, 4 – tapus piliečiu, 5 – niekada negalės naudotis visomis lengvatomis).

Duomenų atrankai naudosime kintamuosius: *agea* – amžių ir *gndr* – lytį (1 – vyr., 2 - mot.). Grafikams pravers respondento identifikacijos numeris *idno.*

Sujungę SPSS failus, išsaugome juos *dta* formatu. Atidarę failą STATA programa, visų pirma paliekame tik tuos kintamuosius, kurių reikės tyrime:

```
keep agea gndr cntry stfgov freehms imsclbn idno
```

Paliksime tyrime reikiamo amžiaus ir lyties respondentus. Kad nekiltų problemų su neteisingai išsaugotais praleistų reikšmių kodais, dar įvesime apribojimus kitų kintamųjų reikšmėms:

```
keep if agea >=20 & agea <= 30 & gndr == 1 &  stfgov <= 9 & freehms <=5 & imsclbn <=5
```

Galima pasižiūrėti informaciją apie failo kintamuosius. Įvykdome

```
describe cntry stfgov freehms  imsclbn
```

```
. describe cntry stfgov freehms  imsclbn

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
cntry           str2   %2s                    Country
stfgov          double %10.0g      stfgov     How satisfied with the national
                                                government
freehms         double %10.0g      freehms    Gays and lesbians free to live
                                                life as they whish
imsclbn         double %10.0g      imsclbn    When should immigrants obtain
                                                rights to social
                                                benefits/services
```

Visi kintamieji yra skaitiniai, o kintamasis *cntry* - simbolinis. Perkoduosime jį į skaitinį kintamąjį (kintamojo *cntry* reikšmės taps naujojo kintamojo *zzz* reikšmių komentarais (angl. labels):

```
encode cntry, generate(zzz)
```

Pagal nutylėjimą STATA programa rodo kintamojo reikšmių komentarus. Todėl paprašę dažnių lentelės, gauname informaciją apie tai, kokioms šalims priklausė respondentai (šalių kodai tapo kintamojo komentarais):

```
tabulate zzz
```

```
. tabulate zzz

    Country |      Freq.     Percent        Cum.
------------+-----------------------------------
         LT |        161       56.69       56.69
         PT |        123       43.31      100.00
------------+-----------------------------------
      Total |        284      100.00
```

Ši lentelė bus naudinga, kai reikės tikrinti klasifikavimo gerumą, tačiau ji nepateikia informacijos apie kintamojo *zzz* reikšmes. Todėl papildome tabulate komandą opcija nolabel.

```
tabulate zzz, nolabel
```

```
. tabulate zzz, nolabel

    Country |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        161       56.69       56.69
          2 |        123       43.31      100.00
------------+-----------------------------------
      Total |        284      100.00
```

Matome, kad *zzz* įgyja reikšmes 1 ir 2. Nors *zzz* yra dvireikšmis kintamasis, logistinei regresijai jis netinka, nes STATA programa reikalauja, kad viena iš kintamojo reikšmių būtų 0. Todėl sukuriame naują kintamąjį, kuris jau įgis reikšmes 0 ir 1:

```
gen country = zzz - 1
```

Pridėsime naujojo kintamojo reikšmėms komentarus: 0 – LT, 1 – PT. Tam reikia nuosekliai įvykdyti dvi komandas:

```
label define zl  0 "LT" 1"PT"
label values country zl
```

Kintamasis *imsclbn* yra kategorinis.

> Prieš įtraukiant kategorinį kintamąjį į logistinės regresijos modelį, reikia įsitikinti, kad kiekvienai jo kategorijai yra pakankamai daug stebėjimų.

Patikriname, kokias reikšmes įgyja *imsclbn*:

```
tabulate country imsclbn
```

	When should immigrants obtain rights to social benefits/services					
country	Immediate	After a y	After wor	Once they	They shou	Total
LT	10	9	58	75	9	161
PT	22	12	73	13	3	123
Total	32	21	131	88	12	284

Matome, kad Lietuvos imtyje mažai respondentų rinkosi 1 – 2 ir 5. Todėl regresorių *imsclbn* keisime kategoriniu kintamuoju, įgyjančiu mažiau reikšmių. Sukursime du tokius kintamuosius: *imsb3* (*imsb3* = 0, kai *imsclbn* = 1 arba 2; *imsb3* = 1, kai *imsclbn* = 3; *imsb3* = 2, kai *imsclbn* = 4 arba 5) ir *imsb2* (*imsb2* = 0, kai *imsclbn* = 1 , 2 arba 3; *imsb2* = 1, kai *imsclbn* = 4 arba 5):

```
recode  imsclbn (1 2 = 0) (3 = 1) (4 5 = 2), gen(imsb3)
recode  imsclbn (1/3 = 0)  (4 5 = 1), gen(imsb2)
```

Mažesnės *imsb2, imsb3* reikšmės rodo palankesnę nuostatą imigrantų atžvilgiu. Įsitikiname, kad dabar yra pakankamai daug stebėjimų[13] kiekvienai regresorių kategorijai. Įvykdome

```
tabulate country imsb2
tabulate country imsb3
```

[13] Ką reiškia pakankamai daug, sprendžia pats tyrėjas. Mūsų nuomone, jeigu kažkokį atsakymą renkasi mažiau nei 10 % respondentų, vertėtų pamąstyti apie kintamojo perkodavimą, paliekant mažiau kategorijų.

country	RECODE of imsclbn (When should immigrants obtain rights to social benefits/servi 0	1	Total
LT	77	84	161
PT	107	16	123
Total	184	100	284

country	RECODE of imsclbn (When should immigrants obtain rights to social benefits/servi 0	1	2	Total
LT	19	58	84	161
PT	34	73	16	123
Total	53	131	100	284

4.3.2. Modelio tyrimas

Logistinę regresiją STATA programa galima atlikti, naudojant *logit* ir *logistic* komandas. Didelio skirtumo tarp šių komandų nėra, todėl apsiribosime komanda *logit*. Tirsime modelį *country = f(stfgov, freehms, imsb2).* Modelyje yra vienas dvireikšmis regresorius *imsb2* (reikšmės 0 ir 1). Įvykdome komandą:

```
logit  country freehms stfgov imsb2
```

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(3)      =     118.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -134.82313                       Pseudo R2       =     0.3061
```

country	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
freehms	-.81604	.151165	-5.40	0.000	-1.112318	-.5197622
stfgov	.3504268	.0753233	4.65	0.000	.2027959	.4980578
imsb2	-1.708391	.3430608	-4.98	0.000	-2.380778	-1.036004
_cons	1.306911	.427897	3.05	0.002	.4682487	2.145574

Chi kvadrato kriterijaus statistika ir *p* reikšmė pateikiamos lentelės viršuje dešinėje pusėje. Matome, kad chi kvadrato statistka yra didelė (lygi 118,96) ir statistiškai reikšminga (jos *p* reikšmė 0,000..). Faktiškai tai informacija, kad su pasirinktais regresoriais modelis gerai tinka duomenims, nors neaišku, ar visi kintamieji modelyje reikalingi. Kuo chi kvadrato reikšmė didesnė, tuo labiau tikėtina, kad logistinės regresijos modelis tinka. Ji ypač informatyvi tada, kai turime labai daug

duomenų. Tada net ir statistiškai reikšminga, bet maža chi kvadrato reikšmė, kelia įtarimų dėl modelio gerumo.

Determinacijos pseudokoeficiento reikšmė nėra didelė (**Pseudo R2 = 0.3061**), bet logistinei regresijai šis rodiklis yra nelabai svarbus. Juolab, kad vis tiek viršija 0,20 – tradicinę determinacijos koeficientams taikomą tinkamumo ribą. STATA pateikia Makfadeno determinacijos pseudokoeficientą.

Užrašysime modelį, įstatydami koeficientų įverčius (exp{x}, tai trumpas e^x žymėjimas).

$$\frac{\widehat{\mathrm{P}}(country = PT)}{\widehat{\mathrm{P}}(country = LT)} = \exp\{1{,}31 + 0{,}35\, stfgov - 0{,}81\, freehms\ - 1{,}71 imsb2\}.$$

Įsitikiname modelio logiškumu. Didėjant skliaustuose esančiam reiškiniui (tikimybių santykio logit funkcijai), kartu didėja ir tikimybė, kad respondentas iš Portugalijos. Ir priešingai, mažėjant skliaustuose esančiam reiškiniui, didėja tikimybė, kad respondentas iš Lietuvos. Koeficientas prie *freehms* ir yra neigiamas, o koeficientas prie *stfgov* teigiamas. Todėl, didėjant *freehms* reikšmei didėja tikimybė, kad respondentas iš Lietuvos. Taigi, kuo respondentas priešiškiau nusistatęs seksualinių mažumų atžvilgiu (*freehms* didesnis), tuo tikėtiniau, kad jis iš Lietuvos. Analogiškai darome išvadą, kad kuo repondentas palankiau vertina vyriausybę, tuo didesnė tikimybė, kad jis iš Portugalijos. Koeficientas prie *imsb2* yra neigiamas. Kai įstatome vienetą (*imsb2* = 1 atitinka neigiamą požiūrį į imigrantus), visas reiškinys sumažėja ir padidėja tikimybė, kad respondentas iš Lietuvos.

Modelio regresorių galimybių santykiai gaunami, įvykdžius komandą:

```
logit  country freehms stfgov imsb2, or
```

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(3)      =     118.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -134.82313                       Pseudo R2       =     0.3061

------------------------------------------------------------------------------
     country | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     freehms |   .4421792    .066842    -5.40   0.000      .328796     .594662
      stfgov |   1.419673   .1069345     4.65   0.000     1.224822    1.645522
       imsb2 |   .1811571   .0621479    -4.98   0.000     .0924786    .3548699
------------------------------------------------------------------------------
```

Kintamojo *imsb2* galimybių santykis yra lygus 0,181. Todėl galime padaryti tokią išvadą: nepalankus požiūris į socialines lengvatas imigrantams (*imsb2* = 2), lyginant su palankiu, sumažina santykį

$$\frac{\widehat{\mathrm{P}}(country = \mathrm{PT})}{\widehat{\mathrm{P}}(country = \mathrm{LT})}$$

0,181 karto. Kaip jau minėta aukščiau, labiau tampa tikėtina, kad respondentas buvo iš Lietuvos. Analogiškai interpretuojami galimybių santykiai ir tolydiems kintamiesiems. Pavyzdžiui, kiekvienas papildomas balas, palankiau vertinant vyriausybę (*stfgov*) padidina tikimybių santykį 1,42 karto. Labiau tampa tikėtina, kad respondentas iš Portugalijos. Kiekvienas papildomas balas neigiamiau vertinant seksualines mažumas *freehms*, sumažina tikimybių santykį 0,44 karto. Tampa labiau tikėtina, kad respondentas iš Lietuvos. Be pačių galimybių santykių lentelėje dar pateikiami ir jų 95% pasikliautinųjų intervalų apatiniai ir viršutiniai rėžiai. Matome, kad intervaliniams kintamiesiems rėžiai yra pakankamai maži.

Norėdami gauti klasifikacinę lentelę, įvykdome komandą:

```
lstat
```

```
Logistic model for country

                 -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        89            33  |        122
     -     |        34           128  |        162
-----------+--------------------------+-----------
   Total   |       123           161  |        284

Classified + if predicted Pr(D) >= .5
True D defined as country != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   72.36%
Specificity                     Pr( -|~D)   79.50%
Positive predictive value       Pr( D| +)   72.95%
Negative predictive value       Pr(~D| -)   79.01%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   20.50%
False - rate for true D         Pr( -| D)   27.64%
False + rate for classified +   Pr(~D| +)   27.05%
False - rate for classified -   Pr( D| -)   20.99%
--------------------------------------------------
Correctly classified                        76.41%
--------------------------------------------------
```

D kategorija atitinka Portugaliją (*country* $\neq$ 0). Matome, kad iš 123 Portugalijos atstovų, teisingai atpažinti 89. Tai sudaro 72,36 % teisingų klasifikavimų. Analogiškai nustatome, kad iš 161 Lietuvos atstovų teisingai klasifikuoti 128 (79,5%). Bendrasis teisingų klasifikavimų procentas (89+128 iš 123+161) yra 76,41%. Ankstesniame skyrelyje buvo nustayta, kad imtyje respondentai iš Portugalijos sudaro 43,31% , o iš Lietuvos – 56,69%. Būtent toks būtų ir prognozės tiklsumas, jeigu visą laiką, kaip papūgos kartotume “PT” (“LT”). Taigi, taikant logistinę regresiją, klasifikavimas labai pagerėja. Vienas svarbiausių reikalavimų keliamų geriems logistinės regresijos modeliams patenkintas.

Norėdami gauti informaciją apie Hosmerio – Lemešou statistiką, įvykdome komandą:

```
lfit, group(10)
```

```
Logistic model for country, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       284
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        15.44
                 Prob > chi2 =         0.0512
```

Gerai, kai Hosmerio – Lemešou statistikos p ≥ 0,05. Šiuo atveju p = 0,0512. Šis rodiklis irgi rodo priimtiną modelio tikimą duomenims. Šiaip jau Hosmerio – Lemešou kriterijus yra pagalbinis ir taikomas tik nedidelėms imtims. Imtis n = 284, nėra itin maža, todėl galima buvo apsieiti ir be jo.

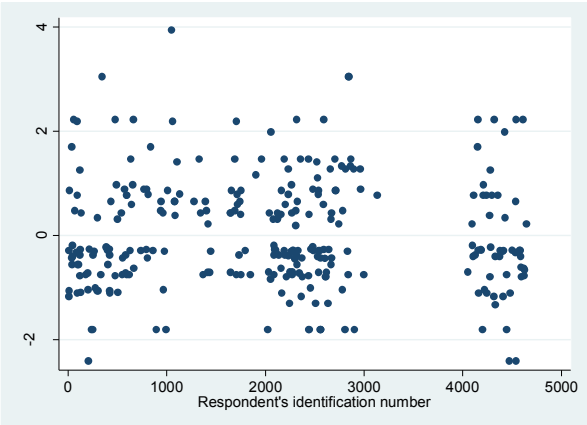
Išskirčių nustatymas STATA programoje nėra labai gerai realizuotas ir taikomas retai (išskirtis turi būti labai „piktybinė“, kad logistinė regresija į ją reaguotų). Todėl žemiau pateikiamas aprašymas skaitytinas tik, *kaip labai neprimygtina rekomendacija*. Visa ankstenė informacija rodo, kad *logistinės regresijos modelis duomenims gerai tinka*.

Vienas išskirčių nustatymo būdų – sukurti kintamuosius, kuriuose atsispindėtų liekamosios paklaidos ir pažiūrėti, ar nėra labai išsiskiriančių reikšmių. Ištirsime standartizuotasias paklaidas ir deviacijos statistikos paklaidas. Deviacija – rodiklis, taikomas apibendrintuose tiesiniuose modeliuose (logistinė regresija – atskiras tokių modelių atvejis). Įvykdome komandas:

```
predict dev, deviance
predict rstandart, rstandard
```

Nusibraižome liekamųjų paklaidų grafikus:

```
scatter rstandart idno
scatter  dev idno
```



Abudu grafikai nerodo kažkokių labai išsiskiriančių reikšmių. Vis dėlto, jeigu laikysime, kad standartinės paklaidos viršijančios 3 rodo išskirtį (ši labai pritempta prielaida siejasi su regresorių normalumu, kuris logistinei regresijai nėra svarbus), tai keletą tokių taškų rasime. Kaip nustatyti, kur jie? Galima tiesiog įvykdyti komandą:

```
list rstandart if rstandart> 3
```

```
     | rstand~t |
     |----------|
 30. | 3.050217 |
191. | 3.942029 |
275. | 3.050217 |
276. | 3.050217 |
     |----------|
```

Galima šias išskirtis pašalinti ir pažiūrėti, ar labai keičiasi modelis. Alternatyva aptarta kitame skyrelyje.

4.3.3. Atsparioji logistinė regresija

Jeigu kilo abejonių dėl duomenyse esančių išskirčių, galima suskaičiuoti modelio koeficientus, naudojantis atspariosios (robastinės) logistinės regresijos opciją.

```
logit country freehms stfgov imsb2, robust
```

```
. logit country freehms stfgov imsb2, robust

Iteration 0:   log pseudolikelihood = -194.30391
Iteration 1:   log pseudolikelihood = -139.71847
Iteration 2:   log pseudolikelihood = -135.06235
Iteration 3:   log pseudolikelihood =  -134.8241
Iteration 4:   log pseudolikelihood = -134.82313

Logistic regression                               Number of obs   =        284
                                                  Wald chi2(3)    =      62.44
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -134.82313                 Pseudo R2       =     0.3061

------------------------------------------------------------------------------
             |               Robust
     country |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     freehms |    -.81604     .16484    -4.95   0.000     -1.13912   -.4929596
      stfgov |   .3504268   .0761087     4.60   0.000     .2012565    .4995972
       imsb2 |  -1.708391   .3492739    -4.89   0.000    -2.392955   -1.023827
       _cons |   1.306911   .4261149     3.07   0.002     .4717415    2.142081
------------------------------------------------------------------------------
```

Gautoji modelio lygtis

$$\frac{\widehat{\mathrm{P}}(country = PT)}{\widehat{\mathrm{P}}(country = LT)} = \exp\{1{,}31 + 0{,}35\, stfgov - 0{,}81\, freehms\ - 1{,}71 imsb2\}.$$

Matome, kad ji nė kiek nesiskiria, nuo gautosios, taikant klasikinę logistinę regresiją. Tai dar kartą patvirtina, kad anksčiau nagrinėtas modelis gerai tinka duomenims.

Dar viena situacija, kai siūloma taikyti robastinius liekamųjų paklaidų įverčius, yra vadinamosios *klasterizuotos imtys.* Duomenys natūraliai grupuojasi pagal kažkokį požymį, ir įtariame liekamųjų paklaidų priklausomumą grupėse. Pavyzdžiui, išlaikys egzaminą studentas, ar neišlaikys, sprendžiame tikrindami 7 grupių atstovus. Priklausymas tai pačiai grupei (darbinė atmosfera, tie patys dėstytojai) daro tikėtiną prielaidą, kad rezultatuose tai atsispindės. Tada geriau taikyti atspariosios logistinės regresijos modelį. [14]

4.3.4. Reikšmių prognozavimas

Tarkime, kad norime prognozuoti respondento šalį, žinodami, kad jam(jai) *freehms* = 3, *stgov* = 5, *imsb2* = 0. Tą padaryti galima, įvykdžius komandą:

```
adjust freehms = 3 stfgov = 5 imsb2 = 0, pr
```

```
     Dependent variable: country     Command: logit
Covariates set to value: freehms = 3, stfgov = 5, imsb2 = 0
_____________________________________________________________________________

_____________
     All |        pr
_____________
         |  .648152
_____________
     Key:  pr  =  Probability
```

Tikimybė, kad respondentas iš Portugalijos P(*country* = 1) = P(*country* = PT) =0,65. Aišku, kad P(*country* = LT) = 0,35. Tikėtiniau, kad iš Portugalijos.

Galima padaryti prognozes abiem *imsb2* atvejams:

```
adjust freehms = 3 stfgov = 5, by(imsb2) pr
```

```
benefits/
servi    |        pr
_________|__________
       0 |  .648152
       1 |  .250215
_________|__________
     Key:  pr  =  Probability
```

Kai *imsb2* = 0, tikimybė P(*country* = PT) =0,65 sutampa su jau anksčiau apskaičiuotąja. Kai *imsb2* = 1, tikimybė P(*country* = PT) =0,25. Tikimybė sumažėja daugiau, nei du kartus ir tampa

[14] Turint informaciją apie klasterių savybes, reikėtų taikyti HLM, žr. pratarmę.

labiau tikėtina, kad respondentas ij Lietuvos. Vaizdžiai matome požiūrio į imigrantus vaidmenį. Priešiškesnė nuostata imigrantų atžvilgiu labai stipriai keičia tikėtiną respondento šalį.

Galima skaičiuoti ne pačias tikimybes, bet galimybes (tikimybių santykius). Įvykdome:

```
adjust freehms = 3 stfgov = 5 imsb2 = 0, exp
```

```
     Dependent variable: country     Command: logit
Covariates set to value: freehms = 3, stfgov = 5, imsb2 = 0
```

All	exp(xb)
	1.84214

```
     Key:  exp(xb)  =  exp(xb)
```

Kadangi jau žinome, jog P(*country* = PT) =0,65, P(*country* = LT) = 0,35, tai nesunku patikrinti

$$\frac{\widehat{\mathrm{P}}(country = PT)}{\widehat{\mathrm{P}}(country = LT)} = \frac{0{,}65}{0{,}35} = 1{,}8\ldots$$

4.3.5. Kategoriniai regresoriai

Dvireikšmio regresoriaus atvejį jau aptarėme ankstesniuose skyreliuose. Situacija sudėtingesnė, kai regresorius gali įgyti daugiau nei dvi kategorines reikšmes. Tada jį reikia pakeisti keletu dvireikšmių pseudokintamųjų. STATA programa tą galima padaryti automatiškai. Ištirsime modelį, kai vietoje *imsb2* į modelį įtrauktas *imsb3.*Įvykdome komandą:

```
xi: logit  country  freehms stfgov i.imsb3
```

Opcija i.imsb3 reiškia, kad bus sukurti du (vienu mažiau nei yra kategorijų) pseudokintamieji, įgyjantys reikšmes 0 ir 1. Be to, 1 atitiks attinkamą *imsb3* reikšmę.

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(4)      =     119.65
                                                  Prob > chi2     =     0.0000
Log likelihood = -134.47869                       Pseudo R2       =     0.3079
```

country	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
freehms	-.8117534	.1514513	-5.36	0.000	-1.108592	-.5149144
stfgov	.3533061	.0757091	4.67	0.000	.204919	.5016931
_Iimsb3_1	-.3309373	.4009415	-0.83	0.409	-1.116768	.4548937
_Iimsb3_2	-1.953385	.4556013	-4.29	0.000	-2.846347	-1.060423
_cons	1.532331	.5110018	3.00	0.003	.5307859	2.533876

Gavome statistiškai reikšmingai tinkantį modelį. Vis dėlto, vienas iš naudotųjų pseudokintamųjų (atitinkantis *antrąją* regresoriaus *imsb3* reikšmę) yra statistiškai nereikšmingas. Pažvelkime į klasifikacinę lentelę (įvykdome lstat):

```
Logistic model for country

                 -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        87            27  |        114
     -     |        36           134  |        170
-----------+--------------------------+-----------
   Total   |       123           161  |        284

Classified + if predicted Pr(D) >= .5
True D defined as country != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   70.73%
Specificity                     Pr( -|~D)   83.23%
Positive predictive value       Pr( D| +)   76.32%
Negative predictive value       Pr(~D| -)   78.82%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   16.77%
False - rate for true D         Pr( -| D)   29.27%
False + rate for classified +   Pr(~D| +)   23.68%
False - rate for classified -   Pr( D| -)   21.18%
--------------------------------------------------
Correctly classified                        77.82%
--------------------------------------------------
```

Bendrasis teisingų klasifikavimų procentas (77,82%) kiek didesnis, nei modelyje su *imsb2* (76,41%). Kita vertus, mažiau teisingai klasifikuoti Portugalijos atstovai (70,73%, o buvo 72,36 %), ir kiek geriau Lietuvos respondentai (83,23%, buvo 79,5%).). Galime daryti išvadą, kad apskritai šis modelis nėra geresnis už buvusįjį.

4.3.6. Multikolinearumas

Parodysime, kaip gali kilti (ir likti nepastebėta) multikolinearumo problema, į modelį įtraukdami kintamojo *freehms* kvadratą. Paaiškinti tokį įtraukimą nesunku – manome, kad šis kintamsis toks svarbus, kad priklausomybė nuo jo ne tiesinė, o kvadratinė. Naująjį regresorių (pavadinsime jį *freesquare*) sukuriame komanda:

```
gen freesquare = freehms*freehms
```

Sudarome modelį su visais kintamaisiais:

```
logit country freehms stfgov imsb2  freesquare
```

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(4)      =     126.26
                                                  Prob > chi2     =     0.0000
Log likelihood =  -131.1732                       Pseudo R2       =     0.3249

------------------------------------------------------------------------------
     country |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     freehms |  -2.612501   .7065961    -3.70   0.000    -3.997404   -1.227598
      stfgov |   .3687341   .0762968     4.83   0.000     .2191951    .5182732
       imsb2 |  -1.759636   .3542756    -4.97   0.000    -2.454004   -1.065269
  freesquare |   .3123296   .1161209     2.69   0.007     .0847368    .5399223
       _cons |   3.460992   .9527517     3.63   0.000     1.593633    5.328351
------------------------------------------------------------------------------
```

Matome, kad modelyje visi kintamieji statistiškai reikšmingi, chi kvadrato reikšmė didelė (126, p = 0,000...). Vis dėlto, modelyje yra akivaizdžiai neteisinga priklausomybė nuo *freesquare*: šio regresoriaus koeficientas yra teigiamas, priešingai nei *freehms.* Iškart susiduriame su interpretacijos problema. Multikolinearumą galima pastebėti, užmetus akį į kintamųjų koreliaciją:

```
corr freesquare freehms
```

```
. corr  freesquare freehms
(obs=284)

             | freesq~e  freehms
-------------+------------------
  freesquare |   1.0000
     freehms |   0.9791   1.0000
```

Koreliacija lygi 0,97, t.y. nepaprastai stipri. Įsitikinti, kad regresijos lygtyje atsirado neteisingas koeficiento ženklas, galima ir tiesiog atliekant logistinę *country* regresiją pagal vienintelį kintamąjį *freesquare*. Įvykdome

```
logit country  freesquare
```

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(1)      =      52.98
                                                  Prob > chi2     =     0.0000
Log likelihood = -167.81396                       Pseudo R2       =     0.1363

------------------------------------------------------------------------------
     country |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  freesquare |  -.1487757   .0249306    -5.97   0.000    -.1976387   -.0999126
       _cons |   .8404502   .2056653     4.09   0.000     .4373536    1.243547
------------------------------------------------------------------------------
```

Matome, kad koeficiento prie *freesquare* ženklas, kaip ir tikėjomės yra neigiamas. Beje, multikolinearumo problemą galima išspręsti iš modelio pašalinant ir *freehms*. Čia, kaip ir dažnai būna regresijoje – nėra vienintelio teisingo duomenims tinkamo modelio. Gali būti net keli

statistikos požiūriu tinkami modeliai. Geresnis iš jų tas, kuris geriau atspindi realią situaciją. O tą nuspręsti padeda ne statistika, o geras tiriamo dalyko išmanymas.

Praktiniuose logistinės regresijos taikymuose multikolinearumas tikrinamas tik tada, kai tyrėjo nuomone atsiranda netikėti koeficientų ženklai. Jeigu tyrėjas savo nuomonės neturi (o kitų nuomonių nežino), tai jis gali pažiūrėti, ar intervalinių regresorių koreliacijos nėra didelės. Jeigu tyrėjui neaišku, kokia koreliacija yra didelė (jis ir sriubos pasūdyti nesugeba, jeigu neranda instrukcijos, kiek gramų druskos į litrą sriubos dėti, ką jau čia apie koreliaciją bekalbėti), tai jis gali sudaryti logistinės regresijos modelį su *vieninteliu* regresoriumi. Regresoriaus koeficiento ženklas (teigiamas/neigiamas) turėtų išlikti ir jungtiniame modelyje. Operaciją kartojame su antruoju regresoriumi. Tada su trečiuoju. Ir t.t.

4.3.7. Žingsninė logistinė regresija

Viskas, kas pasakyta apie žingsninės regresijos trūkumus, kalbant apie tiesinę regresiją, tinka ir logistinės regresijos atveju. Trumpai galima reziumuoti taip: jokios formalios statistinės procedūros neatstos gilaus dalyko išmanymo. Pasitikint vien statistiniais skaičiavimais, galima prikurti labai neteisingų interpretacijų.

Po tokios pesimistinės įžangos, pažiūrėkime, kaip su STATA atlikti žingsninę logistinę regresiją. Kaip ir visada, galima žingsniuoti į priekį (angl. stepwise forward regresssion) arba atgal (angl. stepwise backward regression). Tirsime modelį su *imsb3*. Kaupiamoji (akumuliuojanti, forward) žingsninė regresija atliekama komanda

```
xi: sw logit country freehms stfgov i.imsb3, pe(0.05)
```

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(3)      =     118.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -134.82313                       Pseudo R2       =     0.3061

------------------------------------------------------------------------------
     country |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     freehms |    -.81604    .151165    -5.40   0.000    -1.112318   -.5197622
   _Iimsb3_2 |  -1.708391   .3430608    -4.98   0.000    -2.380778   -1.036004
      stfgov |   .3504268   .0753233     4.65   0.000     .2027959    .4980578
       _cons |   1.306911    .427897     3.05   0.002     .4682487    2.145574
------------------------------------------------------------------------------
```

Matome, kad modelyje liko kintamasis, kuris iš esmės sutampa su *imsb2*, t.y. netiesiogiai siūloma pereiti prie modelio su *imsb2*. O žiūrint formaliai, nieko gero negavome – nes neišeina į modelį įtraukti tik pusę kategorinio kintamojo (*_Iimsb3_1* modelyje nebeliko).

Šalinamoji (eliminuojanti, backward) žingsninė regresija atliekama komanda

```
xi: sw logit country freehms stfgov i.imsb3, pr(0.05)
```

```
Logistic regression                               Number of obs   =        284
                                                  LR chi2(3)      =     118.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -134.82313                       Pseudo R2       =     0.3061

------------------------------------------------------------------------------
     country |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     freehms |    -.81604    .151165    -5.40   0.000    -1.112318   -.5197622
      stfgov |   .3504268   .0753233     4.65   0.000     .2027959    .4980578
   _Iimsb3_2 |  -1.708391   .3430608    -4.98   0.000    -2.380778   -1.036004
       _cons |   1.306911    .427897     3.05   0.002     .4682487    2.145574
------------------------------------------------------------------------------
```

Gavome tiksliai tą patį modelį, kaip ir pirmuoju atveju (taip įvyksta ne visada).

Uždaviniai

1. Failas **ESS4LT1.sav**. Socialinis Lietuvos respondentų tyrimas 2009 m. spalis – 2010 m. ir analogiškas 2008 m. tyrimas Estijoje **ESS4EE.sav**. Kintamieji:

- *ageyo30* (amžiaus grupės, *ageyo30* = 1, jeigu iki 29 m.)
- *gndr* (lytis, 1 – vyr., 2 – mot.),
- *cntry* (šalis, LT – Lietuva, EE – Estija),
- *pray* (maldos ne religinių apeigų metu, 1 – kasdien, 2 – dažniau nei kartą per sav., 3 – kartą per sav., 4 – bent kartą per mėn., 5 – tik per didžiąsias šventes, 6 – dar rečiau, 7 – niekada),
- *freehms* (gėjai ir lesbietės gali gyventi, kaip nori, 1 – pilnai sutinku, ... 5 – visiškai nesutinku),
- *happy* (laimingumas, 0 – ypatingai nelaimingas, 10 – ypatingai laimingas),
- *stfdem* (kiek respondentas yra patenkintas demokratijos padėtimi savo šalyje),
- *stfeco* (kiek respondentas yra patenkintas savo šalies ekonomine padėtimi),
- *stfedu* (kiek respondentas yra patenkintas savo šalies švietimo sistema),
- *stfgov* (kiek respondentas yra patenkintas savo šalies vyriausybe)
- *stfhlth* (kiek respondentas yra patenkintas savo šalies sveikatos apsauga),
- *stflife* (kiek respondentas yra patenkintas savo gyvenimu apskritai).

Kintamieji *stfdem, stfeco, stfgov, stfhlth, stflife* matuojami 10 balų skale nuo 0 (itin nepatenkintas) iki 10 (itin patenkintas).

Atrinkus moteris iki 29 m, ištirti modelį:

cntry = f(pray, freehms,happy,stfdem, stfeco, stfgov, stfhlth, stflife).

Atsakyti į klausimus:

1) Ar didžiausio tikėtinumo chi kvadratas statistiškai reikšmingas?
2) Ar visi pradinio modelio kintamieji statistiškai reikšmingi?
3) Ar yra multikolinearumo problema?
4) Ar pakankamas klasifikavimo tikslumas?
5) Modelį patobulinti, naudojantis žingsnine regresija.

2. Failas **ZTLT** . Tyrimas: Lietuvos tolerancijos profiliai, 2003 m. Klausimų grupės:

- *r02_01 – r02_04* ranginėje skalėje įvertinta, koks procentas nurodytos grupės (kitos rasės, kitos tautybės, kitos religijos, kitos seksualinės orientacijos) gyventojų nekeltų Lietuvoje problemų.
- *r04_01 – r04_09* Lietuvoje turėtų būti toleruojama (skalėje 1 – visiškai sutinku, 2 – sutinku, 3 – nesutinku, 4 – visiškai nesutinku) kitos tautybės, kitos rasės, kitos krikščioniškos konfesijos, nekrikščioniškos religijos, suteikiamas prieglobstis politiniams pabėgėliams, suteikiamas prieglobstis ekonominiams pabėgėliams, visi Lietuvos gyv. turi turėti lygias galimybes įsidarbinti ir socialines garantijas).
- *s01* – lytis,
- *s02* – amžius,
- *r01_01* respondento požiūris (skalėje 1 – teigiamas, 2 – neutralus, 3 – neigiamas, 4 – sunku pasakyti) į juodaodžius.
- *r01_02* respondento požiūris (skalėje 1 – teigiamas, 2 – neutralus, 3 – neigiamas, 4 – sunku pasakyti) į musulmonus.

Sukurti naujus kintamuosius:

1) *a2 = r02_01 + ...+ r01_04*. Kintamasis *a2* atspindi respondento pakantumą „kitokiems“. Kuo didesnė *a2* reikšmė, tuo pakantumas didesnis.
2) *a4 = r04_01 + ...+ r04_09*. Kintamasis *a4* – tai respondento nuomonė apie valstybės tolerantiškumo laipsnį. Kuo mažesnė *a4* reikšmė, tuo valstybė tolerantiškesnė.

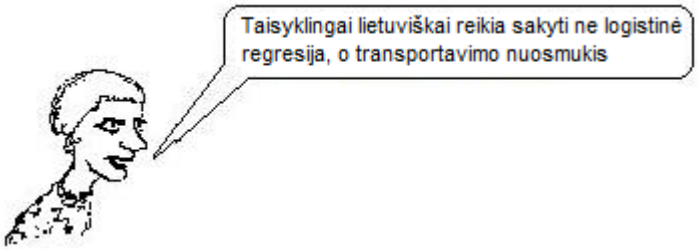
Atsirinkti ne vyresnes nei 35 m. moteris, kurios turi savo nuomonę apie juodaodžius ir musulmonus (*r01_01* ir *r01_02* nelygūs 4). Įsitikinti, kad tarp šių moterų yra tik viena, kuri nepalankiai vertina juodaodžius. Pašalinti ją iš tolesnio tyrimo. Sukurti naują kintamąjį *juodas* (*juodas* = 0, kai *r01_01* = 1 ir *juodas* = 1, kai *r01_01* = 2).

Ištirti logistinės regresijos modelį *juodas = f(a2, a2, r01_02).*

3. Failas **ZTLT1**. Vyrai ir moterys, 2006 m. spalis. Kintamieji

- *s1* – lytis (1 – vyr., 2 – mot.),
- *s12* – regionas (*s12* = 1, jeigu Vilniaus; *s12* = 2, jeigu Kauno, ...),
- *G6_1* – dažniausiai ruošia maistą,
- *G6_2* – dažniausiai plauna indus,
- *G6_3* – dažniausiai perka maisto produktus,
- *G6_4* – dažniausiai plauna grindis, šluosto dulkes, valo kilimus,
- *G6_5* – dažniausiai taiso sugedusius buitinius prietaisus,
- *G6_6* – dažniausiai apmoka sąskaitas,
- *G6_7* – dažniausiai organizuoja laisvalaikį.

Visi kintamieji *G6_1 – G6_7* matuojami skalėje: 1 – visada aš, 2 – dažniausiai aš, 3 – vienodai abudu sutuoktiniai/partneriai, 4 – dažniausiai sutuoktinis/partneris, 5 – visada sutuoktinis/partneris, 6 – visada arba dažniausiai kitas kartu gyvenantis žmogus. Taigi, kuo kintamieji *G6_1 – G6_7* mažesni, tuo labiau problemą sprendžia pats(pati) respondentas(ė). Atsirinkus Kauno regiono gyventojus, su žingsnine logistine regresija ištirti modelį *s_1 = f(G6_1, G6_2, G6_3, G6_4, G6_5, G6_6, G6_7).* Ar gautasis modelis tinkamas? Kokius alternatyvius modelius galima pasiūlyti?

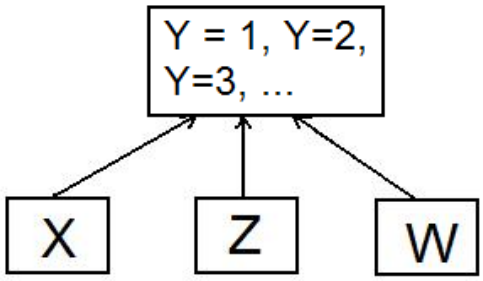


5. DAUGIANARĖ LOGISTINĖ REGRESINĖ ANALIZĖ

Tipinės situacijos, kai taikoma daugianarė logistinė regresinė analizė yra šios:

- Norime įvertinti, kaip perkamo automobilio spalva priklauso nuo pirkėjo amžiaus, lyties ir išsilavinimo.
- Tiriame, kas charakterizuoja parlamentaro balsavimą (pritaria-nepritaria-susilaiko).
- Norime išsiaiškinti, kaip vienuolikos metrų baudinio realizavimą svarbiose futbolo rungtynėse įtakoja nervingumas, patirtis ir nuovargis.
- Tiriame, kaip vairavimo egzamino praktinės dalies (išlaiko iš karto – išlaiko antru laikymu – pirmus du kartus laiko nesėkmingai) priklauso nuo vairuotojo amžiaus ir turėtų praktinių užsiėmimų skaičiaus. Viena iš tiriamų problemų yra tokia: kaip kiekviena papildoma su vairavimo instruktoriumi praleista valanda pakeičia tikimybių iškart išlaikyti egzaminą / neišlaikyti egzamino santykį.

Modelio schema:



Kintamasis *Y* vadinamas *priklausomu* (regresuojamu) *kintamuoju*, kintamieji *X, Z, W* vadinami *nepriklausomais kintamaisiais* (*regresoriais*). Akivaizdu, kad daugianarė logistinė regresija yra dvinarės logistinės regresijos apibendrinimas, kai priklausomas kintamasis įgyja daugiau nei dvi skirtingas reikšmes.

5.1. Daugianarės logistinės regresijos modelis

5.1.1. Modelio lygtys

Supaprastintai daugianarę logistinę regresiją galima įsivaizduoti taip: sumodeliuojame kiekvienos *Y* reikšmės įgijimo tikimybes. Jas ir nagrinėjame. Realiai lyginamos ne visos tikimybės tarpusavyje, o visos su viena – iš anksto pasirinktos *Y* reikšmės tikimybe. Todėl daugianarės logistinės regresinės analizės matematinis modelis aprašomas keliomis lygtimis. Pasirenkama viena vadinamoji kontrolinė *Y* kategorija (dažniausiai tai didžiausia *Y* reikšmė), ir sudaroma daug dvinarės logistinės regresijos dalinių modelių kitų *Y* kategorijų tikimybių ir kontrolinės kategorijos

tikimybės santykiams. Parodysime, kaip sudaromi modeliai, kai Y įgyja keturias reikšmes $Y = 1, 2, 3, 4$. Tegul kontrolinė kategorija atitinka reikšmę $Y = 4$. Sudarome tris matematinius modelius tikimybių santykiams.

$$\frac{\mathrm{P}(Y=1)}{\mathrm{P}(Y=4)} = \exp\{C_1 + b_{11}X + b_{12}Z + b_{13}W\},$$
$$\frac{\mathrm{P}(Y=2)}{\mathrm{P}(Y=4)} = \exp\{C_2 + b_{21}X + b_{22}Z + b_{23}W\},$$
$$\frac{\mathrm{P}(Y=3)}{\mathrm{P}(Y=4)} = \exp\{C_3 + b_{31}X + b_{32}Z + b_{33}W\}.$$

Koeficientai $C_1, b_{11}, b_{12}, b_{13}, C_2, b_{21}, b_{22}, b_{23,}C_3, b_{31}, b_{32}, b_{33}$ nėra žinomi. Jų įverčiai $\hat{C}_1, \hat{b}_{11}, \hat{b}_{12}, \hat{b}_{13}$ .... gaunami, panaudojus imties duomenis. Teigiamas (neigiamas) koeficientas rodo, kad kintamajam didėjant, priklausymo konkrečiai Y kategorijai tikimybė*, lyginant ją su kontrolinės kategorijos tikimybe*, išauga (sumažėja). Žinoma, jeigu kitų kintamųjų reikšmės lieka fiksuotos.

Jeigu $b_{11} > 0$, tai X didėjant, tampa labiau tikėtina, kad $Y = 1$, o ne $Y = 4$.

Jeigu $b_{11} < 0$, tai X didėjant, tampa labiau tikėtina, kad $Y = 4$, o ne $Y = 1$.

Pavyzdžiui, jeigu $Y = 4$ žymi atvejį, kai pacientas sveikas, o $Y = 1$ – susirgimą tuberkulioze, tai $\hat{b}_{11} > 0$, rodo, kad X didėjant, tikimybė susirgti tuberkulioze, lyginant ją su tikimybe išlikti sveikam, išauga. Su $Y = 4$ lyginame, tik todėl, kad minėtame modelyje tai buvo kontrolinė kategorija. Jeigu kontrolinė kategorija būtų $Y = 3$, tai viską lygintume su $Y = 3$.

Aprašant regresijos modelį, kaip taisyklė, apsiribojama lyginimais su kontroline kategorija. Vis dėlto pažymėtina, kad galima pagal koeficientų reikšmes padaryti išvadą ir apie dviejų nekontrolinių kategorijų tikimybes. Tereikia atsižvelgti į atitinkamų koeficientų skirtumą. Fiksuokime visų kintamųjų, išskyrus X, reikšmes.

Jeigu $b_{11} - b_{21} > 0$, tai X didėjant, tampa labiau tikėtina, kad $Y = 1$, o ne $Y = 2$.

Jeigu $b_{11} - b_{21} < 0$, tai X didėjant, tampa labiau tikėtina, kad $Y = 2$, o ne $Y = 1$.

Analogiškai visiems regresoriams galima palyginti kitas Y kategorijas.

5.1.2. Prognozavimas

Tikimybių P(Y = k) prognozės konkrečioms X, Z, W reikšmėms skaičiuojamos taip: pasižymime tikimybių santykių logaritmus (kurie dar vadinami tikimybių santykių logit funkcijomis) z_1, z_2, z_3 ir pagal formules

$$z_1 = \hat{C}_1 + \hat{b}_{11}X + \hat{b}_{12}Z + \hat{b}_{13}W, \quad z_2 = \hat{C}_2 + \hat{b}_{21}X + \hat{b}_{22}Z + \hat{b}_{23}W,$$
$$z_3 = \hat{C}_3 + \hat{b}_{31}X + \hat{b}_{32}Z + \hat{b}_{33}W$$

surandame konkrečias jų reikšmes. Tam užtenka įstatyti tas *X*, *Z* ir *W* reikšmes, kurioms norime gauti prognozę. Tikimybės įverčiai skaičiuojami pagal formules:

$$\widehat{\mathrm{P}}(Y=1) = \frac{\mathrm{e}^{z_1}}{1+\mathrm{e}^{z_1}+\mathrm{e}^{z_2}+\mathrm{e}^{z_3}}, \quad \widehat{\mathrm{P}}(Y=2) = \frac{\mathrm{e}^{z_2}}{1+\mathrm{e}^{z_1}+\mathrm{e}^{z_2}+\mathrm{e}^{z_3}},$$
$$\widehat{\mathrm{P}}(Y=3) = \frac{\mathrm{e}^{z_3}}{1+\mathrm{e}^{z_1}+\mathrm{e}^{z_2}+\mathrm{e}^{z_3}}, \quad \widehat{\mathrm{P}}(Y=4) = \frac{1}{1+\mathrm{e}^{z_1}+\mathrm{e}^{z_2}+\mathrm{e}^{z_3}}.$$

Čia e = 2,7183...

Pavyzdys. Tarkime, kad atlikę daugialypę logistinę regresiją, gavome tokius logit funkcijų įverčius

$$z_1 = 1 + 2X - Z + 3W, \quad z_2 = 2 + X + Z + 2W, \quad z_3 = 1 + X + Z + W.$$

Tarkime, kad *X* = 0, *Z* =1, *W* =0. Įstatome šias reikšmes ir gauname $z_1 = 0$, $z_2 = 3$, $z_3 = 2$. Nesunkiai surandame $\mathrm{e}^0 = 1$, $\mathrm{e}^3 = 20{,}1$, $\mathrm{e}^2 = 7{,}39$, $\mathrm{e}^0 + \mathrm{e}^3 + \mathrm{e}^2 = 29{,}49$. Todėl

$$\widehat{\mathrm{P}}(Y=1) = \widehat{\mathrm{P}}(Y=4) = \frac{1}{29{,}49} = 0{,}0339, \ \widehat{\mathrm{P}}(Y=2) = \frac{20{,}1}{29{,}49} = 0{,}68, \ \widehat{\mathrm{P}}(Y=3) = \frac{7{,}39}{29{,}49} = 0{,}25.$$

Taigi, šiam repondentui labiausiai tikėtina, kad *Y* =2.

Kartais reikalinga *Y* reikšmės prognozė. Pavyzdžiui, kokios spalvos automobilį pirks dvidešimtmetė pankė? Prognozė daroma taip: suskaičiuojamos visų galimų *Y* kategorijų tikimybės. Prognozuojame tą *Y* reikšmę, kurios įgijimo tikimybė didžiausia. Jeigu didžiausia tikimybė, kad pankė pirks raudoną automobilį, tai ir prognozuojame, kad ji pirks raudoną automobilį.

5.1.3. Galimybių santykiai

Konkretiems stebėjimams apskaičiavus tikimybes, galima surasti jų santykius P(*Y*=j) / P(*Y*=m), kurie vadinami *galimybėmis* (angl. odds). Galimybė parodo, kiek kartų kintamajam *Y* yra tikėtiniau įgyti reikšmę *j*, nei įgyti reikšmę *m*. Aišku, kad galimybių yra viena mažiau, nei *Y* kategorijų.

Galimybių santykis (angl. odds ratio) parodo, kiek kartų pasikeis tikimybių santykis (galimybė), kai atitinkamas kintamasis padidės vienetu, fiksavus kitų kintamųjų reikšmes. Praktiškai galimybių santykiai skaičiuojami tik lyginant su kontroline kategorija. Statistinės

programos dažniausiai kiekvienam kintamajam pateikia ne tik patį galimybių santykį, bet ir šio santykio 95% pasikliautinąjį intervalą.

Tarkime, kad kontrolinė kategorija yra Y =3 ir tiriame P(Y=1)/P(Y=3). Kai regresorius X padidėja vienetu, ankstesnis tikimybių santykis pasikeičia per $e^{\hat{b}_{11}}$.

Skaičius $e^{\hat{b}_{11}}$ yra galimybių santykis (odds ratio) galimybei P(Y=1)/P(Y=3).

$$\left(\frac{P(Y=1)}{P(Y=3)}\right)_{naujas} = e^{\hat{b}_{11}} \left(\frac{P(Y=1)}{P(Y=3)}\right)_{senas}$$

Pavyzdys. Tarkime, kad tiriame respondentus, kurių abudu tėvai sirgo diabetu, Y žymi respondento sveikatos būseną: Y = 1 (pirmo tipo diabetas), Y = 2 (antro tipo diabetas), Y = 3 (sveikas). Tegul Y = 3 yra kontrolinė kategorija. Tarkime, kad vienas iš regresorių yra kūno masės indeksas. Jeigu kūno masės indekso galimybių santykis modeliui P(Y=1)/P(Y=3) yra 1,2, tai reikš, kad kūno masės indekso padidėjimas padidina galimybę susirgti pirmos rūšies diabetu 1,2 karto. Analogiškai interpretuojamas galimybių santykis modeliui P(Y=2)/P(Y=3).

Nors galimybių santykiai ir atspindi regresorių svarbą, bendriausiu atveju jie neleidžia nustatyti, kurie regresoriai modelyje svarbiausi. Daugianarėje logistinėje regresijoje iš viso tokių rodiklių (tiesinės regresijos standartizuotųjų beta koeficientų analogų) nėra. Jeigu regresoriai palyginami pagal savo reikšmes (nelyginame tonų su metais), tai dar gal ir galima kalbėti apie regresorių svarbą *daliniuose* modeliuose. Tada tiesiog reikia prisiminti dvinarės logistinės regresijos galimybių santykių interpretacijas, žr. 4.1.2 skyrelį.

5.1.4. Duomenys

Duomenys turi tenkinti tokius reikalavimus:

- Priklausomas kintamasis Y kategorinis, regresoriai – intervaliniai arba kategoriniai. Visi kategoriniai regresoriai keičiami dvireikšmių (0 ir 1) regresorių rinkiniu. SPSS ir STATA programos tą daro automatiškai. Primename, kad ranginis kintamasis su mažai įgyjamų reikšmių (pvz., mažiau nei 5 reikšmėmis) irgi traktuotinas, kaip kategorinis. Jis irgi keičiamas dvireikšmių regresorių rinkiniu.
- Intervaliniai regresoriai neturi stipriai koreliuoti.
- Duomenyse negali kažkurių Y reikšmių būti labai mažai. Turime turėti pakankamai informacijos apie įvairių kategorijų respondentus. Tai kiek gi stebėjimų turi tekti kiekvienai

Y kategorijai? Griežtos instrukcijos neegzistuoja. Idealiu atveju visų kategorijų stebėjimų yra po lygiai. Žiūrint į turimą imtį, tenka patiems apsispręsti, ar nukrypimai nuo idealios imties struktūros yra maži. Jeigu *Y* įgyja 3– 4 reikšmes, tai imtyje kiekvienos kategorijos stebėjimai turėtų sudaryti bent 15% – 20% visų stebėjimų. Jeigu *Y* įgyja 5 – 6 reikšmes, tai imtyje kiekvienos kategorijos stebėjimai turėtų sudaryti bent 10% – 15% visų stebėjimų. Šių konspektų autoriui iš viso neteko matyti, kad praktikoje daugianarė logistinė regresija būtų taikoma, kai *Y* įgyja daugiau nei 6 skirtingas reikšmes.

- Jeigu modelyje yra kategorinių regresorių, tai reikalaujama, kad kiekvienam jų reikšmių deriniui būtų bent keli *Y* stebėjimai. Pavyzdžiui, jeigu regresoriai *X* ir *Z* įgyja tik dvi reikšmes (0 ir 1), tai imtyje turi būti *Y* stebėjimų, kai ($X = 0$, $Z = 0$), kai ($X = 0$, $Z = 1$), kai ($X = 1$, $Z = 0$), kai ($X = 1$, $Z = 1$). Praktikoje šis reikalavimas keičiamas švelnesniu – turi būti ne daugiau 5 % kategorinių regresorių reikšmių derinių, kuriems neturime *Y* stebėjimų. Intervaliniams regresoriams tokio reikalavimo nekeliame.

5.1.5. Modelio tinkamumas

Modelio tinkamumas nustatomas, atsižvelgus į keletą rodiklių. Svarbiausieji yra šie:

a) *Klasifikacinė lentelė.* Tai viena svarbiausių modelio tinkamumo duomenims charakteristikų. Jeigu jau sudarėme logistinės regresijos modelį duomenims, kuriems žinomos *Y* reikšmės, tai natūralu pasitikrinti, ar gerai veikia klasifikavimas. Konkretiems stebėjimams prognozuojama *Y* reikšmė ir žiūrima, ar spėjimas sutapo su tikrąja *Y* reikšme. Kuo daugiau sutapimų, tuo modelis geresnis. Įprasta reikalauti, kad teisingai klasifikuojamų kiekvienos kategorijos stebėjimų procentas būtų ne mažesnis, nei tos kategorijos procentas tarp *Y* reikšmių. Kodėl? Logika čia tokia: tarkime visą laiką spėjame, kad $Y = 1$. Teisingų sprendimų bus toks procentas, kokį tarp visų *Y* reikšmių sudaro vienetai. Modelis turėtų užtikrinti ne blogesnį klasifikavimą.

b) *Modelio tikimo didžiausio tikėtinumo santykio chi kvadrato statistika*. (angl. model fit likelihood ratio Chi-square). Parodo, ar modelyje yra bent vienas reikalingas regresorius. Jeigu *p* reikšmė didesnė už 0,05, tai regresijos modelio tinkamumas labai abejotinas.

c) *Pirsono ir deviacijos chi kvadratai.* Tai alternatyvos didžiausio tikėtinumo kriterijui. Modelis tinka duomenims, kai Pirsono ir deviacijos chi kvadratų $p \geq 0{,}05$. Kai modelyje yra intervalinių regresorių, Pirsono ir deviacijos chi kvadrato statistikos taikomos retai.

d) *Regresorių didžiausio tikėtinumo santykio chi kvadratas*. Padeda nuspręsti, ar kintamasis šalintinas iš modelio. Jeigu *p* reikšmė $< 0{,}05$, tai sakome, kad *kintamasis yra statistiškai reikšmingas* ir dažniausiai jį modelyje paliekame. Jeigu *p* reikšmė $\geq 0{,}05$, tai kintamasis yra

statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais (kai be to kintamojo labai suprastėja teisingų klasifikavimų procentas).

e) *Voldo testai*. Aukščiau buvo parodyta, kad daugianarės logistinės regresijos matematinį modelį sudaro daug dalinių modelių skirtų visoms *Y* kategorijoms, išskyrus kontrolinę kategoriją. Voldo kriterijus taikomas kiekvienam iš šių dalinių modelių (submodelių). Jis padeda nustatyti, kurie kintamieji statistiškai reikšmingi, tuose daliniuose modeliuose (jiems $p < 0,05$). Pavyzdžiui, gali būti, kad modelyje, skirtam santykiui P(*Y*=1)/P(*Y*=4), kintamasis *X* bus statistiškai reikšmingas, o modelyje skirtam P(*Y*=3)/P(*Y*=4), kintamasis *X* bus statistiškai nereikšmingas. Tai leis padaryti išvadą, kad kintamasis *X* labiau skiria *Y*=1 nuo *Y*= 4, nei *Y*=3 nuo *Y*= 4. Manoma, kad Voldo kriterijus yra silpnesnis už didžiausio tikėtinumo chi kvadrato kriterijų. Todėl, darant išvadas apie bendrąjį kintamųjų statistinį reikšmingumą, reikėtų remtis ankstesniu punktu.

f) *Determinacijos (pseudo) koeficientai* (angl. pseudo R square). Parodo bendrąjį modelio tikimą. Įgyja reikšmes iš intervalo [0, 1]. Kuo reikšmė didesnė, tuo modelis geriau tinka duomenims. Nelabai gerai, kai $R^2 < 0,20$. Daugianarėje logistinėje regresijoje determinacijos pseudokoeficientai vaidina tik pagalbinį vaidmenį. Koeficientų yra net keli: Kokso-Snelo, Nagelkerkės, Makfadeno. Kurį iš jų naudoti, pasirenka pats tyrėjas (STATA naudoja Makfadeno). Aprašant rezultatus privalu nurodyti tikslų determinacijos koeficiento pavadinimą.

5.1.6. Daugianarės logistinės regresinės analizės etapai

Išsami modelio analizė susideda iš tokių etapų:

1) Patikriname ar modelio didžiausio tikėtinumo chi kvadrato statistikos *p* reikšmė < 0,05. Jeigu $p \geq 0,05$, – modelis netinkamas.
2) Patikriname, ar visi regresoriai statistiškai reikšmingi (visos didžiausio tikėtinumo chi kvadrato kriterijaus *p* reikšmės < 0,05). Jeigu ne – modelis taisytinas.
3) Pažiūrėję į galimybių santykius, nustatome, kaip regresorių pokyčiai veikia konkrečios *Y* reikšmės įgijimo tikimybę, lyginant ją su kontrolinės kategorijos įgijimo tikimybe.
4) Peržiūrime ar pakankamai daug teisingų klasifikavimų klasifikacinėje lentelėje. Jeigu nedaug – modelis netinkamas.
5) Papildomai galima ištirti, ar deviacijos chi kvadrato statistikos *p* reikšmė ≥ 0,05 (taip turi būti geruose modeliuose, kuriuose nedaug intervalinių regresorių). Su Voldo kriterijumi

patikriname, daliniuose modeliuose visi regresoriai statistiškai reikšmingi, t.y., ar jų *p* reikšmės < 0,05. Konstantos statistinio reikšmingumo netiriame.

6) Jeigu viskas gerai – modelį aprašome. Jeigu ne – tobuliname.

Labai gerai duomenims tinkančiame modelyje:

- Modelio didžiausio tikėtinumo chi kvadrato p < 0,05.
- Visiems regresoriams didžiausio tikėtinumo chi kvadrato p < 0,05.
- Daliniuose modeliuose daugumai regresorių Voldo kriterijaus p < 0,05.
- Teisingai klasifikuojama kiekvienos kategorijos stebėjimų procentas turi būti ne mažesnis, nei tos kategorijos procentas tarp *Y* reikšmių.
- Jeigu imtis nėra labai didelė, ir visoms įmanomoms regresorių reikšmių kombinacijoms yra pakankamai daug *Y* stebėjimų (nulinius dažnius turi ne daugiau 5% langelių), tai deviacijos ir Pirsono chi kvadrato statistikų p reikšmės ≥ 0,05.
- Visų duomenų Kuko matai ≤ 1.
- Pasirinktasis determinacijos koeficientas ≥ 0,20.

5.1.7. Modelio tobulinimas

Modelio tobulinimas, kaip taisyklė, reiškia kintamųjų šalinimą arba jų transformaciją.

Pašalinus bent vieną kintamąjį, keičiasi visos modelio charakteristikos: determinacijos koeficientai, visos *p* reikšmės, klasifikacinė lentelė.

Reikia nepamiršti, kad ne visada pavyksta sudaryti duomenims tinkamą modelį. Yra duomenų, kuriems logistinės regresija (o dažnai ir jokia kita regresija) tiesiog netinka. Modelis tobulinamas, jeigu yra problemų:

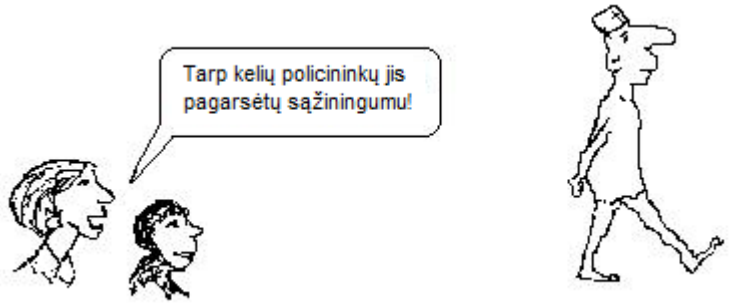
- *Yra statistiškai nereikšmingas kintamasis* (jo didžiausio tikėtinumo chi kvadrato *p* reikšmė didesnė už 0,05). Tada šis kintamasis pašalinamas iš modelio, ir logistinė regresinė analizė pakartojama be jo. Viskas gerai, jeigu klasifikacinėje lentelėje teisingai klasifikuotų atvejų sumažėjo nežymiai. Pristatydami galutinį modelį, tada rašysime, kad modelyje liko tik statistiškai reikšmingi kintamieji. Jeigu teisingai klasifikuotų atvejų sumažėja radikaliai, tai

kintamąjį į modelį gražiname, o aprašyme konstatuojame, kad „kintamasis buvo statistiškai nereikšmingas. Vis dėlto, jis paliktas modelyje, nes be jo labai sumažėja teisingų klasifikavimų procentas“.

- *Yra multikolinearumo problema.* Faktiškai ši problema susideda iš dviejų dalių – priežasties nustatymo ir jos pašalinimo. Multikolinearumą įtarsime tik tada, jeigu modelyje atsiras „keistas“ (mums visai nelauktas) regresoriaus koeficiento ženklas. Tada galima atskirai suskaičiuoti ir peržiūrėti intervalinių regresorių koreliacijas. Jeigu kažkurie regresoriai stipriai koreliuoja, derėtų modelyje palikti vtik kelis iš jų arba juos pakeisti jų vidurkiu. Žinoma, tik tada, kai tas vidurkis yra prasmingas, t.y. nesudėjome metus su kilogramais. Galima sudaryti atskiras daugianares logistines regresijas *Y = f(regresorius)* kiekvienam regresoriui. Tada bent jau išsiaiškinsime, kokie turi būti modelio koeficientų ženklai. O po to vis tiek pradedame šalinti iš modelio kintamuosius, arba sugalvojame kuo juos pakeisti. Daugianarėje logistinėje regresijoje multikolinearumas tikrinamas labai retai.

5.1.8. Pastabos apie multinominės regresijos taikymą

- Kontrolinė *Y* reikšmė dažniausiai yra ta, kuri tyrėjui atrodo svarbiausia.
- Aprašymuose neretai apsiribojama bendruoju teisingų klasifikavimų procentu. Tai nėra blogai. Tik reikia nepamiršti, kad gerai atpažinti turi *visas Y* kategorijas. Pats tyrėjas sprendžia, kurios kategorijos atpažinimas yra svarbiausias.
- Tyrimuose dar taikomos žingsninės logistinės regresijos modeliai, kai modelyje automatiškai paliekami patys statistiškai reikšmingiausi iš pradinių kintamųjų.
- Jeigu kitos charakteristikos geros, bet netikęs teisingai klasifikuotų atvejų procentas, tai modelis vis tiek nelaikomas geru.
- Gali būti ir vienas regresorius. Juo turi būti intervalinis kintamasis.
- Kartais šio skyrelio daugianarė logistinė regresija dar vadinama **nominaliąja daugianare logistine regresija**, skiriant ją nuo ranginės daugianarės logistinės regresijos.
- Pseudokintamieji aprašomi tiap pat, kaip ir tiesinėje regresijoje.
- Galima modelio gerumą įvertinti ir naudojant informacinius indeksus (kriterijus), pavyzdžiui, Akaikės informacinį kriterijų AIC. Gero modelio AIC yra daug mažesnis už AIC modelio be kintamųjų (tik su konstanta). Vis dėlto, toks argumentas gana silpnas, nes skamba panašiai, kaip pasakymas teisme – lyginant su žmogžudžiais-maniakais, kaltinamasis dar visai doras.



5.1.9. Standartiniai daugianarės logistinės regresijos modelių aprašymai

Visada aprašomas galutinis, po visų tobulinimų gautasis, modelis. Pateiksime keletą galimų aprašymų fragmentų.

- *Taikytas daugianarės logistinės regresijos modelis duomenims tiko. Modelio didžiausio tikėtinumo chi kvadrato statistika $\chi^2 = 23,27$; $p < 0,05$. Visi kintamieji buvo statistiškai reikšmingi. Teisingai klasifikuota 80,3% imties respondentų. Nagelgerkės determinacijos koeficientas lygus 0.69.*
- *Iš modelio buvo pašalintas statistiškai nereikšmingas kintamasis X. Galutinis modelis buvo statistiškai reikšmingas, $\chi^2 = 45,2$; $p < 0, 01$. Taikant modelį imties duomenims, teisingai klasifikuoti 73% pirmos rūšies diabetu sergantys pacientai, 65% antros rūšies diabetu sergantys respondentai ir 80% sveikų respondentų. Makfadeno determinacijos koeficientas $R^2 = 0.70$. Kintamajam W padidėjus vienetu galimybių santykis susirgti pirmos rūšies diabetu padidėja 1,5 karto.*
- *Modelyje paliktas statistiškai nereikšmingas kintamasis Z, nes jį pašalinus labai sumažėja teisingai klasifikuotų atvejų procentas.....*

5.2. Daugianarė logistinė regresinė analizė su SPSS

5.2.1. Duomenys

Kaip atlikti daugianarę logistinę regresinę analizę parodysime, tirdami 2008 metų Europos Sąjungos socialinio tyrimo (European Social Survey) Čekijos, Izraelio ir Švedijos duomenis **ESS4CZ, ESS4IL, ESS4SE**. Tyrime naudosime tokius kintamuosius:,

- *cntry* – šalies kodas (CZ – Čekija, IL – Izraelis, SE – Švedija),
- *stfedu* – respondento požiūris į savo šalies švietimo sistemos būklę (reikšmės nuo 0 – itin bloga, iki 10 – itin gera),

- *imsclbn* – nuo kada imigrantai gali pilnai naudotis šalies socialinėmis lengvatomis (matuojamas nominalia skale: 1 – iškart po atvykimo, 2 – po vienerių metų, nepriklausomai nuo to, ar dirbo, 3 – po vienerių darbo ir mokesčių mokėjimo metų, 4 – tapus piliečiu, 5 – niekada negalės naudotis visomis lengvatomis),
- *trstprl* – pasitikėjimas savo šalies parlamentu (matuojamas 10 balų skale nuo 0 – itin nepasitikiu, iki 9 – itin pasitikiu),
- *pray* – kaip dažnai meldžiamasi (matuojamas 7 balų skale nuo 1 – kasdien iki 7 – niekada),
- *hhmmb* – pastoviai kartu gyvenančių namų ūkio narių skaičius.

Bandysime nustatyti, ar išvardinti kintamieji padeda charakterizuoti respondento šalį. Tirsime 20 – 30 metų vyrus. Duomenų atrankai naudosime kintamuosius *agea* (amžius) ir *gndr* (lytis, 1 – vyr., 2 - mot.).

Visus duomenų failus sujungsime į vieną. Atsidarome ESS4CZ ir renkamės **Data → Merge Files→ Add Cases** opciją. Atsidariusiame lange renkamės **Browse**, nurodome failą ESS4IL ir spaudžiame **Continue**. Atsidariusiame lange spaudžiame **OK**. Viską pakartojame prijungdami failą ESS4SE.

Peržiūrėję kintamojo *imsclbn* įgyjamų reikšmių dažnius, įsitikiname, kad Čekijoje labai nedidelis procentas respondentų renkasi reikšmes 1, 2 ir 5. Todėl kintamąjį *imsclbn*, keisime dvireikšmiu kintamuoju *imigrantbf* (*imifrantbf = 0*, kai *imsclbn ≤ 3* ir *imigrantbf = 1*, kai *imsclbn ≥ 4*). Pasirenkame **Transform → Recode into Different Variables** ir perkeliame *imsclbn* į dešinį laukelį. Laukelyje **Name** užrašome *imigrantbf* ir paspaudžiame **Change.** Pasirenkame **Old and New Value**. Naudodami Old **value – New Value – Add** parinktis nurodome, kaip turi atrodyti naujasis regresorius (1 to 3 → 0, 4 to 5 → 1). Po to spaudžiame **Continue** ir **OK**. Norėdami paaiškinti naujo kintamojo prasmę papildomai (**Variable View–Value Labels**) pridedame kintamojo reikšmių komentarus (0 – positive attitude, 1 – negative attitude).

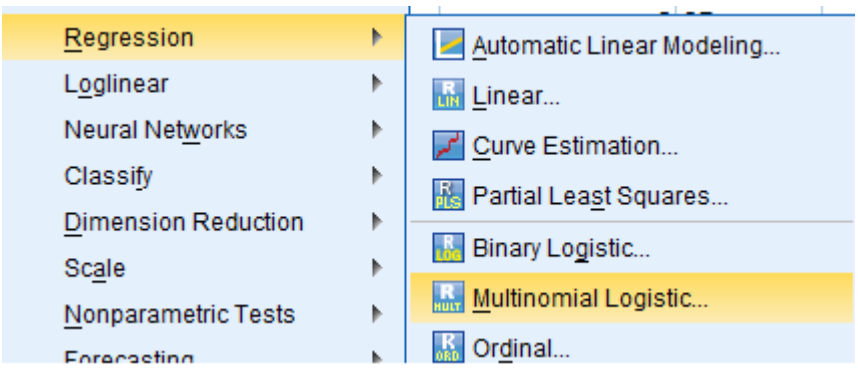
Reikiamus respondentus atrenkame, naudodami **Select Cases** opciją (atrankos sąlyga: agea <= 30 & agea >= 20 & gndr = 1).

5.2.2. SPSS parinktys daugianarei logistinei regresijai

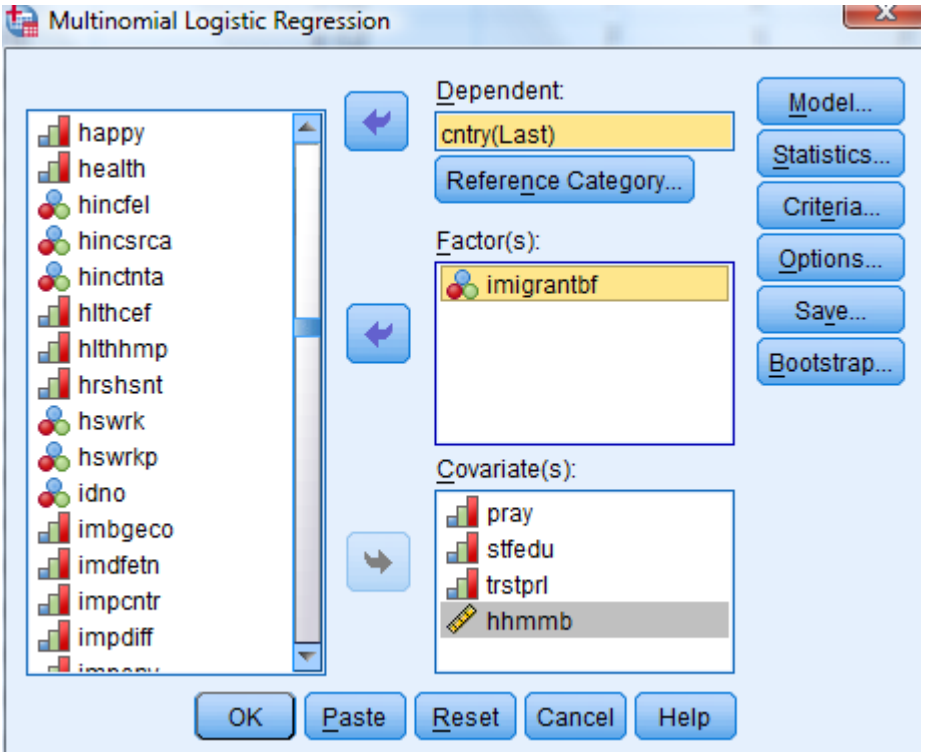
Tirsime modelį:

cntry = f (imigrantbf, pray, stfedu, trstprl, hhmmb).

Atsidarome langelį **Analyze** ir renkamės **Regression → Multinomial Logistic.**

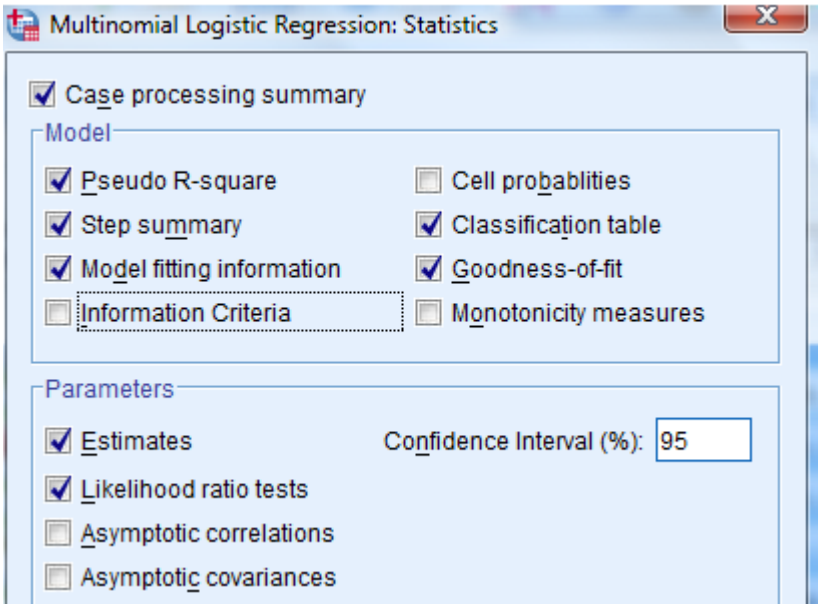


Atsidariusiame meniu perkeliame kintamąjį *cntry* į laukelį **Dependent**. Kintamuosius *pray*, *stfedu, trstprl, hhmmb* įkeliame į laukelį **Covariates**. Kintamasis *imigrantbf* yra kategorinis, todėl jį keliame į **Factor(s)**.



Pagal nutylėjimą SPSS programoje kontroline priklausomo kintamojo kategorija tampa ta, kurios kodas yra didžiausias. Kintamasis *cntry* simbolinis, surikiavus jo reikšmes pagal abėcėlę, kontroline valstybe taps Švedija. **Kontrolinė kategorija – ta kategorija su kuria lyginsime visas likusias**. Dažniausiai – tai svarbiausia priklausomo kintamojo kategorija. Ką pasirinkti kontroline kategorija yra paties tyrėjo reikalas (kaip tą padaryti žr. žemiau). Mes kiekvieną valstybę lyginsime su Švedija ir tai atsispindės išvadose.

Spaudžiame klavišą **Statistics** ir papildomai pažymime **Goodness-of-fit** ir **Classification table**. Visos kitas, pagal nutylėjimą pažymėtas opcijas paliekame ramybėje.



Grįžę į prieš tai buvusį meniu, spaudžiame **OK**.

5.2.3. Rezultatai

Rezultatų išklotinė parsideda įspėjimu, kad 899 (65%) langelių turi nulinius dažnius.

Warnings

There are 899 (65.0%) cells (i.e., dependent variable levels by subpopulations) with zero frequencies.

Nesigilinant į detales, galima pasakyti, kad tai įspėjimas jog Pirsono ir deviacijos chi kvadrato kriterijai **šiam modeliui netaikytini**. Tiesiog buvo sudarytos visos įmanomos regresorių reikšmių kombinacijos ir patikrinta, ar joms yra pakankamai *cntry* reikšmių. Pirsono ir deviacijos kriterijai taikytini tik tada, kai nulinius dažnius turi mažai (pvz., ne daugiau 5%) regresorių rekšmių kombinacijų. Kai modelyje yra bent keli intervaliniai regresoriai, tokia situacija yra reta. Taigi, spręsdami apie modelio tinkamumą duomenims, Pirsono ir deviacijos chi kvadrato kriterijus ignoruosime. Beje, žemiau suradę lentelę **Goodness-of-Fit**, galime įsitikinti abiejų kriterijų neadekvačiu elgesiu: Pirsono chi kvadrato *p* reikšmė lygi nuliui (modelis duomenims absoliučiai netinka), o deviacijos chi kvadrato *p* reikšmė lygi vienetui (modelis duomenims absoliučiai tinka).

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1072.259	910	.000
Deviance	611.343	910	1.000

Lentelėje **Case Processing Summary** pateikiama informacija apie kategorinių kintamųjų reikšmes. Žinome, kad nė viena kategorija neturi labai stipriai dominuoti. Įsitikiname, kad taip ir yra: imtyje yra 158 Čekijos gyventojai, 221 izraelietis ir 142 Švedijos gyventojai. Iš tirtųjų respondentų palankią nuomonę apie socialines lengvatas imigrantams pareiškė apie pusę.

Case Processing Summary

		N	Marginal Percentage
cntry Country	CZ Czech Republic	158	30.3%
	IL Israel	221	42.4%
	SE Sweden	142	27.3%
imigrantbf Social benfits for immigrants	.00 positive attitude	283	54.3%
	1.00 negative attitude	238	45.7%
Valid		521	100.0%
Missing		71	
Total		592	
Subpopulation		461[a]	

a. The dependent variable has only one value observed in 439 (95.2%) subpopulations.

Lentelė **Case Processing Summary** dar naudojama, sprendžiant apie klasifikavimo gerumą. Klasifikacinėje lentelėje teisingų klasifikavimų procentas turi būti ne mažesnis, nei jų procentas imtyje (šie procentai yra lentelėje **Case Processing Summary** stulpelyje **Marginal Percentage**). Logika čia tokia: jeigu užsimerksime ir apie kiekvieną respondentą sakysime „švedas“, tai teisingai klasifikuosime 27,3 % respondentų. Tinkamas diaugianarės logitinės regresijos modelis turi duoti didesnį švedų atpažinimo procentą. Pačiame rezultatų išklotinės gale yra klasifikacinė lentelė:

Classification

	Predicted			
Observed	CZ Czech Republic	IL Israel	SE Sweden	Percent Correct
CZ Czech Republic	111	23	24	70.3%
IL Israel	23	177	21	80.1%
SE Sweden	27	19	96	67.6%
Overall Percentage	30.9%	42.0%	27.1%	73.7%

Iš jos išplaukia, kad, taikydami daugianarės logistinės regresijos modelį, teisingai atpažinome (klasifikavome) 70,3% Čekijos gyventojų, 80,1% izraeliečių ir 67,6% Švedijos gyventojų. Taigi, gavome vieną svarbiausių modelio tinkamumo patvirtinimo įrodymų. Trumpai aptarsime kitus

lentelės **Classification** skaičius. Pirmoje eilutėje parašyta, kad iš visų imties Čekijos atstovų, modelis teisingai atpažino 111, klaidingai 23 Čekijos respondentus palaikė izraeliečiais, o 24 – Švedijos gyventojais. Analogiškai interpretuojamos likusios eilutės. Bendras teisingai klasifikuotų imties respondent procentas 73,7%.

Lentelėje **Pseudo R-Square** yra determinacijos pseudokoficientai. Kaip ir dvinarėje logistinėje regresijoje jie nevaidina bent kiek svarbesnio vaidmens. Vis dėlto, koeficientai turėtų būti nelabai maži. Pavyzdžiui, jie turėtų būti nemažesni už 0,20 (ar bent nedaug mažesni). Kaip taisyklė, modelio aprašyme cituojamas tik vienas determinacijos pseudokoeficientas (kuris – tyrėjo įgeidis). Pavyzdžiui, šiuo atveju galima konstatuoti, kad „gerą modelio tikimą duomenims rodo ir pakankamai didelis Nagelkerkės determinacijos pseudokoeficientas $R^2 = 0,65$“.

Pseudo R-Square

Cox and Snell	.575
Nagelkerke	.650
McFadden	.397

Modelio tikimo didžiausio tikėtinumo santykio chi kvadrato statistikos *p* reikšmė yra lentelės **Model Fitting Information Sig.** stulpelyje. Darome išvadą, kad modelis duomenims tinka (bent vienas regresorius modelyje reikalingas), nes $p = 0,000 < 0,05$.

Model Fitting Information

	Model Fitting Criteria	Likelihood Ratio Tests		
Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1090.376			
Final	644.039	446.337	10	.000

Kurie regresoriai modelyje reikalingi, sprendžiama pagal lentelę **Likelihood Ratio Tests**. Joje yra didžiausio tikėtinumo santykio chi kvadrato kriterijaus rezultatai kiekvienam regrersoriui. Jeigu *p* reikšmė (jos visos yra stulpelyje **Sig**,) yra mažesnė už 0,05, tai sakysime, kad regresorius (kintamasis) yra statistiškai reikšmingas. Toks regresorius, kaip taisyklė, modelyje paliekamas. Jeigu modelyje būtų statistiškai nereikšmingų kintamųjų, juos reikėtų pabandyti iš modelio pašalinti. Pašalinus bent vieną regresorių, visa analizė pradedama iš pradžių, nes keičiasi ir klasifikacinė lentelė ir visos *p* reikšmės. Regresorių galima šalinti iš modelio, jeigu teisingų klasifikavimų procentas keičiasi nedaug. Mūsų nagrinėjamo pavyzdžio atveju visi regresoriai yra statistiškai reikšmingi. Tai dar vienas įrodymas, kad modelis tinka.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	644.039	.000	0	.
pray	713.761	69.722	2	.000
stfedu	771.837	127.798	2	.000
trstprl	723.584	79.545	2	.000
hhmmb	706.825	62.786	2	.000
imigrantbf	663.956	19.917	2	.000

Lentelėje **Parameter Estimates** yra informacija apie dalinius modelius (submodelius), kai kiekviena kategorija lyginama su kontroline.

Parameter Estimates

cntry Country[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B): Lower Bound	Upper Bound
CZ Czech Republic	Intercept	.502	.977	.263	1	.608			
	pray	.011	.122	.008	1	.927	1.011	.796	1.284
	stfedu	.299	.069	18.908	1	.000	1.348	1.178	1.543
	trstprl	-.487	.062	61.777	1	.000	.614	.544	.694
	hhmmb	.216	.105	4.252	1	.039	1.241	1.011	1.524
	[imigrantbf=.00]	-1.212	.281	18.661	1	.000	.298	.172	.516
	[imigrantbf=1.00]	0[b]	.	.	0	.	.	.	.
IL Israel	Intercept	5.077	.877	33.477	1	.000			
	pray	-.573	.108	28.098	1	.000	.564	.456	.697
	stfedu	-.379	.072	27.700	1	.000	.685	.595	.789
	trstprl	-.250	.065	14.987	1	.000	.779	.686	.884
	hhmmb	.704	.106	43.903	1	.000	2.023	1.642	2.491
	[imigrantbf=.00]	-.821	.306	7.226	1	.007	.440	.242	.801
	[imigrantbf=1.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: SE Sweden.

b. This parameter is set to zero because it is redundant.

Mūsų atveju viršutinė lentelės dalis yra skirta tikimybės, kad respondentas yra iš Čekijos, palyginimui su tikimybe, kad jis – iš Švedijos. Visų pirma, peržvelgę Voldo kriterijaus *p* reikšmes (konstantai p reikšmės nežiūrime), įsitikiname, kad kintamasis *pray* yra statistiškai nereikšmingas. Darome išvadą, kad maldingumo laipsnis nėra reikšmingas faktorius, padedantis atskirti Čekijos

gyventojus nuo Švedijos gyventojų. Ar reikia bandyti iš modelio šalinti kintamąjį *pray*? Nesunku pastebėti, kad šis kintamasis yra statistiškai reikšmingas antrajame modelyje, t.y. padeda atskirti izraelietį nuo Švedijos gyventojo. Todėl tikėtina, kad apskritai modelyje kintamasis *pray* reikalingas. Visi kiti kintamieji statistiškai reikšmingi. Pasižiūrėję į stulpelį B ir prisiminę kaip koduojamos šalys, užrašome pirmąjį dalinį logistinį modelį taip:

$$\frac{\mathrm{P}(cntry = \mathrm{CZ})}{\mathrm{P}(cntry = \mathrm{SE})} = \mathrm{e}^{z_1}.$$

Čia $z_1 = 0{,}502 + 0{,}011pray + 0{,}299stfedu - 0{,}487trstprl + 0{,}216hhmmb$

$$+\begin{cases} 0, & \text{jei } imigrantbf = 1 \\ -1{,}212, & \text{jei } imigrantbf = 0 \end{cases}.$$

Koeficientų ženklai rodo, kad didėjant *stfedu* ir *hhmmb* reikšmėms (respondentui išreiškiant didesnį pasitikėjimą švietimo sistema ir gyvenant didesnėje šeimoje) didėja tikimybė, kad jis iš Čekijos, o ne iš Švedijos. Didėjant pasitikėjimui parlamentu (*trstprl*), tikimybė, kad respondentas iš Čekijos, o ne iš Švedijos mažėja. Galų gale dviguba koeficiento *imigrantbf* reikšmė, rodo, kad repondentui pareiškus palankesnę nuomonę apie socialines lengvatas imigrantams (*imigrantbf* = 0), tikimybė, kad jis iš Čekijos, o ne iš Švedijos mažėja. Šią informaciją galima patikslinti, pateikiant galimybių santykius (stulpelis **Exp(B)**) kartu su jų pasikliautiniais intervalais (stulpelis **95% Confidence Interval for Exp(B)**). Pavyzdžiui, galimybių santykis (angl. odds ratio) kintamjam *stfedu* yra 1,348 (95% pasikl. int. 1,178 – 1,543). Tai reiškia, kad kintamajam *sfedu* padidėjus vienetu tikimybių santykis P(*cntry* = CZ) / P(*cntry* =SE) padidės 1,348 karto:

$$\left(\frac{\mathrm{P}(cntry = \mathrm{CZ})}{\mathrm{P}(cntry = \mathrm{SE})}\right)_{\text{kai } stfedu \text{ padidėja vienetu}} = 1{,}348\left(\frac{\mathrm{P}(cntry = \mathrm{CZ})}{\mathrm{P}(cntry = \mathrm{SE})}\right)_{senas}$$

Analogiškai interpretuojami visi likę galimybių santykiai.

Visiškai analogiškai ištiriamas antrasis lentelės **Parameter Estimates** modelis, skirtas Izraelio ir Švedijos lyginimui. Gauname

$$\frac{\mathrm{P}(cntry = \mathrm{IL})}{\mathrm{P}(cntry = \mathrm{SE})} = \mathrm{e}^{z_2},$$

Čia $z_2 = 5{,}077 - 0{,}573pray - 0{,}379stfedu - 0{,}25trstprl + 0{,}704hhmmb$

$$+\begin{cases} 0, & \text{jei } imigrantbf = 1 \\ -0{,}821, & \text{jei } imigrantbf = 0 \end{cases}.$$

Parametrų ženklai ir galimybių santykiai aptariami analogiškai Čekijos-Švedijos modeliui.

5.2.4. Prognozavimas

Prognozės atliekamos tik esant poreikiui. Atitinkamos tikimybės skaičiuojamos pagal formules pateiktas, ankstesniame skyrelyje. Pateiksime vieną pavyzdį. Tarkime, kad respondentas pasirinko tokius atsakymus: *pray* = 2, *stfedu* = 4, *trstprl* =5, *hhmb* = 4, *imigrantbf* = 1. Įvertinsime tikimybes, kad jis iš Čekijos (Izraelio, Švedijos). Pradžioje pagal aukščiau pateiktas formules surandame z_1 ir z_2:

$$z_1 = 0{,}502 + 0{,}011 * 2 + 0{,}299 * 4 - 0{,}487 * 5 + 0{,}216 * 4 + 0 = 0{,}149,$$

$$z_2 = 5{,}077 - 0{,}573 * 2 - 0{,}379 * 4 - 0{,}25 * 5 + 0{,}704 * 4 + 0 = 3{,}981.$$

Tada suskaičiuojame (primename, kad e = 2,71828....)

$$\mathrm{e}^{z_1} = 1{,}160 \ldots, \qquad \mathrm{e}^{z_2} = 53{,}570, \quad \mathrm{e}^{z_1} + \mathrm{e}^{z_2} = 54{,}73\,.$$

Beliko įvertinti tikimybes:

$$\mathrm{P}(cntry = \mathrm{SE}) = \frac{1}{1 + 54{,}73} = 0{,}01794 \ldots, \qquad \mathrm{P}(cntry = \mathrm{CZ}) = \frac{1{,}16}{1 + 54{,}73} = 0{,}0208 \ldots$$

$$\mathrm{P}(cntry = \mathrm{IL}) = \frac{53{,}57}{1 + 54{,}73} = 0{,}9612..$$

Matome, kad labai didelė tikimybė, jog atsakinėjo izraelietis.

5.2.5. Rezultatų aprašymas

Mes tikėjomės, kad požiūris į šalies švietimo būseną, imigrantus, parlamentą bei pamaldumas gerai charakterizuoja respondento šalį. Šis spėjimas pasitvirtino. Pateikdami aprašymus, pradžioje nurodome ką tyrėme, kokius kintamuosius naudojome ir kaip juos perkodavome. Tada tiesiog kompaktiškai surašome visa tai, ką išsamiai aptarėme ankstesniame skyrelyje. Pateikiame vieną iš daugybės galimų tokio aprašymo variantų.

Taikėme daugianarę logistinę regresiją. Kontroline kategorija pasirinkome Švediją – visuotinos socialinės gerovės šalį. Modelis gerai tiko duomenims. Buvo teisingai klasifikuota 70,3% Čekijos, 80,1% Izraelio ir 67,6% Švedijos gyventojų. Bendras teisingai klasifikuotų imties respondentų procentas 73,7%. Nagelkerkės determinacijos pseudokoeficientas $R^2 = 0{,}65$. Modelio didžiausio tikėtinumo santykio chi kvadrato statistika $\chi^2 = 446{,}337$, $p < 0{,}01$. Pagal didžiausio tikėtinumo santykio chi kvadrato kriterijų visi regresoriai yra statistiškai reikšmingi.

Pritaikius Voldo kriterijų, paaiškėjo, kad submodelyje, skirtame Izraelio ir Švedijos respondentų skirtumams visi kintamieji (*pray, stfedu, trsprl, hhmmb, imigrantbf*) yra statistiškai reikšmingi. Tikimybių santykio logaritmo lygtis yra lygi

$$\ln \frac{\mathrm{P}(cntry = \mathrm{IL})}{\mathrm{P}(cntry = \mathrm{SE})} = 5{,}077 - 0{,}573pray - 0{,}379stfedu - 0{,}25trstprl + 0{,}704hhmmb$$

$$+\begin{cases}0, & \text{jei } imigrantbf = 1\\ -0{,}821, & \text{jei } imigrantbf = 0\end{cases}$$

Matome, kad mažėjant pamaldumui ir augant pasitikėjimui parlamentu bei švietimo sistema (didėjant *pra, stfedu* ir *trstprl*) ir, esant teigiamam požiūriui į imigrantus, mažėja tikimybė, kad respondentas iš Izraelio, o ne iš Švedijos. O priklausymas didesnėms šeimoms (*hhmmb*) tikimybę, kad respondentas iš Izraelio, o ne iš Švedijos didina. Galimybių santykiai ir jų 95% pasikliautinieji intervalai yra tokie: kintamajam *pray* – 0,564 ([0,456; 0,697]), kintamajam *stfedu* – 0,685 ([0,595; 0,697]), kintamajam *trstprl* – 0,779 ([0,686; 0,884]), kintamajam *hhmmb* – 2,023 ([1,642; 2,491]). Lyginant su nepalankia nuomone apie imigrantus (*imigrantbf* = 1), palanki nuomonė (*imigrantbf* = 0) sumažina galimybę, kad respondentas iš Izraelio, o ne iš Švedijos, 0,44 karto (95% galimybių santykio pasikliautinis intervalas [0,242; 0,801]).

Pritaikius Voldo kriterijų, paaiškėjo, kad submodelyje, skirtame Čekijos ir Švedijos respondentų skirtumams visi kintamieji išskyrus pamaldumą (*pray*) yra statistiškai reikšmingi. Tikimybių santykio logaritmo lygtis yra lygi

$$\ln \frac{\mathrm{P}(cntry = \mathrm{CZ})}{\mathrm{P}(cntry = \mathrm{SE})} = 0{,}502 + 0{,}011pray + 0{,}299stfedu - 0{,}487trstprl + 0{,}216hhmmb$$

$$+\begin{cases}0, & \text{jei } imigrantbf = 1\\ -1{,}212, & \text{jei } imigrantbf = 0\end{cases}.$$

Matome, kad augant pasitikėjimui švietimo sistema ir gyvenant didesnėje šeimoje (didėjant *hhmmb* ir *stfedu*), didėja tikimybė, kad respondentas iš Čekijos, o ne iš Švedijos. O teigiamas požiūris į parlamentą ir į imigrantus tikimybę, kad respondentas iš Čekijos, o ne iš Švedijos mažina. Galimybių santykiai ir jų 95% pasikliautinieji intervalai yra tokie: kintamajam *stfedu* – 1,348 ([1,178; 1,543]), kintamajam *trstprl* – 0,614 ([0,544; 0,694]), kintamajam *hhmmb* – 1,241 ([1,011; 1,524]). Lyginant su nepalankia nuomone apie imigrantus (*imigrantbf* = 1), palanki nuomonė (*imigrantbf* = 0) sumažina galimybę, kad respondentas iš Čekijos, o ne iš Švedijos, 0,298 karto (95% galimybių santykio pasikliautinis intervalas [0,172; 0,516]).

5.2.6. SPSS komandų sintaksė daugianarei logistinei regresinei analizei

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam daugianarės logistinės regresijos modeliui. Jas galima naudoti sintaksės lange, pasirenkant opciją **Run**.

Kintamojo *imigrantbf* kūrimas:

```
RECODE imsclbn (1 thru 3=0) (4 thru 5=1) INTO imigrantbf.
```

```
EXECUTE.
```

Tinkamų respondentų atranka:

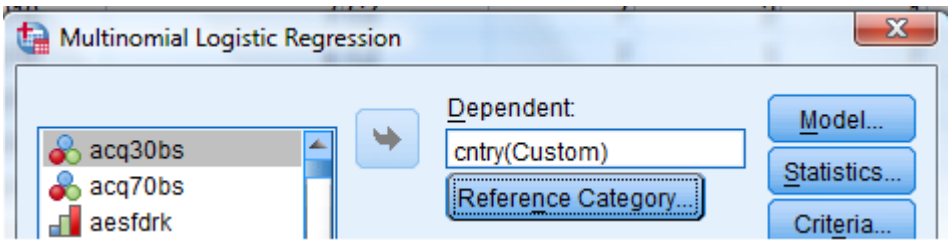
```
USE ALL.
COMPUTE filter_$=(agea >= 20 & agea <= 30 & gndr = 1).
VARIABLE LABELS filter_$ 'agea >= 20 & agea <= 30 & gndr = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

Daugianarė logistinė regresija:

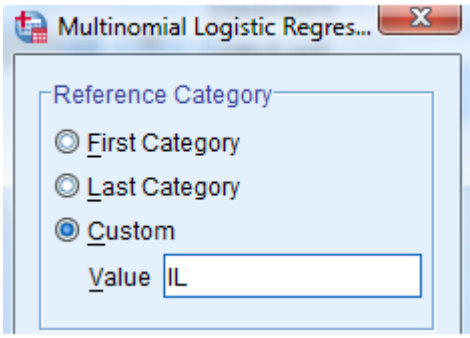
```
NOMREG cntry (BASE=LAST ORDER=ASCENDING) WITH pray stfedu hhmmb trstprl
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5)
 CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE)
ENTRYMETHOD(LR) REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
 /PRINT=CLASSTABLE FIT PARAMETER
SUMMARY LRT CPS STEP MFI.
```

5.2.7. Kitos kontrolinės kategorijos pasirinkimas

Jeigu norime lyginimams pasirinkti kitą priklausomo kintamojo kategoriją, reikia meniu **Multinomial Logistic Regression** paspausti klavišą **Reference Category.**



Atsidariusiame lange renkamės **Custom** ir langelyje **Value** nurodome norimą priklausomo kintamojo kodą. Pavyzdžiui, norėdami kitas valstybes lyginti su Izraeliu, rašome IL.



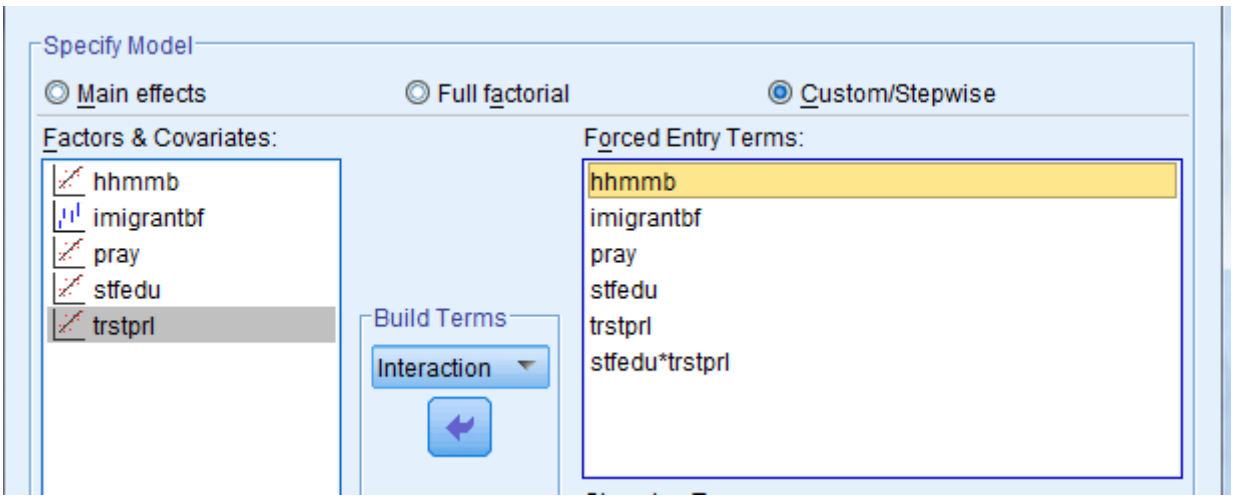
Beje, visų statistikų ir klasifikacinių lentelių reikšmės nepriklauso nuo kontrolinės kategorijos pasirinkimo. Keičiasi tik lentelė **Parameter Estimates**.

5.2.8. Kintamųjų sąveikos įtraukimas

Analogiškai dvinarei logistinei regresijai į modelį galima įtraukti ir regresorių sandaugas. Tada sakoma, kad modelyje buvo atsižvelgta į kintamųjų tarpusavio sąveiką. Tarkime, kad norime ištirti modelį

*cntry = f (imigrantbf, pray, stfedu, trstprl, hhmmb, stfedu*trstprl).*

Visos SPSS parinktys išlieka tokios pačios, kaip ir pirminiame modelyje. Vienintelis skirtumas tas, kad pagrindiniame meniu pasirenkame **Model**. Atsidariusiame lange pažymime **Custom Stepwise** ir į langelį **Forced Entry Terms** sukeliame visus kintamuosius. Kintamųjų sandauga atsiras, kai kelsime *stfedu* ir *trstprl* kartu. Pasirenkame **Continue** ir **OK.**



Rezultatų išklotinėje matyti, kad regresorių sąveika statistiškai nereikšminga.

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	641.201[a]	.000	0	.
hhmmb	704.311	63.111	2	.000
imigrantbf	661.190	19.989	2	.000
pray	709.903	68.702	2	.000
stfedu	707.976	66.775	2	.000
trstprl	653.156	11.955	2	.003
stfedu * trstprl	644.039	2.838	2	.242

Beje, Voldo testai abiem modeliams (su interakcija ir be interakcijos) tokie patys. Klasifikacinė lentelė irgi nepagerėja. Darome išvadą, kad regresorių sąveika *stfedu*trstprl* modelyje nereikalinga.

5.2.9. Žingsninė daugianarė logistinė regresija

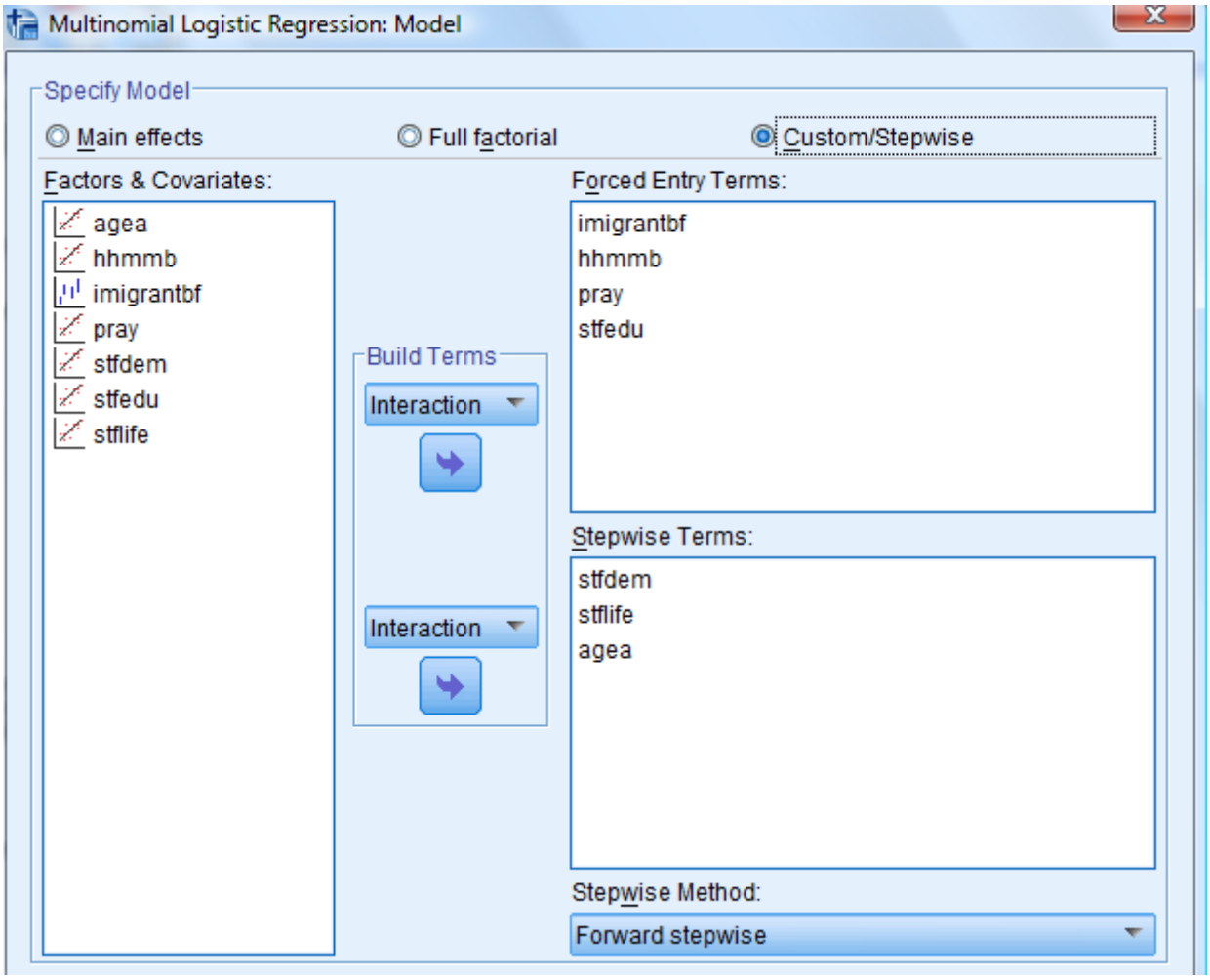
Kaip ir kituose regresijos modeliuose galima taikyti žingsninę (keliaetapę) regresiją, ir leisti modelį tobulinti SPSS programai. Primename, kad tada modelis bus tobulinamas pagal formalius statistinius rodiklius. Gautas galutinis modelis bus statistiškai geriausias, bet nebūtinai geriausias tyrimo prasme. Yra keli žingsninės regresijos variantai – kai modelis vis pildomas naujais regresoriais, arba kai jie iš modelio šalinami. Be to, žingsninę regresiją galima taikyti tik daliai regresorių, kuriais norime papildyti jau turimą modelį. Ją šiame skyrelyje ir aptarsime. Tarkime, kad į modelį

cntry = f (imigrantbf, pray, stfedu, hhmmb)

norime įtraukti kintamuosius

- *stfdem* (pasitenkinimas demokratijos lygiu šalyje),
- *stflife* (pasitenkinimas gyvenimu),
- *agea* (amžių).

Taikysime žingsninę regresiją. Visi meniu užpildomi, kaip ir pirminio tyrimo atveju. Skirtumai atsiranda tik **Model** lange. Regresorius *imigrantbf, pray, stfedu, hhmml* perkeliame į lauką **Forced Entry Terms.** Šie regresoriai bus visuose modeliuose. Regresorius *stfdem*, *stflife*, *agea* perkeliame į langą **Stepwise Terms**. Šiais regresoriais modelis bus po truputį pildomas. Statistiškai nereikšmingi regresoriai į modelį nebus įtraukiami. Pasirenkame **Stepwise Method: Forward stepwise**.



Gautoje rezultatų išklotinėje atsirado lentelė **Step Summary.**

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
				-2 Log Likelihood	Chi-Square[a,b]	df	Sig.
Step 0	0	Entered	Intercept, imigrantbf, hhmmb, pray, stfedu	756.278	.		
Step 1	1	Entered	Stfdem	708.037	48.241	2	.000
Step 2	2	Entered	Stflife	691.149	16.888	2	.000

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the likelihood ratio test.

b. The chi-square for removal is based on the likelihood ratio test.

Lentelėje pažymėta, kad pradinis modelis su regresoriais *imigrantbf, pray, stfedu, hhmml* (Step 0) buvo pradžioje papildytas regresoriumi *stfdem* (Step 1), o po to – regresoriumi *stflife* (Step 2). Regresorius *agea* buvo nepanaudotas. Tai reiškia, kad įtraukus *agea* į modelį, jis buvo statistiškai nereikšmingas. Galutinis modelis

cntry = f (imigrantbf, pray, stfedu, hhmml, stfdem, stflife).

Visos jo charakteristikos pateikiamos lentelėse, kurios nė kiek nesiskiria nuo aptartųjų skyrelyje 4.2.3. Beje, lange **Stepwise Terms** galima patalpinti ir regresorių sandaugas.

5.3. Daugianarė logistinė regresinė analizė su STATA

5.3.1. Duomenys

Failas **ZTLT** . Tyrimas: Lietuvos tolerancijos profiliai, 2003 m. Respondentai atsakė, koks jų požiūris (1 – teigiamas, 2 – neutralus, 3 – neigiamas, 4 – sunku pasakyti)

- *r01_01* – į juodaodžius,
- *r01_02* – į musulmonus,
- *r01_03* – į žydus,
- *r01_04* – į čigonus (romus),
- *r01_05* – į migrantus,
- *r01_06* – į pabėgėlius.

Norėsime ištirti, kaip požiūrį į romus įtakoja požiūris į kitus nelietuvius. Visų pirma, pasižiūrime, kiek kokių reikšmių šis kintamasis įgyja. Įvykdome komandą

```
tabulate r01_04
```

```
       grupes? |      Freq.     Percent        Cum.
---------------+-----------------------------------
     Teigiamas |        147       14.08       14.08
     Neutralus |        402       38.51       52.59
     Neigiamas |        446       42.72       95.31
Sunku pasakyti |         49        4.69      100.00
---------------+-----------------------------------
         Total |      1,044      100.00
```

Matome, kad nuomonės neturėjo tik 4.69% respondentų. Panaši situacija su atsakymais į kitus klausimus. Todėl iš tyrimo pašaliname visus respondentus, neturinčius nuomonės apie svetimtaučius:

```
keep if  r01_01<=3 & r01_02 <=3 & r01_03 <=3 &  r01_04 <=3 & r01_05 <=3 &  r01_06 <=3
```

Sudėdami visus kintamuosisu, išskyrus *r01_04,* sukuriame jungtinį kintamąjį *kiti.* Šis kintamsis atspindi respondentų priešiškumą, visiems nelietuviams, išskyrus romus. Kuo didesnė kintamojo *kiti* reikšmė, tuo priešiškumas yra didesnis:

```
gen kiti =  r01_01+ r01_02 + r01_03+ r01_05+ r01_06
```

5.3.2. STATA parinktys

Tikrinsime modelį: *r01_04 = f(kiti)*. Kontroline kategorija pasirenkame reikšmę *r01_04* = 3 (neigiamą požiūrį). Įvykdome

```
mlogit  r01_04 kiti, base(3)
```

```
Multinomial logistic regression                   Number of obs   =        863
                                                  LR chi2(2)      =     478.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -640.75012                       Pseudo R2       =     0.2717

------------------------------------------------------------------------------
      r01_04 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Teigiamas    |
        kiti |   -1.28939   .0833927   -15.46   0.000    -1.452836   -1.125943
       _cons |   11.32658   .7826248    14.47   0.000     9.792668     12.8605
-------------+----------------------------------------------------------------
Neutralus    |
        kiti |  -.4749833   .0498444    -9.53   0.000    -.5726766     -.37729
       _cons |   5.285512   .5570161     9.49   0.000     4.193781    6.377244
------------------------------------------------------------------------------
(r01_04==Neigiamas is the base outcome)
```

Modelio tikimo didžiausio tikėtinumo santykio chi kvadrato statistikos *p* reikšmė yra grafoje **Prob > chi2.** Darome išvadą, kad modelis duomenims tinka, nes p = 0,000 < 0,05. Pateiktasis (Makfadeno) determinacijos pseudokoeficientas nėra itin didelis R^2 = 0,271, bet jo vaidmuo nėra itin svarbus (ir jis vis tiek > 0,20). Voldo testai rodo, kad visuose submodeliuose tik konstanta kartais yra statistiškai nereikšminga. Užrašykime pirmąjį dalinį modelį:

$$\frac{\mathrm{P(Teigiamas)}}{\mathrm{P(Neigiamas)}} = \mathrm{e}^{z_1}.$$

Čia $z_1 = 11{,}326 - 1{,}289kiti.$

Regresoriaus *kiti* koeficiento ženklas yra neigiamas. Todėl, didėjant *kiti* reikšmėms (respondentui išreiškiant didesnį priešiškumą nelietuviams), didėja tikimybė, kad ir romų atžvilgiu jis nusistatęs neigiamai. Analogiškai galima sudaryti ir pakomentuoti antrąjį modelį:

$$\frac{\mathrm{P(Neutralus)}}{\mathrm{P(Neigiamas)}} = \mathrm{e}^{z_2}.$$

Čia $z_2 = 5{,}285 - 0{,}7498kiti.$

Norėdami gauti galimybių santykius, įvykdome:

```
mlogit  r01_04 kiti, base(3) rrr
```

r01_04	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Teigiamas						
kiti	.2754389	.0229696	-15.46	0.000	.2339059	.3243465
Neutralus						
kiti	.6218955	.030998	-9.53	0.000	.5640138	.6857172

(r01_04==Neigiamas is the base outcome)

Galimybių santykiai pateikti kartu su jų pasikliautiniais intervalais. Paimkime pirmąjį dalinį modelį. Galimybių santykis regresoriui *kiti* lygus 0,275 (95% pasikliautinis intervalas 0,234 – 0,324). Tai reiškia, kad kintamajam *kiti* padidėjus vienetu, tikimybių santykis P(Teigiamas) / P(Neigimas) sumažės 0,275 karto:

$$\left(\frac{\mathrm{P(Teigiamas)}}{\mathrm{P(Neigiamas)}}\right)_{\text{kai } kiti \text{ padidėja vienetu}} = 0{,}275\left(\frac{\mathrm{P(Teigiamas)}}{\mathrm{P(Neigiamas)}}\right)_{senas}$$

Analogiškai interpretuojamas ir antrasis galimybių santykis.

Norėdami pamatyti alternatyvius determinacijos koeficientus, Akaikės indeksą ir pan. įvykdome:

```
fitstat
```

```
Measures of Fit for mlogit of r01_04

Log-Lik Intercept Only:       -879.802   Log-Lik Full Model:           -640.750
D(859):                       1281.500   LR(2):                         478.104
                                         Prob > LR:                       0.000
McFadden's R2:                   0.272   McFadden's Adj R2:               0.267
ML (Cox-Snell) R2:               0.425   Cragg-Uhler(Nagelkerke) R2:      0.489
Count R2:                        0.673   Adj Count R2:                    0.429
AIC:                             1.494   AIC*n:                        1289.500
BIC:                         -4525.696   BIC':                         -464.583
BIC used by Stata:            1308.542   AIC used by Stata:            1289.500
```

5.3.3. Prognozavimas

Prognozavimui patogiausia naudoti SPost priedo komandą *prvalue*. Tarkime, kad norime padaryti *r01_04* prognozę respondentui, kurio *kiti* = 10. Įvykdome

```
prvalue, x(kiti=10)
```

```
Confidence intervals by delta method

                                 95% Conf. Interval
  Pr(y=Teigiama|x):   0.0715   [ 0.0485,    0.0946]
  Pr(y=Neutralu|x):   0.5857   [ 0.5426,    0.6288]
  Pr(y=Neigiama|x):   0.3428   [ 0.3011,    0.3845]

    kiti
x=    10
```

Jeigu regresorių būtų daugiau, tai nurodytume daugiau reikšmių.

5.3.4. Klasifikacinė lentelė

Tiesiogiai klasifikacinės lentelės gauti su STATA 10 negalima. Todėl ją sukursime, įdėję šiek tiek darbo. Visų pirma, kiekvienam respondentui, taikydami logistinės regresijos modelį, sukuriame ir išsaugome visų kategorijų tikimybes:

```
predict pre1 pre2 pre3
```

Duomenyse atsirado trys nauji stulpeliai. Stulpelyje *pre1* kiekvienam respondentui suskaičiuota tikimybė, kad apie romus išreiškė teigiamą požiūrį. Analogiškai, stulpelyje *pre2* yra tikimybės, kad požiūris buvo neutralus, o *pre3* – neigiamas. Sukuriame „tuščią“ stulpelį:

```
gen cat = 0
```

Dabar nuosekliai pakeičiame nulius į 1, kai labiausiai tikėtina teigiamas požiūris. Po to, pakeičiame nulius į 2, kai labiausiai tikėtinas neutralus požiūris ir į 3, kai požiūris neigiamas.

```
replace cat = 1 if pre1>pre2 & pre1 > pre3
replace cat = 2 if pre2>pre1 & pre2 > pre3
replace cat = 3 if pre3>pre1 & pre3 > pre2
```

Suteikiame reikšmėms paaiškinamuosius komentarus. Įvykdome:

```
label define nopsas 1 "Teigiamas" 2 "Neutralus" 3 "Neigiamas"
label values cat nopsas
```

Dabar jau galime sudaryti klasifikacinę lentelę. Įvykdome:

```
tab r01_04 cat, row
```

Poziuris i zmones	cat Teigiamas	Neutralus	Neigiamas	Total
Teigiamas	101 74.26	29 21.32	6 4.41	136 100.00
Neutralus	18 4.88	274 74.25	77 20.87	369 100.00
Neigiamas	12 3.35	140 39.11	206 57.54	358 100.00
Total	131 15.18	443 51.33	289 33.49	863 100.00

Matome, kad iš 136 respondentų, kurie pareiškė teigiamą požiūrį į romus, teisingai klasifikuoti 101 (74,26%) . Teisingai buvo klasifikuoti 74,25% neutralų požiūrį išreiškę respondentai ir 57,54% – išreiškę neigiamą požiūrį. Palyginkime su pradine nuomonių dažnių lentele:

```
tab r01_04, freq
```

grupes?	Freq.	Percent	Cum.
Teigiamas	136	15.76	15.76
Neutralus	369	42.76	58.52
Neigiamas	358	41.48	100.00
Total	863	100.00	

Klasifikacinės lentelės procentai daug didesni, nei gautume aklai spėliodami. Taigi modelį galima laikyti visai tinkamu.

5.3.5. Kategorinis kintamasis

Kategorinis kintamasis turi būti keičiamas dvireikšmiais pseudokintamaisiais. Tą galima padaryti automatiškai, tačiau viena ir kategorinio kintamojo reikšmių turi būti lygi nuliui. Ištirsime

daugianarės logistinės regresijos modelį *r01_04 = f(kiti, s11)*. Kintamajame *s11* nurodyta respondento gyvenamoji vieta (1 – Vilnius, 2 – Kaunas, Klaipėda, Šiauliai, Panevėžys, 3 – rajono centras, 4 – kiti miestai, 5 – kaimas). Tarp *s11* reikšmių nėra nulio, todėl sukuriame surogatinį kintamąjį *zzz* = *s11* – 1. Įvykdome:

```
gen zzz=s11-1
```

Daugianarė regresija atliekama įvykdžiui:

```
xi:mlogit r01_04 kiti i.zzz
```

```
Multinomial logistic regression                   Number of obs   =        863
                                                  LR chi2(10)     =     494.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -632.38796                       Pseudo R2       =     0.2812

------------------------------------------------------------------------------
      r01_04 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Teigiamas    |
        kiti |  -.8333296   .0722538   -11.53   0.000    -.9749445   -.6917148
     _Izzz_1 |   .3625495   .4233289     0.86   0.392    -.4671599    1.192259
     _Izzz_2 |   .9054994   .4391138     2.06   0.039     .0448521    1.766147
     _Izzz_3 |    1.36664   .9268267     1.47   0.140    -.4499069    3.183187
     _Izzz_4 |   .2707802   .4247504     0.64   0.524    -.5617153    1.103276
       _cons |   5.771609    .632821     9.12   0.000     4.531303    7.011916
-------------+----------------------------------------------------------------
Neigiamas    |
        kiti |   .4934852   .0509341     9.69   0.000     .3936562    .5933143
     _Izzz_1 |  -.4684803   .2538883    -1.85   0.065    -.9660923    .0291317
     _Izzz_2 |   -.299621   .2555699    -1.17   0.241    -.8005288    .2012868
     _Izzz_3 |   1.650373   .9318776     1.77   0.077    -.1760737    3.476819
     _Izzz_4 |  -.5188331   .2421728    -2.14   0.032    -.9934831   -.0441831
       _cons |  -5.146229   .5764447    -8.93   0.000     -6.27604   -4.016418
------------------------------------------------------------------------------
(r01_04==Neutralus is the base outcome)
```

Kategorinis regresoriu *zzz* buvo pakeistas dvireikšmiais regresoriais, kurie įgyja nulius ir vienetus: *_Izzz_1* = 1, kai *zzz* = 1, *_Izzz_2* = 1, kai *zzz* = 2, ir t.t. Bendroji formulė pirmajam daliniam modeliui atrodo taip:

$$\frac{\mathrm{P(Teigiamas)}}{\mathrm{P(Neigiamas)}} = \mathrm{e}^{z_1}.$$

Čia

$z_1 = 5{,}77 - 0{,}83\ kiti + 0{,}36\ \_Izzz\_1 + 0{,}91\ \_Izzz\_2 + 1{,}36\ \_Izzz\_3 + 0{,}27\ \_Izzz\_4.$

Norėdami gauti lygtį vilniečiams, vietoje visų *_Izzz_j* įstatome nulius. Norėdami gauti lygtį kaimui, įstatome *_Izzz_1* = *_Izzz_2* = *_Izzz_3* = 0, *_Izzz_4* = 1 ir t.t., žr. 5.2.4 skyrelį.

Uždaviniai

1. Failas **ZTLT** . Tyrimas: Lietuvos tolerancijos profiliai, 2003 m. Respondentai atsakė, koks jų požiūris (1 – teigiamas, 2 – neutralus, 3 – neigiamas, 4 – sunku pasakyti) į musulmonus (kintamasis *r01_02*). Kintamasis *r02_04* (koks procentas Lietuvos gyventojų būdami homoseksualūs nekeltų grėsmės visuomenei) yra ranginis, bet rangų daug. Todėl tyrime jį galima traktuoti, kaip intervalinį. Atsakymai į klausimą, ar galima pateisinti darbdavį, atsisakiusį priimti į darbą kitos rasės žmogų (*r06_01*) ir homoseksualų žmogų (*r06_07*) koduoti skalėje: 1 – Niekada, 2 – Kartais, 3 – Visada, 4 – sunku pasakyti. Kintamasis *s01* žymi lytį (1 – vyr., 2 – mot.). Reikia perkoduoti kintamuosius *r06_01* ir *r06_07* į *ras* ir *homo* : 1 → 1, 2 → 2, 3 → 2, 4 → 3. Suteikti naujosioms reikšmėms „etiketes“ : 1 – niekada, 2 – kartais/visada, 3 – sunku pasakyti. Vyrams ištirti daugianarės logistinės regresijos modelį: *homo = f(ras,r01_02,r02_04).*

2. Failai **ESS4LT1, ESS4PT, ESS4IL**. Kintamieji: *cntry* – šalies kodas, *agea* – amžius, *gndr* – lytis (1 – vyr., 2 – mot.), *pray* – kaip dažnai meldžiamasi (matuojamas 7 balų skale nuo 1 – kasdien iki 7 – niekada), *hhmmb* – pastoviai kartu gyvenančių namų ūkio narių skaičius,

- *stfedu, stflife, stfeco, stfgov, stfdem, stfhlth* – respondento požiūris į savo šalies švietimo sistemos būklę, savo gyvenimą, ekonominę padėtį, vyriausybę, demokratijos būklę, sviekatos apsaugą (reikšmės nuo 0 – itin bloga, iki 10 – itin gera),
- *trstprl, trstlgl, trstplc, trstplt, trstprt, trstep, trstun* – pasitikėjimas savo šalies parlamentu, teisine sistema, policija, politikais, partijomis, Europarlamentu, Jungtinėmis tautomis, (matuojamas 10 balų skale nuo 0 – itin nepasitikiu, iki 9 – itin pasitikiu),

Pritaikius daugianarę logistinę regresiją, ištirti kokie kintamieji padeda charakterizuoti respondentės šalį. Tirti 35 – 40 metų moteris.

3. Pritaikyti daugianarės logistinės regresijos modelį 6.2 skyrelio problemai. Kuris modelis geriau tinka?

dvinarė daugianarė

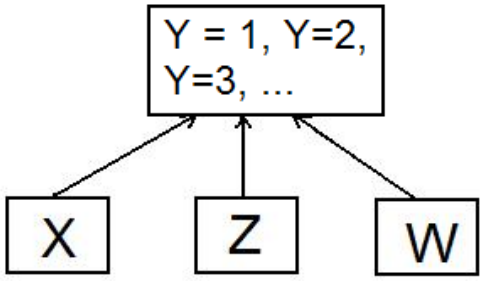
Logistinės regresijos

6. RANGINĖ LOGISTINĖ REGRESINĖ ANALIZĖ

Tipinės situacijos, kai taikoma ranginė logistinė regresinė analizė yra šios:

- Norime įvertinti, kaip respondento požiūris į gyvenimą (nykuma – rutina – smaguma) priklauso nuo jo išsilavinimo, lyties ir turtinės padėties.
- Tiriame, kas charakterizuoja pritarimą vykdomai užsienio politikai (visai nepritariu – labiau nepritariu, nei pritariu – labiau pritariu, nei nepritariu – pritariu).
- Manome, kad teniso turnyruose užimta prizinė vieta priklauso nuo vidutinio smūgiuoto kamuoliuko greičio.

Ranginė logistinė regresija dar vadinama daugialyge logistine regresija arba proporcingų galimybių (angl. proportional odds) modeliu. Kartais naudojamas trumpesnis terminas – ranginė regresija, praleidžiant žodį logistinė. Modelio schema:



Ranginis kintamasis *Y* vadinamas *priklausomu* (regresuojamu) *kintamuoju*, kintamieji *X, Z, W* vadinami *nepriklausomais kintamaisiais* (*regresoriais*). Ranginė logistinė regresija yra daugianarės logistinės regresijos analogas, kai išnaudojama informacija apie *Y* ranginę struktūrą.

6.1. Ranginės logistinės regresijos modelis

6.1.1. Modelio lygtys

Analogiškai daugianarei regresijai ranginė logistinė regresija leidžia įvertinti *Y* reikšmių įgijimo tikimybes. Matematinis modelis aprašomas keliomis lygtimis. Nors kintamojo *Y* reikšmės gali būti ir simbolinės, paprasčiau, kai jos yra skaitinės. Todėl toliau, aprašydami teorinį modelį remsimės pavyzdžiu, kai priklausomas kintamasis *Y* įgyja 4 rangines reikšmes ($Y = 1, 2, 3, 4$). Visi modeliai sudaromi tam tikrų tikimybių santykiams, pasirenant įvairias vadinamasias *jungties* funkcijas. Tų funkcijų yra įvairių. Aprašysime pačią populiariausią *logit* funkciją. Tarkime, kad *Y* priklauso nuo intervalinių arba dvireikšmių regresorių *X, Z, W*. Sudaromi trys matematiniai modeliai priklausomo kintamojo tikimybių santykių logaritmams (logit funkcijoms):

$$\ln\frac{\mathrm{P}(Y \le i)}{\mathrm{P}(Y > i)} = C_i - b_1\,X - \ b_2Z - b_3W, \ \ i = 1, 2, 3,$$

t.y.

$$\ln\frac{\mathrm{P}(Y \le 1)}{\mathrm{P}(Y > 1)} = C_1 - b_1\,X - b_2Z - b_3W, \ \ \ln\frac{\mathrm{P}(Y \le 2)}{\mathrm{P}(Y > 2)} = C_2 - b_1\,X - b_2Z - b_3W,$$

$$\ln\frac{\mathrm{P}(Y \le 3)}{\mathrm{P}(Y > 3)} = C_3 - b_1\,X - b_2Z - b_3W.$$

Konstantos $C_1, C_2, C_3, b_1, b_2, b_3$ nėra žinomos. Jų įverčiai $\hat{C}_1, \hat{C}_2, \hat{C}_3, \hat{b}_1, \hat{b}_2, \hat{b}_3$ gaunami, panaudojus imties duomenis. Atkreipiame dėmesį, kad visose lygtyse regresorių koeficientai yra tie patys (pvz., visose lygtyse prie *X* yra tas pats b_1) , o skiriasi tik konstantos C_1, C_2, C_3. Be to, kiek neįprastai, prieš regresorius yra minuso ženklai. Kartais sudaromi modeliai atvirkštiniam tikimybių santykiui. Tada visi teorinio modelio koeficientai užrašomi be minuso ženklų. Modelio koeficientai $C_1, C_2, C_3, b_1, b_2, b_3$ naudojami prognozavimui, galimybių santykiams ir kintamųjų įtakos interpretavimui.

Tarkime, kad gavome tokią lygtį:

$$\ln\frac{\mathrm{P}(Y \le i)}{\mathrm{P}(Y > i)} = C_i + 3\,X - \ 4Z.$$

Čia *i* – bet kuri *Y* reikšmė.Tada, didėjant *X* , didėja tikmybė, kad *Y* įgis mažesnes reikšmes. Analogiškai, didėjant *Z* , didėja tikimybė, kad *Y* įgis didesnes reikšmes (didėja ir tikimybė P(*Y* > *i*), lyginant ją su P(*Y*≤ *i*)).

6.1.2. Galimybės ir jų santykis

Analogiškai dvinarei logistinei regresijai galima apibrėžti *galimybę* (angl. odds), t.y. tikimybių santykį P(*Y* ≤*j*) / P(*Y*> *j*). Žinoma, galima apibrėžti ir atvirkščią galimybę.

Galimybė P(*Y* ≤*j*) / P(*Y*> *j*) parodo, kiek kartų kintamajam *Y* yra tikėtiniau įgyti mažesnę, nei didesnę reikšmę.

Galimybė P(*Y*> *j*) / P(*Y* ≤*j*) parodo, kiek kartų kintamajam *Y* yra tikėtiniau įgyti didesnę, nei mažesnę reikšmę

Galimybių santykis (angl. odds ratio) parodo, kaip pasikeis galimybė (tikimybių santykis), kai atitinkamas kintamasis padidės vienetu, fiksavus kitų kintamųjų reikšmes. Galimybių santykis

priklauso tik nuo modelio koeficientų. Tarkime, kad nagrinėjame P(Y> j) / P($Y \leq j$). Tada, regresoriui X padidėjus vienetu, šis santykis pasikeičia per $e^{\hat{b}_1}$:

$$\left(\frac{P(Y > j)}{P(Y \leq j)}\right)_{naujas} = e^{\hat{b}_1} \left(\frac{P(Y > j)}{P(Y \leq j)}\right)_{senas}$$

Daugiklis $e^{\hat{b}_1}$, kuris ir yra vadinamas galimybių santykiu, yra kiekybinis kintamojo X įtakos modelyje įvertinimas. Jeigu $\hat{b}_1 > 0$, tai žinome, kad X didėjant, taip pat didėja tikimybė, kad Y įgis didesnes reikšmes. Tai kokybinė charakteristika. O kiek padidės ta tikimybė? Atsakyti į šį klausimą neįmanoma, nežinant pačios pradinės tikimybės. Galimybių santykis yra kompromisinis sprendimas – parodo, kiek pasikeis tikimybių santykis. Pavyzdžiui, jeigu $e^{\hat{b}_1} = 2$, tai santykis padvigubės. Tai **nereiškia**, kad tikimybė $P(Y > j)$ padvigubės. Tiesą sakant, šiuo konkrečiu atveju naujoji tikimybė visada padidės mažiau, nei dvigubai. Jei senoji tikimybė yra 0,5 (santykis 1:1), tai naujoji bus 2/3 (santykis 2:1); jei senoji tikimybė 0,25 (santykis 1:3), tai naujoji 0,40 (santykis 2:3) ir pan.

6.1.3. Prognozavimas

Atkreipiame dėmesį į tai, kad P(Y > i) = 1 – P($Y \leq$ i). Todėl, naudojantis aprašytaisiais modeliais, nesunku rasti P($Y \leq$ i). O šias tikimybes jau galima panaudoti, skaičiuojant P(Y = i). Kadangi Y įgyja tik sveikas reikšmes, tai nesunku suprasti, kad

$$\boxed{P(Y= j) = P(Y \leq j) - P(Y \leq j - 1).}$$

Konkrečioms X, Z, W reikšmėms tai daroma taip: surandame tikimybių santykių logaritmus z_1, z_2, z_3 ir pagal formules

$$z_1 = \hat{C}_1 - \hat{b}_1 X - \hat{b}_2 Z - \hat{b}_3 W, \;\; z_2 = \hat{C}_2 - \hat{b}_1 X - \hat{b}_2 Z - \hat{b}_3 W, z_3 = \hat{C}_3 - \hat{b}_1 X - \hat{b}_2 Z - \hat{b}_3 W$$

surandame konkrečias jų reikšmes. Tam užtenka įstatyti tas X, Z ir W reikšmes, kurioms norime gauti prognozę. Tikimybių įverčiai skaičiuojami pagal formules:

$$\boxed{\hat{P}(Y = 1) = \frac{1}{1 + e^{-z_1}}, \;\; \hat{P}(Y \leq 2) = \frac{1}{1 + e^{-z_2}}, \hat{P}(Y \leq 3) = \frac{1}{1 + e^{-z_3}}.}$$

Čia e = 2,7183...

Tarkime, kad norime rasti $\hat{P}(Y = 1)$, $\hat{P}(Y = 2)$, $\hat{P}(Y = 3)$, $\hat{P}(Y = 4)$. Tikimybių įvertčius $\hat{P}(Y = 1)$, $\hat{P}(Y \leq 2)$, $\hat{P}(Y \leq 3)$ apskaičiuojame pagal aukščiau pateiktas formules. Tada pasinaudojame formulėmis:

$$\hat{P}(Y = 2) = \hat{P}(Y \leq 2) - \hat{P}(Y = 1), \qquad \hat{P}(Y = 3) = \hat{P}(Y \leq 3) - \hat{P}(Y \leq 2),$$

$$\hat{P}(Y = 4) = 1 - \hat{P}(Y \leq 3).$$

Pavyzdys. Tarkime, kad atlikę ranginę logistinę regresiją, gavome tokius parametrų įverčius: $\hat{C}_1 = 1, \hat{C}_2 = 2, \hat{C}_3 = 4,\ \hat{b}_1 = 1,\ \hat{b}_2 = -1, \hat{b}_3 = 2$. Tarkime, kad turime respondentą, kuriam $X = 2$, $Z = 1$, $W = 0$. Tada $\hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W = 1$, $z_1 = 1 - 1 = 0$, $z_2 = 2 - 1 = 1$, $z_3 = 4 - 1 = 3$. Nesunkiai surandame $\mathrm{e}^0 = 1$, $\mathrm{e}^{-1} = 0{,}3678$, $\mathrm{e}^{-3} = 0{,}04978$. Todėl

$$\widehat{\mathrm{P}}(Y = 1) = 0{,}5;\ \widehat{\mathrm{P}}(Y \le 2) = \frac{1}{1{,}3678} = 0{,}731,\ \widehat{\mathrm{P}}(Y \le 3) = \frac{1}{1{,}04978} = 0{,}953$$

ir

$$\widehat{\mathrm{P}}(Y = 1) = 0{,}5;\ \widehat{\mathrm{P}}(Y = 2) = 0{,}731 - 0{,}5 = 0{,}231,$$

$$\widehat{\mathrm{P}}(Y = 3) = 0{,}953 - 0{,}731 = 0{,}222, \qquad \widehat{\mathrm{P}}(Y = 4) = 1 - 0{,}953 = 0{,}047.$$

Kartais reikalinga grubi prognozė, kurį iš galimų Y atsakymų rinksis respondentas? Prognozė daroma taip: suskaičiuojamos visų galimų Y reikšmių tikimybės. Prognozuojame tą reikšmę, kurios įgijimo tikimybė didžiausia. Pavyzdžio atveju prognozuotume, kad Y įgis 1.

6.1.4. Duomenys

Duomenys turi tenkinti tokius reikalavimus:

- Priklausomas kintamasis Y turi būti ranginis. Kai tų rangų yra labai daug, tai maža šansų, kad modelis duomenims tiks. Kai rangų pakankamai daug (bent 7 arba, kaip sako itin liberalūs statistikai, bent 5), o Y vidurkis visiškai prasmingas, tai galima bandyti taikyti tiesinės regresijos modelį.
- Regresoriai – intervaliniai, ranginiai arba kategoriniai kintamieji. Nereikalaujama regresorių normalumo. Kategoriniai regresoriai keičiami dvireikšmiais pseudokintamaisiais.
- Duomenyse negali kažkurių Y reikšmių būti labai mažai. Turime turėti informacijos apie įvairių kategorijų respondentus, kad galėtume nustatyti, kas būdinga kiekvienai kategorijai. Iš tikrųjų reikalavimas yra kiek stipresnis: pakankamai (dažniausiai bent 5) reikšmės turi būti kiekvienam kategorinių regresorių ir priklausomo kintamojo Y reikšmių deriniui. Todėl į modelį įtraukdami daug kategorinių regresorių, galime susidurti su problemomis.
- Turi būti tenkinama *lygiagrečių tiesių* prielaida. Matematiškai šią prielaidą galime nusakyti, kaip reikalavimą, kad modeliuose gali skirtis tik konstantos C_1, C_2, C_3, o daugikliai b_1, b_2, b_3 visuose modeliuose yra tie patys. Ar tenkinama ši prielaida tikrinama specialiu chi kvadrato kriterijumi. Lygiagrečių tiesių prielaida – tai reikalavimas, kad „atstumai“ tarp Y rangų būtų vienodi. Kai yra keli regresoriai, grafiškai lygiagrečių tiesių prielaidą pavaizduoti sudėtinga – ji virsta lygiagrečių hiperplokštumų prielaida.

6.1.5. Modelio tinkamumas

Modelio tikimą duomenims parodo:

a) *Modelio tikimo didžiausio tikėtinumo santykio chi kvadrato statistika* (angl. model fit likelihood ratio Chi-square test) . Vertinamas bendrasis modelio tikimas duomenims. Gerai duomenims tinkančiam modeliui $p < 0,05$. Jeigu $p \geq 0,05$, tai regresijos modelio tinkamumas labai abejotinas.

b) *Pirsono ir deviacijos chi kvadrato statistikos.* Tai alternatyvos didžiausio tikėtinumo kriterijui. Modelis gerai tinka duomenimis, kai Pirsono ir deviacijos chi kvadratų p reikšmės yra didelės ($p \geq 0,05$). Kai modelyje yra intervalinių regresorių, Pirsono ir deviacijos chi kvadratų p reikšmės beveik visada yra nepatikimas modelio gerumo indikatorius. Be to, šios charakteristikos jautrios imties didumui. Didelėms imtims p reikšmės gali būti nepagrįstai mažos. Todėl į šias charakteristikas atsižvelgiama toli gražu ne visada.

c) *Voldo testas*. Voldo testas padeda nuspręsti, kurie kintamieji modelyje reikalingi (yra statistiškai reikšmingi, jų $p < 0,05$). Gerame modelyje visi kintamieji yra statistiškai reikšmingi. Kaip taisyklė, konstantai šio kriterijaus netaikome ir iš modelio nešaliname, net ir statistiškai nereikšmingos konstantos.

d) *Determinacijos (pseudo) koeficientai* (angl. R square). Parodo bendrąjį modelio tikimą. Įgyja reikšmes iš intervalo [0, 1]. Kuo koeficiento reikšmė didesnė, tuo modelis geriau tinka duomenims. Nelabai gerai, kai $R^2 < 0,20$. Ranginėje logistinėje regresijoje determinacijos koeficientai vaidina tik pagalbinį vaidmenį. Jų yra net keli (Kokso-Snelo, Nagelkerkės, Makfadeno). Kurį iš jų naudoti dažniausiai nusprendžia (sufantazuoja) pats tyrėjas. Privalu nurodyti tikslų determinacijos koeficiento pavadinimą.

e) *Tiesių lygiagretumo hipotezės chi kvadrato statistika.* Tiesių lygiagretumo prielaida galioja, kai statistikos p reikšmė $\geq 0,05$. Geruose modeliuose tiesių lygiagretumo prielaida galioja.

f) *Y reikšmių įgijimo tikimybės ir klasifikacinė lentelė.* Konkretiems stebėjimams prognozuojamos Y reikšmės įgijimo tikimybės, pasirenkama ta Y reikšmė, kurios įgijimo tikimybė didžiausia ir tikrinama, ar spėjimas sutapo su tikrąja Y reikšme. Kuo daugiau sutapimų, tuo modelis geresnis. Ranginėje logistinėje regresijoje ši charakteristika nelaikoma labai svarbia. Kai kurie paketai jos net neskaičiuoja. Vis dėlto, jeigu visos kategorijos vienodai svarbios, tai kiekvienos jų teisingų atpažinimų procentas neturi būti mažesnis už tos kategorijos duomenų procentą imtyje.

Apibendrindami galime nusakyti, kokie turi būti gero regresijos modelio rodikliai.

Labai gerai duomenims tinkančiame modelyje:

- Didžiausio tikėtinumo chi kvadrato kriterijaus $p < 0,05$.
- Visų regresorių Voldo kriterijaus $p < 0,05$.
- Tiesių lygiagretumo testo chi kvadrato $p \geq 0,05$.
- Klasifikacinėje lentelėje teisingų kiekvienos kategorijos klasifikavimų procentas ne mažesnis už tos kategorijos duomenų procentą imtyje.
- Nelabai didelėms imtims, turinčioms nedaug (arba neturinčioms) intervalinių regresorių deviacijos chi kvadrato testo ir Pirsono chi kvadrato testo $p \geq 0,05$.

6.1.6. Modelio tyrimas ir tobulinimas

Patikriname visas svarbiausias charakteristikas.

1) Jeigu modelio didžiausio tikėtinumo chi kvadrato statistikos *p* reikšmė $\geq 0,05$, tai **modelis netinkamas.**
2) Patikriname, ar visi regresoriai statistiškai reikšmingi (visos Voldo kriterijaus *p* reikšmės < 0,05). Jeigu ne – modelis taisytinas. Konstantos statistinio reikšmingumo netiriame.
3) Patikriname, ar tenkinama tiesių lygiagretumo prielaida. Jeigu atitinkamo chi kvadrato statistikos *p* reikšmė < 0,05, tai reikia papildomos analizės (STATA dažniausiai tenka ieškoti programos papildo *gologit2*, alternatyva – taikyti nominaliąją regresiją).
4) Papildomai galima ištirti, ar deviacijos ir Pirsonos chi kvadratų *p* reikšmės $\geq$ 0,05 (taip turi būti geruose nedidelių imčių modeliuose, turinčiuose nedaug intervalinių regresorių), surasti galimybes ir jų santykius, determinacijos koeficientus ir klasifikacinę lentelę. Jeigu viskas gerai – modelį aprašome. Jeigu ne – tobuliname.

Modelio tobulinimas reiškia kintamųjų šalinimą arba jų transformaciją. Pašalinus bent vieną kintamąjį, keičiasi visos modelio charakteristikos: determinacijos koeficientai, visos p reikšmės. Modelis tobulinamas, jeigu yra problemų:

1) *Yra statistiškai nereikšmingas kintamasis* (jo didžiausio tikėtinumo chi kvadrato *p* reikšmė didesnė už 0,05). Tada šis kintamasis pašalinamas iš modelio, ir logistinė regresinė analizė pakartojama be jo.
2) *Jeigu nepatenkinta tiesių lygiagretumo prielaida,* tai dar pažiūrime, ar pati statistikos reikšmė yra nedidelė. Galime gauti mažą *p* reikšmę vien todėl, kad turime daug duomenų.

Tada modelio atmesti, kaip netinkamo negalima. Alternatyva yra kitokių jungties funkcijų pasirinkimas.

6.1.7. Pastabos apie ranginės regresijos taikymą

- Gali būti ir vienas regresorius.
- Pseudokintamieji aprašomi taip pat, kaip ir tiesinėje regresijoje.
- Jeigu turime abejonių dėl ranginės regresijos modelio tikimo, visada galima išbandyti daugianarės logistinės regresijos modelį. Atrodytų, kad daugianarės logistinės regresijos modelis yra tikslesnis, ir ranginė logistinė regresija iš viso nebereikalinga. Vis dėlto, yra pavyzdžių, kai duomenims ranginės logistinės regresijos modelis tinka, o daugianarės logistinės regresijos modelis – ne .
- Aprašymuose neretai apsiribojama bendruoju teisingų klasifikavimų procentu. Tai nėra blogai. Tik reikia nepamiršti, kad gerai atpažinti turi *visas Y* kategorijas.
- Jeigu nagrinėjame galimybę (tikimybių santykį) P($Y \leq j$)/P($Y > j$), tai galimybių santykiu vadinamas $e^{-\hat{b}_1}$ (prie daugiklio $\hat{b}_1$ atsiranda minusas).
- Galimos ir kitos jungties funkcijos, ne tik logaritminė. Jas galima taikyti klasifikavimui, bet nebebus teisinga galimybių santykių interpretacija. Be to, terminas *logistinė regresija* tada nenaudotinas, reikia kalbėti apie vieną, ar kitą ranginės regresijos modelį. Paminėsime kelias jungties funkcijas, kurias galima naudoti, kai svarbus tik teisingas klasifikavimas:
  - *Dvigubas logaritmas (log-log)*. Modelis atrodo taip:
    $$\ln\left(-\ln(1 - \mathrm{P}(Y > i)\right) = C_i - b_1 X - b_2 Z - b_3 W.$$
    Šią jungties funkcija rekomenduojama naudoti, kai didesnės *Y* reikšmės labiau tikėtinos, nei mažesnės.
  - *Neigiamas dvigubas logaritmas (negative log-log)*. Modelis atrodo taip
    $$-\ln\left(-\ln(1 - \mathrm{P}(Y > i)\right) = C_i - b_1 X - b_2 Z - b_3 W.$$
    Šią jungties funkcija rekomenduojama naudoti, kai mažesnės *Y* reikšmės labiau tikėtinos, nei didesnės.
  - *Atvirkštinis Koši skirstinys (Cauchit).* Tikimybė $\mathrm{P}(Y \leq i))$ išreiškiama per Koši pasiskirstymo funkciją. Modelio lygties nepateiksime nes ji labai sudėtinga. Užtikriname, kad Koši pasiskirstymo funkcija yra pakankamai bjauri, kad praeitų bet koks noras šį modelį skaičiuoti „rankomis“. Koši jungties funkcija rekomenduojama naudoti tada, kai yra daug išskirtinai didelių ir/ar mažų *Y* reikšmių.

6.1.8. Standartiniai daugianarės logistinės regresijos modelių aprašymai

Visada aprašomas galutinis, po visų tobulinimų gautasis, modelis. Pateiksime keletą galimų aprašymų fragmentų.

- *Taikytas daugianarės ranginės regresijos modelis. Modelio didžiausio tikėtinumo chi kvadrato statistika $\chi^2 = 13,27$; $p < 0,05$. Visi kintamieji buvo statistiškai reikšmingi. Teisingai klasifikuota 70,5% imties respondentų. Nagelgerkės determinacijos koeficientas lygus 0.49. Kintamajam X padidėjus 1, galimybių ($P(Y>i)/P(Y\leq i)$) santykis padidėja 2,3 karto.*
- *Modelyje paliktas statistiškai nereikšmingas kintamasis Z, nes jį pašalinus labai sumažėja teisingai klasifikuotų atvejų procentas....*
- *Kintamajam X padidėjus 1, galimybių ($P(Y\leq i)/P(Y>i)$) santykis sumažėja 2,3 karto….*

6.2. Ranginė logistinė regresinė analizė su SPSS

6.2.1. Duomenys

Kaip atlikti ranginę logistinę regresinę analizę parodysime, naudodami Lietuvos aukštojo mokslo tyrimo duomenis **LAMS**. Tyrime naudosime tokius kintamuosius:

- *K32* – asmeninės vidutinės pajamos prieš mėnesį (1 – mažiau 1000, ..., 9 – daugiau 10000),
- *K36_1* – studijų metu įgytų žinių panaudojimo esamame darbe įvertinimas (1 – visiškai nenaudoju, ..., 5 – labai dažnai naudoju, 99 – nestudijavau/negaliu atsakyti),
- *K2* – aukštoji mokykla,
- *D8* – miesto (gyvenvietės), kurioje dirba didumas, (1 – iki 2000, ..., 7 – virš 500000, 8 – dirbu ne Lietuvoje),
- *K25* – pirmojo darbo atitikimas bakalauro studijų krypčiai (1 – tikrai taip, ..., 4 – tikrai ne, 99 – sunku pasakyti),
- *K11_1* – bakalauro studijų rezultatų įvertinimas (1 – dažniausiai pažymių vidurkis būdavo vienas žemiausių kurse, ..., 5 – dažniausiai pažymių vidurkis būdavo vienas aukščiausių kurse, 99 – nestudijavau),
- *K33_1* – išsilavinimo lygio įtaka, pasiekiant esamą padėtį darbo rinkoje (1 – visai nereikšmingas veiksnys, ..., 5 – labai reikšmingas veiksnys, 99 – nestudijavau) .

Norime nustatyti, kaip bakalauro studijas baigusių respondentų studijų metu gautų žinių naudingumo įvertinimas priklauso nuo respondento pažangumo bei nuomonės apie išsilavinimo

įtaką darbo pasiekimams. Tirsime Vilniaus universiteto, VGTU ir LŽŪA absolventus, dirbančius Lietuvoje.

Kadangi verta tirti tik studijavusius žmones ($K36\_1 \neq 99$), kurie baigė VU, VGTU ir LŽŪA ($K2 = 13, 7, 11$), dirba Lietuvoje ($D8 \leq 7$) ir supranta, ar darbas atitinka studijų pobūdį ($K25 \neq 99$), tai su **Select Cases** atrenkame tinkamus respondentus. Sąlyga:

$$K36\_1 \neq 99 \ \& \ K25 \neq 99 \ \& \ (K2 = 13 | \ K2 = 11 \ | \ K2 = 7) \ \& \ (D8 \leq 7).$$

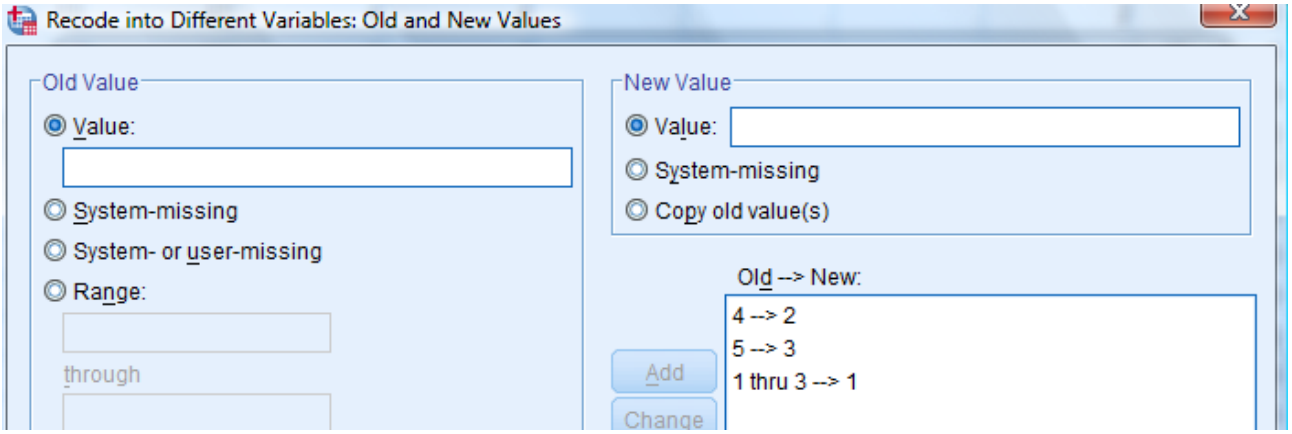
Reikia iš tyrimo pašalinti kintamųjų *K11_1* ir *K33_1* reikšmes 99. Tą padarysime jas įvardindami, kaip trūkstamas reikšmes (angl. missing values). Atidarome **Variable View** langą. Pasirinkę **Missing**, kintamiesiems *K11_1* ir *K33_1* paeiliui nurodome, kad 99 yra praleista reikšmė:

Patikriname, ar turime pakankamai daug skirtingų požiūrių į studijų metu gautas žinias. Pritaikę **Frequences** matome, kad kintamasis *K36_1* įgyja labai nedaug mažų reikšmių, todėl tikslingiau naudoti mažiau kategorijų turintį ranginį kintamąjį.

K36_1|Studijų metu įgytų profesinių žinių panaudojimo esamame darbe įvertinimas

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Visiškai nenaudoju	6	3.6	3.6	3.6
	2	13	7.7	7.7	11.3
	3	34	20.2	20.2	31.5
	4	50	29.8	29.8	61.3
	5 Labai dažnai naudoju	65	38.7	38.7	100.0
	Total	168	100.0	100.0	

Naudodami komandą **Recode into Different Variables,** sukuriame naują kintamąjį *Bakna3*, kuris lygus 1, kai *K36_1* = 1, 2 arba 3 (menkas profesinių žinių naudojimas), *Bakna3* = 2, kai *K36_1* = 4 (vidutinis naudojimas), *Bakna3* = 3, kai *K36_1* = 5 (labai dažnas naudojimas).



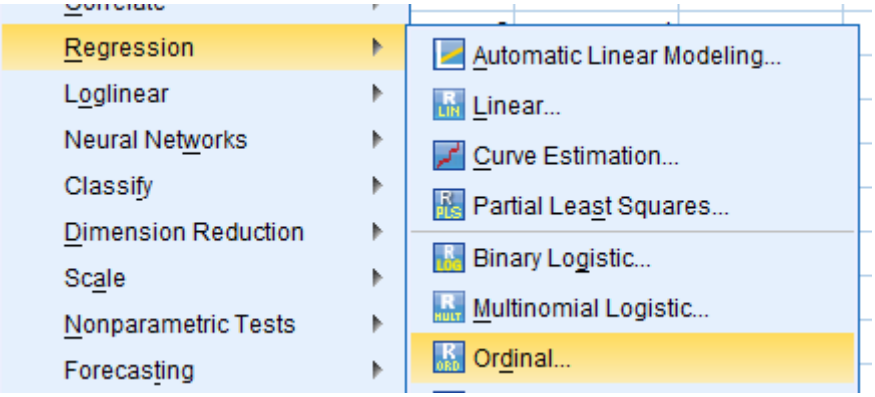
Analogiškai įsitikiname, kad *K32* įgyja labai mažą kiekį didelių reikšmių (mažai absolventų gauna labai didelius atlyginimus). Šį kintamąjį irgi keičiame kintamuoju, turinčiu mažiau kategorijų. Naudodami komandą **Recode into Different Variables,** sukuriame naują kintamąjį *algl2*, kuris visus respondentus padalija į uždirbančius iki 1500 LTL per mėn. (*algl2* = 1) ir uždirbančius daugiau, nei 1500 LTL per mėn. (*algl2* = 2). Beje, nežinančius kiek uždirba (*K32* = 99) į tyrimą neįtraukėme ir perkoduodami, šio kodo nenaudojome.

6.2.2. SPSS parinktys ranginei logistinei regresijai

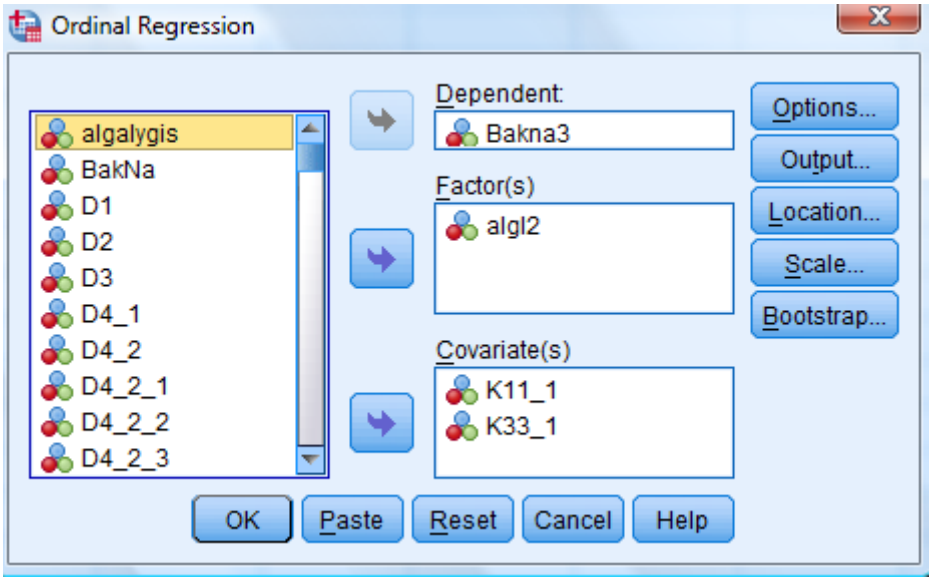
Tirsime modelį:

Bakna3 = f (algl2, K11_1, K33_1).

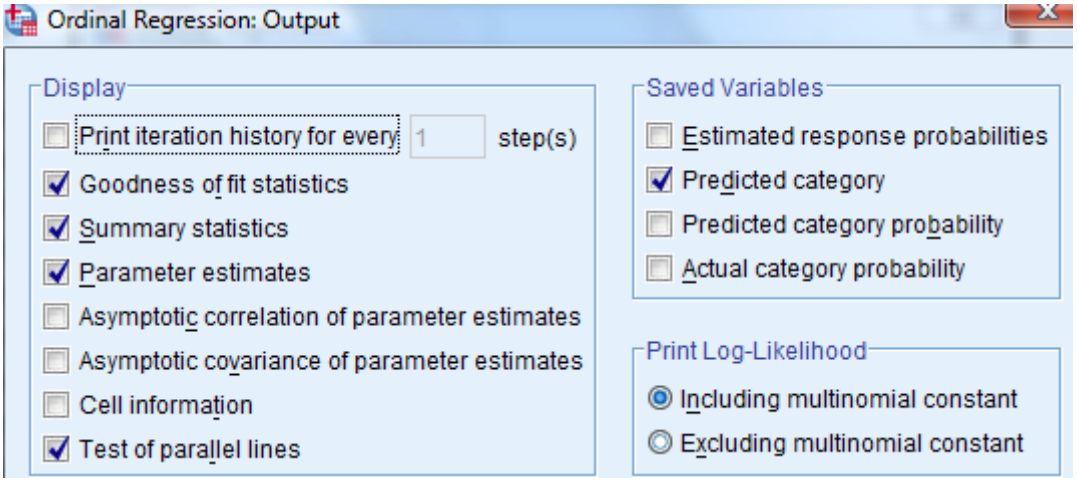
Atsidarome langelį **Analyze** ir renkamės **Regression → Ordinal.**



Atsidariusiame meniu perkeliame kintamąjį *Bakna3* į laukelį **Dependent**. Kintamuosius *K11*_1, *K33_1* įkeliame į laukelį **Covariates**. Kintamasis *algl2* yra kategorinis, todėl jį keliame į **Factor(s)**.



Pasirenkame **Output** ir papildomai pažymime **Test of parallel lines** ir **Predicted category** (tam, kad galėtume sudaryti klasifikacinę lentelę). Pasirekame **Continue** ir **OK.**



6.2.3. Rezultatai

Rezultatų išklotinė prasideda nuo įspėjimo, kad 28,6 % langelių tušti.

Warnings

There are 18 (28.6%) cells (i.e., dependent variable levels by combinations of predictor variable values) with zero frequencies.

Apie kokius langelius čia kalbama? Tai tikrinimas, ar bent kelios *Bakna3* reikšmės yra įgyjamos kiekvienai regresorių reikšmių kombinacijai (pvz., ar yra bent viena *Bakna3* reikšmė, kai *K11_1* = 1, *K33_1* = 1, o *algl2* = 1). Šis rodiklis gali būti informatyvus, kai regresoriai yra kategoriniai. Tada didelis tuščių langelių procentas rodo, kad vertėtų dalį kategorinių kintamųjų iš modelio arba pašalinti, arba sustambinti jų kategorijas. Kai modelyje yra intervalinių kintamųjų, šis įspėjimas ignoruojamas. Nors regresoriai *K11_1*, *K33_1* tik labai „pritemptai“ laikytini intervaliniais, šį įspėjimą mes irgi ignoruosime. Didelis tuščių langelių procentas, taip pat rodo, kad nelabai galima pasitikėti Pirsono ir deviacijos chi kvadrato statistikos *p* reikšmėmis.

Lentelėje **Case Processing Summary** yra informacija apie tai, kokias reikšmes įgyja modelio kategoriniai kintamieji. Visų pirma, įsitikiname, kad pakankamai daug (žr. atitinkamas rekomendacijas duagianarei logistinei regresijai) stebėjimų atitinka kiekvieną kintamųjų *Bakna3* ir regresorius *algl2* kategoriją. Modelis nebūtų labai logiškas, jeigu būtų informacija tik apie kelis respondentus, uždirbančius daugiau nei 1500 LTL, beveik nebūtų dažnai studijų žinias darbe taikiusių respondentų ir pan.

Case Processing Summary

		N	Marginal Percentage
Bakna3 Studijų panaudojimas darbe	1.00 menkas naudojimas	51	30.7%
	2.00 vidutinis naudojimas	50	30.1%
	3.00 dažnas naudojimas	65	39.2%
algl2 alga per men.	1.00 iki 1500	52	31.3%
	2.00 virš 1500	114	68.7%
Valid		166	100.0%
Missing		2	
Total		168	

Priklausomo kintamojo *Bakna3* įgyjamų reikšmių procentai naudotini ir tiriant klasifikavimo tikslumą. Pavyzdžiui, visą laiką spėdami, kad studijų panaudojimas darbe yra vidutinis (*Bakna3* = 2), teisingai klasifikuotume 30,1% atsakymų. Logiška reikalauti, kad pasiūlytas ranginės logistinės regresijos modelis užtikrintų ne blogesnį klasifikavimo procentą.

Lentelėje **Model Fitting Information** yra didžiausio tikėtinumo chi kvadrato kriterijaus statistikos reikšmė ir atitinkama *p* reikšmė. Kadangi $p = 0{,}000... < 0{,}05$, tai gavome vieną iš svarbiausių patvirtinimų, jog modelis duomenims tinka.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	147.472			
Final	82.100	65.372	3	.000

Link function: Logit.

Lentelėje **Goodness-of-Fit** yra Pirsono ir deviacijos chi kvadrato statistikų reikšmės ir *p* reikšmės. Kadangi abidvi $p > 0{,}05$, tai gavome dar vieną patvirtinimą jog modelis duomenims tinka. Primename, kad šios *p* reikšmės svarbios tik tada, kai imtys nedidelės. Taip ir yra tiramu atveju, nes imtį sudaro tik 166 respondentai (žr. lenteles aukščiau). Determinacijos koeficientai (lentelė **Pseudo R-Square)** nėra itin dideli. Vis dėlto du iš jų yra didesni už 0,20, o ir trečiasis nedaug mažesnis. Taigi ir šie rodikliai nerodo labai blogo modelio tikimo.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	25.434	37	.925
Deviance	26.579	37	.898

Link function: Logit.

Pseudo R-Square

Cox and Snell	.326
Nagelkerke	.367
McFadden	.180

Link function: Logit.

Parametrų įverčiai ir jų 95% pasikliautiniai intervalai yra lentelėje **Parameter Estimates**. Prisimename, kad teigiamas kintamojo koeficientas rodo, jog didėjant kintamojo reikšmėms, labiau tikėtinomis taps ir didesnės *Bakna3* reikšmės. Teigiami kintamųjų *K11_1* ir *K33_1* koeficientai (1,182 ir 0,979) visiškai atitinka modelio logiką: kuo geriau respondentas(ė) mokėsi bakalauro studijų metu ir kuo svarbesne laikoma išsilavinimo įtaka karjerai, tuo labiau tikėtina, kad jis(ji) palankiau įvertins studijų medžiagos naudojimą darbe. Analogiškai interpretuojamas ir kategorinio kintamojo *algl2* koeficientas. SPSS pakete ši informacija pateikiama ne itin patogiu pavidalu. Manoma, kad *algl2* = 2 yra pagrindinė kategorija (jos koeficiento nėra), ir stebima, kaip viskas keisis, jeigu *algl2* = 1. Matome, kad prie *algl2* = 1 koeficientas yra neigiamas (-1,277). Prisimename, kad neigiamas koeficientas rodo, jog didėjant kintamojo reikšmėms, labiau tikėtinomis taps ir mažesnės *Bakna3* reikšmės. Taigi, informacija, kad respondentas(ė) uždirba iki 1500 LTL, padaro labiau tikėtiną tikimybę, kad jis(ji) mažiau naudojasi studijų metu sukauptomis žiniomis.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[Bakna3 = 1.00]	7.567	1.426	28.177	1	.000	4.773	10.362
	[Bakna3 = 2.00]	9.286	1.490	38.850	1	.000	6.366	12.206
Location	K11_1	1.182	.264	20.098	1	.000	.665	1.698
	K33_1	.979	.205	22.866	1	.000	.578	1.380
	[algl2=1.00]	-1.277	.360	12.565	1	.000	-1.984	-.571
	[algl2=2.00]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Tikimybių skaičiavimui naudotina tokia formulė (i = 1, 2):

$$\ln\frac{\mathrm{P}(Bakna3 > i)}{\mathrm{P}(Bakna3 \leq i)} = 1{,}182K11_1 + 0{,}979K33\_1$$
$$+\begin{cases} 7{,}567, & \text{jei } i = 1 \\ 9{,}286, & \text{jei } i = 2 \end{cases} - \begin{cases} 1{,}277, & \text{jei } algl2 = 1, \\ 0, & \text{jei } algl2 = 2. \end{cases}.$$

Beliktų į šią formulę įstatyti konkretaus respondento duomenis.

Lentelėje **Test of Parallel Lines** tikrinama hipotezė, kad regresorių koeficientai yra tie patys visiems tikimybių santykiams (taip, kaip aukščiau pateiktoje formulėje – kintant i, keičiasi tik poslinkio konstanta, o daugikliai prie *K11_1* ir *K33_1* nesikeičia). Kadangi p = 0,566 > 0,05, tai ši hipotezė neatmetama.

Test of Parallel Lines[a]

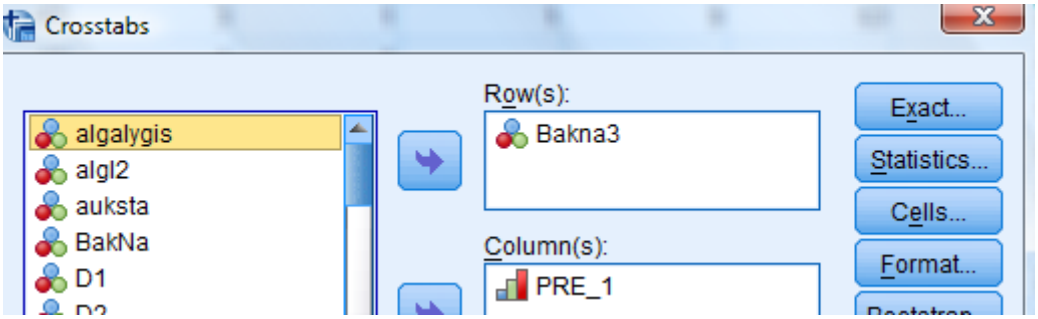
Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	82.100			
General	80.070	2.030	3	.566

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

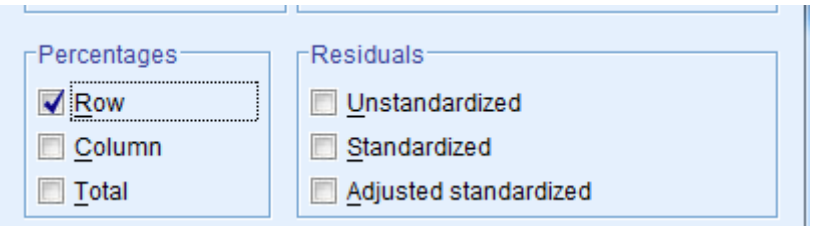
Taigi visos pagrindinės modelio charakteristikos rodo gerą modelio tikimą duomenims. Vis dėlto, rekomenduotina patikrinti, ar modelis tinkamai klasifikuoja respondentus.

6.2.4. Klasifikacinė lentelė

Klasifikacinę lentelę lengviausia gauti, pasinaudojus porinių (kryžminių) dažnių lentele. Kiekvienam respondentui esame išsaugoję prognozuojamą kintamojo *Bakna3* kategoriją (tai įvyko tada, kai SPSS parinktyse pažymėjome **Predicted category**. Duomenyse atsirado naujas stulpelis *PRE_1*, kuriame ir pažymėta toji kategorija). Pasirenkame **Analyze → Descriptive Statistics → Crosstabs**. Kintamąjį *Bakna3* įkeliame į **Row(s)**, o *PRE_1* į **Column(s)**. Pasirenkame **Cells**.



Atsidariusiame lange papildomai pažymime **Row**. Spaudžiame **Continue** ir **OK**.



Gautojoje klasifikacinėje lentelėje matyti, kad teisingai klasifikuota 49% atvejų, kai *Bakna3* = 1, 36% atvejų, kai *Bakna3* = 2 ir 72% atvejų, kai *Bakna3* = 3. Ar patenkinamas toks klasifikavimo tikslumas? Visų pirma, įsitikiname, kad kiekvienai *Bakna3* kategorijai teisingai klasifikuotų atvejų procentas yra ne mažesnis, nei klaidingų klasifikavimų procentas. Be to, įsitikiname, kad taikant modelį, visada klasifikavimo procentas didesnis, ne gautume aklai spėdami

(žr. lentelę **Case Processing Summary**). Pavyzdžiui, visą laiką spėdami, kad *Bakna3* = 2, teisingai klasifikuotume 30,1% atvejų, o taikydami modelį – 36% atvejų. Nors teisingų klasifikavimų procentai nėra labai dideli, jie yra pakankami, kad manytume modelį duomenims esant tinkamą.

Bakna3 Studiju panaudojimas darbe * PRE_1 Predicted Response Category Crosstabulation

			PRE_1 Predicted Response Category			
			1.00 menkas naudojimas	2.00 vidutinis naudojimas	3.00 dažnas naudojimas	Total
Bakna3 Studiju panaudojimas darbe	1.00 menkas naudojimas	Count	25	17	9	51
		% within Bakna3	49.0%	33.3%	17.6%	100.0%
	2.00 vidutinis naudojimas	Count	14	18	18	50
		% within Bakna3	28.0%	36.0%	36.0%	100.0%
	3.00 dažnas naudojimas	Count	2	16	47	65
		% within Bakna3	3.1%	24.6%	72.3%	100.0%
Total		Count	41	51	74	166
		% within Bakna3	24.7%	30.7%	44.6%	100.0%

6.2.5. Rezultatų aprašymas

Pradžioje aprašome, ką tyrėme, kokius kintamuosius naudojome ir kaip juos perkodavome. Tada kompaktiškai surašome visa tai, ką išsamiai aptarėme ankstesniame skyrelyje. Pateikiame vieną iš daugybės galimų tokio aprašymo variantų.

Duomenų analizei taikėme ranginę logistinę regresiją. Tiesių lygiagretumo prielaida buvo patenkinta (χ^2 = 2,03; df = 3; p = 0,566). Pakankamai gerą pasirinkto modelio tikimą duomenims rodo didžiausio tikėtinumo chi kvadrato kriterijus (χ^2 = 65,372; df = 3; p = 0,00) bei Pirsono (χ^2 = 25,434; df = 37; p = 0,925) ir deviacijos (χ^2 = 26,579; df = 37; p = 0,898) chi kvadrato kriterijai. Nagelkerkės determinacijos koeficientas yra lygus 0,367. Taikant modelį teisingai klasifikuojama 49% atvejų, kai *Bakna3* = 1, 36% atvejų, kai *Bakna3* = 2 ir 72% atvejų, kai *Bakna3* = 3. Modelis su įvertintais parametrais atrodo taip:

$$\ln\frac{\mathrm{P}(Bakna3 \le i)}{\mathrm{P}(Bakna3 > i)} = \begin{cases} 7{,}567, & \text{jei } i = 1 \\ 9{,}286, & \text{jei } i = 2 \end{cases} - 1{,}182K11\_1 - 0{,}979K33\_1$$

$$+\begin{cases} 1{,}277, & \text{jei } algl2 = 1, \\ 0, & \text{jei } algl2 = 2. \end{cases}$$

Galima aprašyti ir galimybių santykius.

6.2.6. SPSS komandų sintaksė daugianarei logistinei regresinei analizei

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam ranginės logistinės regresijos modeliui. Jas galima naudoti sintaksės lange, pasirenkant opciją **Run**.

Kintamųjų atranka:

```
COMPUTE filter_$=(K36_1 ~= 99 & K25 ~= 99 & (K2 = 13 | K2 = 11 | K2 =7) & D8 <= 7).
VARIABLE LABELS filter_$ 'K36_1 ~= 99 & K25 ~= 99 & (K2 = 13 | K2 = 11 | K2 =7) & D8 <= 7
(FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

Kintamųjų *Bakna3* ir *algl2* kūrimas:

```
RECODE K36_1 (4=2) (5=3) (1 thru 3=1) INTO Bakna3.
EXECUTE.
RECODE K32 (1 thru 2=1) (3 thru 9=2) INTO algl2.
EXECUTE.
```

Ranginės logistinės regresijos modelis:

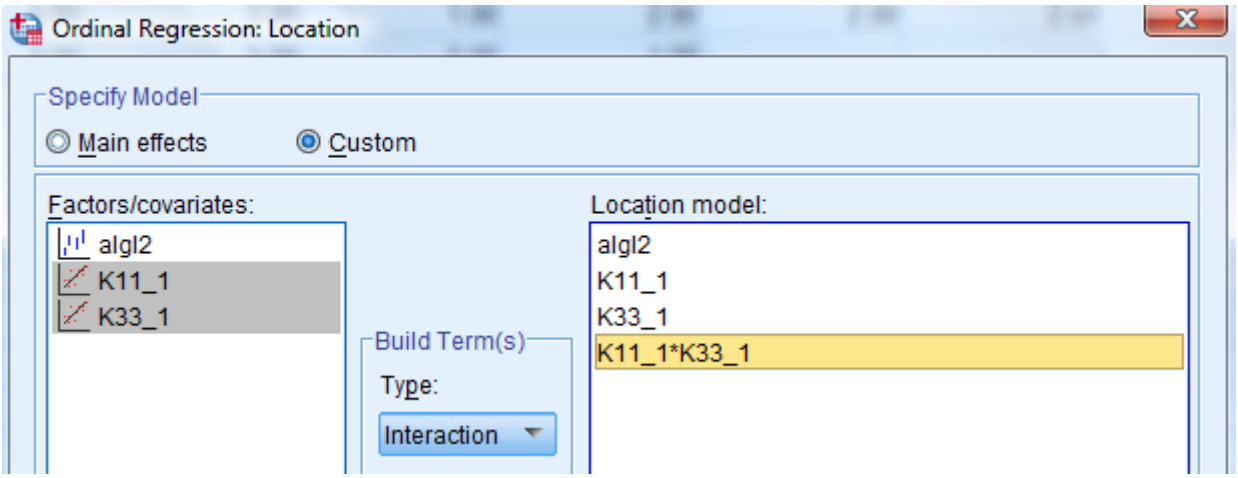
```
PLUM Bakna3 BY algl2 WITH K11_1 K33_1
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100)
MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /LINK=LOGIT
  /PRINT=FIT PARAMETER SUMMARY TPARALLEL
  /SAVE=PREDCAT.
```

Klasifikacinė lentelė:

```
CROSSTABS
  /TABLES=Bakna3 BY PRE_1  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW   /COUNT ROUND CELL.
```

6.2.7. Ranginė logistinė regresija su kintamųjų sąveika

Į modelį galima įtraukti ne tik pačius kintamuosius, bet ir jų sandaugas. Tada sakoma, kad modelyje yra kintamųjų tarpusavio sąveika (interakcija). Parodysime, kaip tą padaryti, naudojantis SPSS programa. Visos parinktys išlieka tokios pat, kaip aprašyta 5.2.2 skyrelyje. Papildomai pagrindiniame lange **Ordinal Regression** paspaudžiame **Location**. Atsidariusiame lange pažymime **Custom**. Iš pateiktų **Type** parinkčių pasirenkame **Interaction** ir į laukelį **Location model** perkeliame visus regresorius po vieną ir regresorių *K11_1* ir *K33_1* kartu.



Faktiškai tiriame modelį *Bakna3 = f (algl2, K11_1, K33_1, K11_1 * K33_1)*, t.y. visose formulėse prisideda kintamųjų *K11_1 * K33_1* sandauga. Patikriname, ar naujasis modelis taip pat gerai tinka duomenims, kaip ir modelis be interakcijos. Nepateiksime visų rezultatų, tik pažymėsime, kad visos modelio tikimo duomenims charakteristikos pagerėja. Išimtį sudaro Voldo kriterijaus reikšmės. Faktiškai visi kintamijei tapo statistiškai nereikšmingais.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[Bakna3 = 1.00]	.561	4.670	.014	1	.904	-8.592	9.714
	[Bakna3 = 2.00]	2.305	4.672	.243	1	.622	-6.852	11.462
Location	[algl2=1.00]	-1.287	.363	12.583	1	.000	-1.999	-.576
	[algl2=2.00]	0[a]	.	.	0	.	.	.
	K11_1	-.587	1.178	.248	1	.618	-2.896	1.722
	K33_1	-.595	1.045	.325	1	.569	-2.644	1.453
	K11_1 * K33_1	.399	.265	2.272	1	.132	-.120	.917

Link function: Logit.

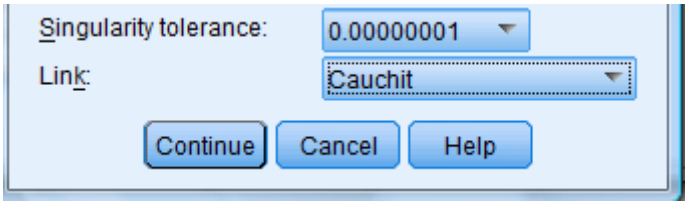
a. This parameter is set to zero because it is redundant.

Taigi modelį papildžius nereikalinga kintamųjų interakcija, ne tik ji tapo statistiškai nereikšminga, bet ir visas modelis tapo visiškai netikęs.

6.2.8. Ranginė regresija su kitomis jungties funkcijomis

Parodysime, kaip ranginės regresijos modelyje pasirinkti kitą jungties funkciją. Pasirinksime ranginės regresijos Koši modelį. Atsidarome langelį **Analyze** ir renkamės **Regression → Ordinal.**

Spaudžiame **Options**. Pasirenkame **Link – Cauchit**.



Paspaudžiame **Continue** ir toliau viską darome taip, kaip aprašyta 5.2.2 ir 5.2.4 skyreliuose. Nepateiksime visų gautų rezultatų, tik pažymėsime, kad praktiškai visos modelio tikimo charakteristikos liko nepakitusios, o klasifikacinė lentelė net pagerėjo: teisingai klasifikuota 47,1% atvejų, kai *Bakna3* = 1, 44% atvejų, kai *Bakna3* = 2 ir 72,3% atvejų, kai *Bakna3* = 3. Tiesa, šiam modeliui nebegalima kalbėti apie galimybių santykius. Kažkokių patikimų rekomendacijų, kada kokią jungties funkciją naudoti – nėra. Matyt, vienintelė protinga rekomendacija – jeigu neišėjo su logaritmine jungtimi gauti gero modelio, tai galima išbandyti kitas jungties funkcijas.

6.3. Ranginė logistinė regresinė analizė su STATA

6.3.1. Duomenys

Pasinaudosime su SPSS jau atliktomis **LAMS** failo duomenų transformacijomis. Tirsime Vilniaus universiteto, VGTU ir LŽŪA absolventus, dirbančius Lietuvoje. Kintamieji

- *K11_1* – bakalauro studijų rezultatų įvertinimas (1 – dažniausiai pažymių vidurkis būdavo vienas žemiausių kurse, ..., 5 – dažniausiai pažymių vidurkis būdavo vienas aukščiausių kurse, 99 – nestudijavau),
- *K33_1* – išsilavinimo lygio įtaka, pasiekiant esamą padėtį darbo rinkoje (1 – visai nereikšmingas veiksnys, ..., 5 – labai reikšmingas veiksnys, 99 – nestudijavau) ,
- *Bakna3* – profesinių žinių naudojimas (1 – menkas, 2 – vidutinis, 3 – labai dažnas),
- *algl2* – algos lygmuo (1 – iki 1500 LTL per mėn., 2 – daugiau, nei 1500 LTL per mėn.).

Reikia iš tyrimo pašalinti kintamųjų *K11_1* ir *K33_1* reikšmes 99. Tą padarome, įvykdę:

```
keep if (K11_1 < 99 & K33_1 < 99)
```

Kiek pritemptai kintamuosius *K11_1* ir *K33_1* galima tarti, esant intervaliniais. Kintamasis *algl2* yra dvireikšmis. Modelyje leistini tik 0-1 įgyjantys dvireikšmiai kintamieji. Todėl *algl2* perkoduojame:

```
gen zalg2 = algl2 - 1
```

6.3.2. Modelio analizė

Tirsime modelį: *Bakna3 = f (algl2, K11_1, K33_1).* Įvykdome

```
ologit Bakna3 zalg2 K11_1 K33_1
```

```
Ordered logistic regression                     Number of obs   =        166
                                                LR chi2(3)      =      65.37
                                                Prob > chi2     =     0.0000
Log likelihood = -148.44438                     Pseudo R2       =     0.1805

------------------------------------------------------------------------------
      Bakna3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       zalg2 |   1.277259   .3600362     3.55   0.000     .5716016    1.982917
       K11_1 |    1.18157   .2665309     4.43   0.000     .6591788    1.703961
       K33_1 |   .9790868   .2020289     4.85   0.000     .5831174    1.375056
-------------+----------------------------------------------------------------
       /cut1 |    8.84465   1.497767                      5.909081    11.78022
       /cut2 |   10.56329   1.569818                      7.486499    13.64007
------------------------------------------------------------------------------
```

Modelio didžiausio tikėtinumo chi kvadrato reikšmė 65,37; p = 0,000. Viena svarbiausių charakteristikų rodo gerą modelio tinkamumą. Pateiktasis (Makfadeno) determinacijos pseudokoeficientas nėra itin didelis R^2 = 0,181, bet jo vaidmuo nėra itin svarbus. Voldo testai rodo, kad visi regresoriai yra statistiškai reikšmingi.

Kitaip, nei SPSS, STATA neužrašo modelio laisvųjų konstanų įverčių, o tik pateikia regresorių daugiklių įverčius. Alternatyva konstantoms yra vadinamieji *pjūvio taškai* (angl. cut points). Tariama, kad tikrasis *Bakna3* yra tolydus kintamasis ir jo reikšmės nuo 0 iki 8,8 žymi menką žinių naudojimą, nuo 8,8 iki 10,5 – vidutinį, o virš 10,5 – didelį naudojimą.Naudos iš tokios interpretacijos nėra daug. Žinoma, pjūvio taškai išsireiškia ir per modelio konstantas. Pavyzdžiui, kai priklausomas kintamasis įgyja tik dvi reikšmes, tai yra vienas pjūvio taškas, kurio reikšmė yra priešinga logit modelio laisvajai konstantai. Kai kategorijų daugiau, nei dvi – išraiškos bjaurėja. Čia

jų nepateikiame, nes nemanome, kad tai turi prasmę. Ranginėje logistinėje regresijoje svarbiausia yra regresorių koeficientai. Modelis su įvertintais koeficientais[15] atrodo taip:

$$\frac{\text{P(dažniau naudoja)}}{\text{P(rečiau naudoja)}} = \exp\{Const + 1{,}28\, zalg2 + 1{,}18K11\_1 + 0{,}98K33\_1\}.$$

Visų kintamųjų koeficientai teigiami. Todėl, didėjant *K11_1* (atsakinėjant geriau besimokiusiam respondentui), tikimybė, kad respondentas dažniau naudoja profesines žinias, išauga. Išauga ši tikimybė ir didėjant *K33_1* (palankiau vertinant išsilavinimo lygio įtaką karjerai). Tikimybę didina ir informacija, kad respondentas uždirba daugiau nei 1500 kitų (*zalg2* = 1, t.y. *algl2* = 2) .

Norėdami gauti informaciją apie galimybių santykius įvykdome

```
ologit  Bakna3 zalg2 K11_1 K33_1, or
```

Bakna3	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
zalg2	3.586797	1.291377	3.55	0.000	1.771101	7.263904
K11_1	3.259487	.8687538	4.43	0.000	1.933204	5.495671
K33_1	2.662024	.5378057	4.85	0.000	1.791615	3.955299
/cut1	8.84465	1.497767			5.909081	11.78022
/cut2	10.56329	1.569818			7.486499	13.64007

Galimybių santykis parodo, kaip pasikeičia tikimybių santykis, kai vienetu didėja tik vieno regresoriaus reikšmė, kitiems regresoriams nesikeičiant. Pavyzdžiui, galimybė, kad respondentas dažniau naudoja profesines žinias išauga 3,59 karto, kai žinome, jog jis uždirba virš 1500 litų (*zalg2* = 1):

$$\left(\frac{\text{P(dažniau naudoja)}}{\text{P(rečiau naudoja)}}\right)_{\text{uždirba >1500}} = 3{,}59\left(\frac{\text{P(dažniau naudoja)}}{\text{P(rečiau naudoja)}}\right)_{\text{uždirba iki 1500}}$$

Analogiškai interpretuojami ir likusių regresorių galimybių santykiai. Alternatyviai galima pasinaudoti SPost programa listcoef.

```
listcoef, help
```

[15] Modelį galima užrašyti ir atvirkščiam tikimybių santykiui. Tada daugikliai prie visų regresorių užrašomi, pakeičiant jų ženklus.

```
ologit (N=166): Factor Change in Odds

  Odds of: >m vs <=m
```

Bakna3	b	z	P>|z|	e^b	e^bStdX	SDofX
zalg2	1.27726	3.548	0.000	3.5868	1.8116	0.4652
K11_1	1.18157	4.433	0.000	3.2595	2.2409	0.6829
K33_1	0.97909	4.846	0.000	2.6620	2.4868	0.9304

```
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

Šioje lentelėje galimybių santykiai yra stulpelyje **e^b** . Klasikinis (standartinis) galimybių santykis atspindi pokyčius, kai regresorius padidėja vienetu. Ne visada, toks požiūris yra logiškas – būna regresorių, kuriems vieneto pokytis – labai mažas, būna regresorių, kuriems vieneto pokytis – labai didelis. Todėl lentelėje dar yra informacija, kaip keisis galimybės, jeigu regresoriaus reikšmė pasikeis per jo standartinį nuokrypį. Toks savotiškas panormavimas ir matavimo skalės nukenksminimas.

Tikrindami tiesių lygiagretumo prielaidą, įvykdome

```
brant
```

```
Brant Test of Parallel Regression Assumption
```

Variable	chi2	p>chi2	df
All	1.98	0.577	3
zalg2	0.02	0.897	1
K11_1	0.13	0.716	1
K33_1	1.96	0.162	1

```
A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

Lentelėje pateikiama informacija, ar lygiagrečių tiesių prielaida tenkinama, kai atsižvelgiama į visus regresorius (**All**) ir kai atsižvelgiama į kiekvieną regresorių atskirai. Visos *p* reikšmės surašytos stulpelyje **p>chi2**. Jeigu regresoriui $p < 0{,}05$, tai gali būti, kad jis modelyje nereikalingas (reikėtų patikrinti modelį be to regresoriaus). Matome, kad regresorių, kuriems $p < 0{,}05$, nagrinėjamame modelyje nėra. Neprieštaraujama ir bendrajai tiesių lygiagretumo prielaidai ($p = 0{,}557 > 0{,}05$). Darome išvadą, kad modelis yra tinkamas.

Beje, norėdami pamatyti alternatyvius determinacijos koeficientus, Akaikės informacinio kriterijaus reikšmę ir kitus rodiklius, tiesiog įvykdome

```
fitstat
```

```
Measures of Fit for ologit of Bakna3

Log-Lik Intercept Only:       -181.131   Log-Lik Full Model:           -148.444
D(161):                        296.889   LR(3):                          65.372
                                         Prob > LR:                       0.000
McFadden's R2:                   0.180   McFadden's Adj R2:               0.153
ML (Cox-Snell) R2:               0.326   Cragg-Uhler(Nagelkerke) R2:      0.367
McKelvey & Zavoina's R2:         0.388
Variance of y*:                  5.380   Variance of error:               3.290
Count R2:                        0.542   Adj Count R2:                    0.248
AIC:                             1.849   AIC*n:                         306.889
BIC:                          -526.141   BIC':                          -50.036
BIC used by Stata:             322.449   AIC used by Stata:             306.889
```

Matome, kad Nagelkerkės R^2 = 0,367 yra gerokai didesnis už Makfadeno R^2.

6.3.3. Prognozavimas

Prognozavimui patogiausia naudoti SPost priedo komandą *prvalue*. Tarkime, kad norime padaryti *Bakna3* prognozę respondentui, kurio *zalg2* = 1, *K11_1* = 2, *K33_1* = 5. Įvykdome

```
prvalue, x( zalg2=1  K11_1=2  K33_1=5)
```

```
ologit: Predictions for Bakna3

Confidence intervals by delta method

                                 95% Conf. Interval
     Pr(y=menkas_n|x): 0.5766  [ 0.3213,    0.8319]
     Pr(y=vidutini|x): 0.3071  [ 0.1488,    0.4653]
     Pr(y=dažnas_n|x): 0.1164  [ 0.0038,    0.2289]

    zalg2  K11_1  K33_1
x=      1      2      5
```

Gavome ne tik galimų *Bakna3* reikšmių tikimybes (pavyzdžiui, P(*Bakna3* = 3) = 0,116), bet ir tų tikimybių pasikliautinius intervalus.

Tarkime norime padaryti prognozę, kai *zalg2* = 1, *K11_1* = 2, o *K33_1* lygus savo galimų reikšmių vidurkiui. Įvykdome

```
prvalue, x( zalg2=1  K11_1=2) rest(mean)
```

```
ologit: Predictions for Bakna3

Confidence intervals by delta method

                                 95% Conf. Interval
     Pr(y=menkas_n|x): 0.7155  [ 0.5052,    0.9258]
     Pr(y=vidutini|x): 0.2180  [ 0.0685,    0.3675]
     Pr(y=dažnas_n|x): 0.0666  [-0.0022,    0.1353]

       zalg2     K11_1     K33_1
x=         1         2  4.373494
```

Be pačių prognozuojamų tikimybių, dar gavome informaciją, kad *K33_1* vidurkis yra 4,37. Vidutinės reikšmės imamos pagal nutylėjimą, todėl tą pačią rezultatų lentelę gautume įvykdę prvalue, x(zalg2=1 K11_1=2). Galima gauti ir išsamesnę informaciją, kaip keičiasi kiekvienos *Bakna3* reikšmės tikimybė, keičiantis regresoriui. Imdami vidutines *K11_1* ir *K33_1* reikšmes ištirsime šią priklausomybę regresoriui *zalg2*. Primename, kad *zalg2* įgyja reikšmę 0, jeigu respondentas uždirba iki 1500 litų per mėnesį. Regresorius *zalg2* = 1, jeigu respondentas uždirba daugiau nei 1500 litų per mėnesį. Įvykdome

```
prtab zalg2
```

```
ologit: Predicted probabilities for Bakna3

Predicted probability of outcome 1 (menkas_naudojimas)

------------------------
    zalg2 | Prediction
-----------+------------
        0 |     0.4539
        1 |     0.1881
------------------------

Predicted probability of outcome 2 (vidutinis_naudojimas)

------------------------
    zalg2 | Prediction
-----------+------------
        0 |     0.3687
        1 |     0.3756
------------------------

Predicted probability of outcome 3 (dažnas_naudojimas)

------------------------
    zalg2 | Prediction
-----------+------------
        0 |     0.1775
        1 |     0.4363
------------------------

        zalg2      K11_1      K33_1
x=  .68674699  4.0180723   4.373494
```

Matome, kad menko profesinių žinių naudojimo tikimybė stipriai sumažėja, kai respondentas uždirba virš 1500 litų. Vidutinio žinių naudojimo tikimybė praktiškai nepasikeičia, o dažno naudojimo tikimybė stipriai išauga.

6.3.4. Klasifikacinė lentelė

Nors klasifikacinė lentelė beveik niekada ranginės logistinės regresijos aprašuose nepridedama, parodysime, kaip ją gauti. Klasifikacinę lentelę tenka susikurti, analogiškai daugianarės regresijos atvejui, žr. 5.3.4 skyrelį. Paeiliui įvykdome komandas:

```
predict pre1 pre2 pre3
gen cat = 0
replace cat = 1 if pre1>pre2 & pre1 > pre3
replace cat = 2 if pre2>pre1 & pre2 > pre3
replace cat = 3 if pre3>pre1 & pre3 > pre2
label define nopsas 1 "Menkai" 2 "Vidutiniškai" 3 "Dažnai"
label values cat nopsas
tab  Bakna3 cat, row
```

```
   frequency
row percentage
```

Studiju panaudojimas darbe	cat Menkai	Vidutinið	Daþnai	Total
menkas naudojimas	25 49.02	17 33.33	9 17.65	51 100.00
vidutinis naudojimas	14 28.00	18 36.00	18 36.00	50 100.00
dažnas naudojimas	2 3.08	16 24.62	47 72.31	65 100.00
Total	41 24.70	51 30.72	74 44.58	166 100.00

Matome, kad iš 51 menkai profesines žinias naudojusio respondento teisingai klasifikuoti 25 (49,02%) . Teisingai buvo klasifikuoti 36% vidutiniai žinių naudotojai ir 72,31% – dažni naudotojai. Išskyrus trečią kategoriją, kiti procentai gana maži. Palyginkime su pradine nuomonių dažnių lentele:

```
tab Bakna3, freq
```

```
. tab  Bakna3

Studiju panaudojimas |
               darbe |      Freq.     Percent        Cum.
----------------------+-----------------------------------
   menkas naudojimas |         51       30.72       30.72
vidutinis naudojimas |         50       30.12       60.84
   dažnas naudojimas |         65       39.16      100.00
```

Klasifikacinės lentelės procentai didesni, nei gautume aklai spėliodami. Taigi, klasifikavimas nėra jau toks blogas.

6.3.5. Kategoriniai regresoriai

Tirtajame modelyje naudojome dvireikšmį kategorinį regresorių. Jeigu regresorius įgyja daugiau, nei dvi reikšmes, tai jį reikia pakeisti keletu dvireikšmių pseudokintamųjų. Su STATA programa tą galima padaryti automatiškai. Reikia prieš *ologit* parašyti *xi:,* o prieš kategorinius kintamuosius *i.* Tiesa kategorinio regresoriaus viena iš įgyjamų reikšmių turi būti lygi 0. Automatinis perkodavimas vyksta taip: kiekviena kita regresoriaus reikšmė keičiama nauju dvireikšmiu regresoriumi (1 – atitinka tą senąją reikšmę, o 0 – visas kitas reikšmes). Kai regresorius ir taip dvireikšmis, tai perkodavę gauname vienintelį jam identišką pseudokintamąjį. Komanda:

```
xi: ologit  Bakna3 K11_1 K33_1 i.zalg2
```

yra ekvivalenti komandai:

```
ologit  Bakna3 K11_1 K33_1 zalg2
```

Uždaviniai

1. Failas **ZTLT**: Lietuvos tolerancijos profiliai, 2003 m. Respondentai atsakė, koks jų požiūris į musulmonus (1 – teigiamas, 2 – neutralus, 3 – neigiamas, 4 – sunku pasakyti) kintamasis, *r01_02*) ir į žydus (kintamasis *r01_03*); ar Lietuvoje turi būti toleruojamos kitos nekrikščioniškos religijos (1 – visiškai sutinku, 2 – sutinku, 3 – nesutinku, 4 – visiškai nesutinku, kintamasis *r04_04*). Kintamasis *s01* žymi lytį (1 – vyr., 2 – mot.).

Tyrimui atrinkti vyrus. Įsitikinti, kad kintamieji *r01_02* ir *r01*_03 įgyja labai mažai ketvertukų. Atrinkti tik respondentus, kuriems kintamieji *r01_02* ir *r01*_03 $\leq$ 3. Ištirti modelį *r01_02 =f(r01*_03, *r04_04).*

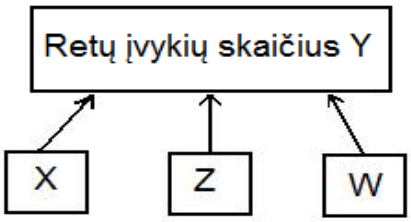
2. Pritaikyti ranginės regresijos modelį problemai nagrinėtai 5.3 skyrelyje. Kuris modelis geriau tinka?

7. PUASONO REGRESINĖ ANALIZĖ

Puasono regresijos modeliai taikomi retų įvykių skaičiui modeliuoti. Tipinės situacijos:

a) Norime ištirti žmogžudysčių skaičiaus per metus priklausomybę nuo miesto gyventojų skaičiaus, vidutinių gyventojų pajamų ir suvartojamo alkoholio kiekio vienam gyventojui.
b) Aiškinamės, kokie socialiniai ir ekonominiai faktoriai lemia vaikų skaičių šeimoje.
c) Tiriame, kaip vyrų, nešiojančių skrybėles, skaičius priklauso nuo įvairių regiono kultūrinių ir ekonominių charakteristikų.

Tarkime, kad skaičiuojame kažkokius pakankamai retus įvykius (Puasono modeliai dar vadinami katastrofų modeliais) ir manome, kad tų įvykių skaičius (pažymėkime jį Y) priklauso nuo kitų stebimų kintamųjų. Modelio schema:



Kintamasis Y vadinamas *priklausomu kintamuoju*, kintamieji X, Z, W vadinami *nepriklausomais* (*aiškinamaisiais*) *kintamaisiais* arba *regresoriais*. Dažnai intuityviai aišku, kokie įvykiai yra reti – draudiminių įvykių skaičius, telefono skambučių kiekis per valandą ir pan. Formalus požiūris – priklausomas kintamasis nėra normalus. Jeigu Y galima laikyti normaliu, tai geriau taikyti tiesinę regresiją. Specialus Puasono regresijos atvejis yra įvykių dažnio (įvykių skaičiaus per metus ar per valandą, įvykių skaičiaus tūkstančiui gyventojų ir pan.) modeliavimas.

7.1. Puasono regresijos modelis

7.1.1. Modelio lygtis ir jos taikymas prognozavimui

Matematinis modelis sudaromas ne pačiam priklausomam kintamajam Y, o jo vidurkiui, kurį pažymėsime simboliu μ. Modelis atrodo taip:

$$\mu = \mathrm{e}^{z},$$

$$z = C + b_1 X + b_2 Z + b_3 W.$$

Naudojant imties duomenis įvertinami modelio koeficientai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$. Jų ženklai padeda suprasti regresorių poveikį Y.

Jeigu koeficientas teigiamas, tai regresoriui didėjant, vidutinė *Y* reikšmė didėja.

Jeigu regresoriaus neigiamas, tai regresoriui didėjant, vidutinė *Y* reikšmė mažėja.

Į *modelio koeficientų eksponentes* (t.y. į skaičių e pakeltą koeficiento laipsniu) atsižvelgiama, kai norima įvertinti *kiek kartų* padidės (sumažės) kintamojo *Y* vidurkis μ, atitinkamam kintamajam padidėjus vienetu. Be pačių modelio koeficientų eksponenčių dar atsižvelgiama į jų pasikliautinuosius intervalus.

Kai *X* padidėja vienetu, kintamojo *Y* vidurkis μ padidėja (sumažėja) $e^{\beta 1}$ karto.

Vidutinio įvykių skaičiaus prognozavimas konkrečioms *X, Z, W* reikšmėms atliekamas taip: į formulę

$$z = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W$$

įstatome tas *X*, *Z* ir *W* reikšmes, kurioms norime gauti prognozę. Vidurkio μ įvertis gaunamas pagal pagal formulę: $\hat{\mu} = e^z$. Čia *e* = 2,718…

Pavyzdys. Tarkime, pavyzdžiui, kad atlikę Puasono regresiją, gavome įverčius $\hat{C} = -21, \hat{b}_1 = -4, \hat{b}_2 = 3, \hat{b}_3 = 1$, ir norime nustatyti, koks turėtų būti vidutinis *Y*, kai *X* = 0, *Z* =7, *W* = 2. Nesunkiai apskaičiuojame *z* = -21+0+21+2 = 2. Todėl $\hat{\mu} = 7{,}3890$ ... Nesunku pastebėti, kad prognozuojame vidutinę reikšmę, kuri dažniausiai nėra sveikas skaičius (ne taip, kaip *Y* reikšmės).

7.1.2. Modelis dažniui

Kartais protingiau modeliuoti ne įvykių skaičių, o jų dažnį. Daugiausia pasaulio genialių vaikų gyvena Kinijoje. Tai dar nereiškia, kad kinai ypatingai gabūs. Tiesiog Kinijoje daugiausia gyventojų. Todėl daugiausia vaikų su vystymosi sutrikimais irgi gyvena Kinijoje. Matome, kad šiuo atveju prasminga kalbėti ne apie bendrąjį genijų skaičių šalyje, o apie jų (tarkime su IQ = 300) skaičių milijonui gyventojų. Tuo atveju naudojamas modifikuotas Puasono regresijos modelis:

$$\frac{\mu}{s} = e^z,$$

$$z = C + b_1 X + b_2 Z + b_3 W.$$

Čia s yra normuojantis dažnio kintamasis (dažniausiai respondentų kiekis, kuriam surinkta informacija). Pavyzdžiui, *s* gali nurodyti, kiek milijonų gyventojų šalyje gyvena. Konstantos *C*, b_1, b_2, b_3 nėra žinomos. Jų įverčiai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$ gaunami, panaudojus imties duomenis. Praktikoje naudojamas ekvivalentus modelio užrašas:

$$\mu = e^{v},\ v = \ln(s) + C + b_1 X + b_2 Z + b_3 W.$$

Faktiškai tai tradicinis Puasono regresijos modelis, tik prie regresorių pridedama (logaritmuota ir su vieneto daugikliu) informacija apie tai, kokiam kiekiui respondentų gauti duomenys.

Prognozės norime vienam milijonui gyventojų, vienai stebėjimų valandai ir pan. Taigi faktiškai norime μ prognozės, kai $s = 1$. Vidutinio įvykių dažnio prognozavimas konkrečioms *X, Z,* W reikšmėms atliekamas taip:

a) Į formulę $v = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W$ įstatome tas *X, Z, W* ir reikšmes, kurioms norime gauti prognozę ($\ln 1 = 0$).
b) Dažnio įvertis gaunamas skaičių *e* pakėlus *v* laipsniu: $\hat{\mu} = e^{v}$.

Pavyzdys. Tarkime, kad skaičiavome, kiek SMS žinučių pasiunčia paaugliai. Žinome, paauglių amžių (*X*) ir lytį (*Z, Z*=0 mot., *Z* = 1 vyr.). Be to, paauglius stebėjome skirtingą valandų skaičių (kintamasis *s*). Šiuo atveju prasminga kalbėti apie vidutinį SMS žinučių skaičių per valandą. Kiekvieno paauglio pasiųstų SMS žinučių skaičių pažymime *Y*, o *Y* vidurkį μ. Tarkime, kad turimiems duomenims pritaikius Puasono regresiją, gavome

$$\ln \hat{\mu} = \ln s + 2 + 0{,}04X + 5Z.$$

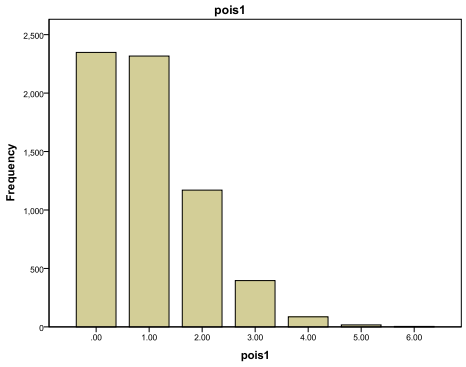
Kiek vidutiniškai SMS per valandą pasiunčia 13 metų paauglė? Įstatę į lygtį s = 1, X = 13, Z = 0, gauname: $\ln \hat{\mu} = \ln 1 + 2 + 0{,}04 \cdot 13 + 0 = 2{,}52$. Prognozuojamas vidutinis SMS per valandą skaičius bus $e^{2{,}52} = 12{,}42$.

7.1.3. Duomenys

Duomenys turi tenkinti tokius reikalavimus:

1) Kintamasis *Y* turi Puasono skirstinį. Tai reiškia, kad jis įgyja reikšmes 0, 1, 2,... , o jo tikimybės skaičiuojamos pagal gana griežtas taisykles. Svarbiausia Puasono kintamojo savybė – jo vidurkis sutampa su dispersija. Primename, kad dispersija charakterizuoja kintamojo reikšmių išsibarstymą, žr. įvadą. Todėl visada reikia pažiūrėti ar priklausomo kintamojo *Y* imties vidurkis nedaug skiriasi nuo dispersijos. Žemiau pateikiama Puasono

kintamojo dažnių diagrama, kai vidurkis lygus 1. Matome, kad skirstinys yra labai asimetriškas, visai nepanašus į normalų ir tiesinei regresijai netinkamas.



2) Dalis regresorių gali būti intervaliniai, dalis dvireikšmiai (įgyti reikšmes 0 arba 1) arba kategoriniai (jie keičiami dvireikšmiais pseudokintamaisiais). Modelį lengviau interpretuoti, kai regresoriai intervaliniai.

7.1.4. Modelio tinkamumas

Modelio tinkamumą duomenims parodančių charakteristikų nėra labai daug:

1) *Deviacija.* Deviacija yra vadinamas skaitinis rodiklis[16], kuris parodo kiek tiriamas modelis skiriasi nuo pilnai duomenis aprašančio (bet labai sudėtingo ir todėl naudoti netinkamo) modelio. Dar reikia atsižvelgti į imties didumą ir regresorių skaičių (juos nusako vadinamieji laisvės laipsniai, angl. *degrees of freedom, df*). Modelis pakankamai geras, jeigu deviacija, padalinta iš savo laisvės laipsnių, nedaug skiriasi nuo vieneto. Intervalas 0,9 – 1,1 visai priimtinas, reikšmė mažesnė už 0,7 (didesnė už 1,3) jau rodo blogą modelio tikimą.
2) *Pirsono chi kvadrato statistika* (Pearson Chi – Square) yra kiek rečiau taikomas deviacijos analogas. Matuoja, ar su modeliu gautųjų ir tikrųjų *Y* reikšmių skirtumai yra maži (statistiškai nereikšmingi). Gerai, kai taip. Tariama, kad modelis pakankamai geras, jeigu chi kvadrato statistika, padalinta iš savo laisvės laipsnių, nedaug skiriasi nuo vieneto (bent jau patenka į intervalą 0,8 – 1, 2).
3) *Tikėtinumo santykio Chi kvadrato statistika* (angl. Likelihood Ratio Chi-Square). Parodo, ar modelyje yra su priklausomu kintamuoju susijusių regresorių. Jeigu statistikos *p* reikšmė didesnė už 0,05, tai regresijos modelio tinkamumas labai abejotinas. Formaliai žiūrint, didelė *p* reikšmė reiškia, kad bandome *Y* elgesį prognozuoti pagal su juo visiškai

[16] Deviacijos išraiška yra pakankamai sudėtinga. Besidomintys ją ras 59 vadovėlio V.Čekanavičius, G.Murauskas „Statistika ir jos taikymai. III“ puslapyje.

nesusijusius kintamuosius. Pavyzdžiui, vidutinį vaikų skaičių lietuviškoje šeimoje pagal Afrikos dramblių tankį Kenijos savanoje. (Matyt, jie ten žiemojančius gandrus trypia).

4) *Informaciniai indeksai* (informaciniai kriterijai). Tai sąlyginiai rodikliai, įgalinantys „iš akies“ palyginti du modelius. Geresnis tas modelis, kurio informacinis indeksas mažesnis. Bene plačiausiai naudojamas Akaikės informacijos kriterijus AIC.
5) *Voldo testai.* Padeda nuspręsti ar kintamasis šalintinas iš modelio. Jeigu p reikšmė $< 0,05$, tai *kintamasis yra statistiškai reikšmingas* ir jį modelyje paliekame. Jeigu $p \geq 0,05$, tai kintamasis yra statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais. Dažniausiai modelio konstanta C paliekama net ir tada, kai ji statistiškai nereikšminga.

7.1.5. Puasono regresinės analizės etapai

Regresinė analizė susideda iš preliminaraus kintamųjų tinkamumo tyrimo ir modelio tinkamumo duomenims tikrinimo:

1) Patikriname, ar Y turi Puasono skirstinį. Kintamojo Y imties dispersija turi nedaug skirtis (idealiu atveju sutapti) nuo Y imties vidurkio. Galima sugeneruoti Puasono kintamąjį, turintį tą patį vidurkį, kaip ir Y, ir vizualiai palyginti abiejų kintamųjų dažnių diagramas.
2) Sudarome Puasono regresijos modelį ir patikriname ar deviacija, padalinta iš savo laisvės laipsnių, nedaug skiriasi nuo vieneto. Alternatyviai galima tikrinti, ar Pirsono chi kvadrato reikšmė, padalinta iš laisvės laipsnių, nedaug skiraisi nuo vieneto.
3) Patikriname ar tikėtinumo santykio chi kvadrato statistikos p reikšmė $< 0,05$. Jeigu p reikšmė $\geq 0,05$, tai modelis yra netinkamas.
4) Patikriname ar visi regresoriai statistiškai reikšmingi (visos Voldo kriterijaus $p < 0,05$). Jeigu ne – modelis taisytinas. Konstantai p reikšmės nežiūrime.
5) Pažiūrėję į koeficientų eksponentes, nustatome, kurie modelio kintamieji svarbiausi.
6) Jeigu viskas gerai – modelį aprašome. Jeigu ne – tobuliname.

Labai gerai duomenims tinkančiame modelyje:

- Deviacija padalinta iš laisvės laipsnių nedaug skiriasi nuo vieneto.
- Didžiausio tikėtinumo chi kvadrato kriterijaus $p < 0,05$.
- Visų regresorių Voldo kriterijaus $p < 0,05$.
- Pirsono chi kvadrato statistika padalinta iš laisvės laipsnių nedaug skiriasi nuo vieneto.

7.1.6. Modelio tobulinimas

Modelio tobulinimas – tai kintamųjų šalinimas arba jų transformacija.

Pašalinus bent vieną kintamąjį, keičiasi visos p reikšmės

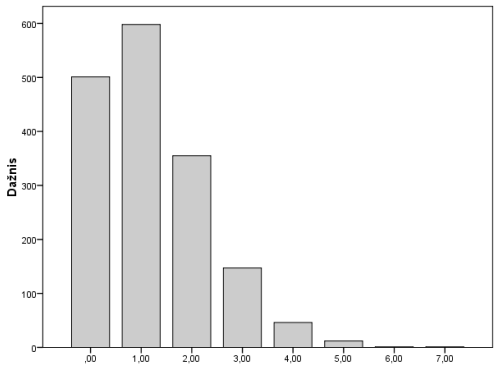
Modelis tobulinamas, jeigu yra problemų:

- *Normuotosios deviacijos (alternatyviai Pirsono chi kvardrato statistikos) ir jos laisvės laipsnių santykis* labai skiriasi nuo 1. Dažniausiai tai reiškia, kad parinkome netinkamus kintamuosius arba, kad priklausomas kintamasis visiškai nepanašus į Puasono kintamąjį. Bandome keisti modelį.
- *Kintamojo Y dispersija daug didesnė už jo vidurkį* (angl. overdispersion). Tada geriau naudoti negiamą binominę regresiją (žr. kitą skyrelį).
- *Per daug nulių* (angl. zero – inflated model). Kintamojo *Y* grafike nulio dažnis daug didesnis, nei turėtų įgyti Puasono atsitiktinis dydis. Kodėl atsiranda būtent nulių perviršis? Tarkime, tiriame respondentų vaikų skaičių. Nulį vaikų turi ne tik respondentai(ės), kurie nusprendė jų neturėti (vienuolės), bet ir tie, kurie dar nespėjo jų pagimdyti. Panašios situacijos iškyla gana dažnai. Kai nulių yra per daug, tai standartinė Puasono regresija netinka. Kai kuriuose statistiniuose paketuose numatyta pataisyta Puasono regresinė analizė, tinkanti ir tokiai situacijai.
- *Yra statistiškai nereikšmingas regresorius* (jam Voldo kriterijaus *p* reikšmė didesnė už 0,05). Regresorius pašalinamas iš modelio ir Puasono regresinė analizė pakartojama be jo. Visi rodikliai peržiūrimi iš naujo.

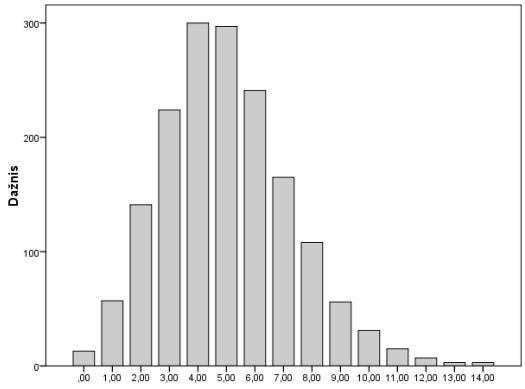
7.1.7. Pastabos apie Puasono regresijos taikymą

- Puasono regresija duomenims tinka gana retai. Dažnai ją galima pakeisti neigiama binomine regresija.
- Primename, kad nors stebime sveikus skaičius įgyjantį dydį, prognozuojame jo vidutinę reikšmę. Vidurkis visai neturi būti sveikas skaičius. Jeigu, taikydami Puasono regresijos prognozę šeimai, kurios abudu tėvai turi bakalaurinį išsilavinimą, gausime, kad šeimoje bus vidutiniškai 1,7 vaiko, tai nereiškia, kad šeimoje bus du vaikai, kurių vienas be kojos. Tai tiesiog prognozė, kad bendras visų šeimų, kuriose tėvai turi bakalaurinį išsilavinimą, vaikų skaičius padalintas iš tokių šeimų skaičiaus bus 1,7.

- Reikia nemaišyti tikėtinumo santykio chi kvadrato (Likelihood Ratio Chi-Square) ir Pirsono chi kvadrato (Pearson Chi-Square). Pastaroji statistika yra deviacijos analogas ir rodo geresnį modelio tikimą, kai yra *maža*. Tikėtinumo santykio chi kvadrato statistika lygina tiriamą modelį su neturinčiu jokių regresorių ir rodo geresnį modelio tikimą, kai yra *didelė* (statistiškai reikšminga).
- Puasono kintamasis, turintis didelį vidurkį, nedaug skiriasi nuo normaliojo, o tuo atveju jau galima taikyti paprastą tiesinę regresiją. Todėl visada rekomenduojama nusibraižyti *Y* dažnių diagramą arba histogramą. Jeigu ji ryškiai asimetrinė (a) pav.), jokių kalbų apie tiesinę regresiją negali būti. Jeigu diagrama primena normaliąją kreivę (b) pav.), galima pamąstyti ir apie tiesinės regresijos taikymą.



a) Vidurkis lygus 1,2



b) Vidurkis lygus 5

- Nereikia suabsoliutinti normuotosios deviacijos ir kitų charakteristikų vaidmens. Nesunku sukonstruoti pavyzdį visai be regresorių, kada normuotoji deviacija maža, o konstanta C statistiškai reikšminga. Modelis turi būti dar ir logiškas.
- Galima dirbtinai transformuoti duomenis, kad jie pasidarytų panašūs į puasoniškus. Tam reikia pritaikyti svorį lygų deviacijos ir laisvės laipsnių santykiui. Mums toks duomenų keitimas atrodo kiek dirbtinas ir šiuose konspektuose jis nagrinėjamas nebus.

7.1.8. Standartiniai Puasono regresijos modelio aprašymai.

Visada aprašomas galutinis, po visų tobulinimų gautasis, modelis. Pateiksime keletą galimų aprašymų fragmentų.

- *Pritaikytas Puasono regresijos modelis gerai tiko duomenims,* $\chi^2 = 13{,}27$; $p < 0{,}01$. *Deviacijos ir jos laisvės laipsnių santykis lygus 1,08. Visi modelio kintamieji statistiškai reikšmingi. Kintamojo Y vidurkio regresinis modelis atrodo taip:* $\ln \hat{\mu} = -2 + 0{,}46X + 5Z$. *Kai X padidėja 1, kintamojo Y vidurkis išauga 1,58 karto.....*

- *Iš modelio buvo pašalintas statistiškai nereikšmingas kintamasis X. Galutinis modelis buvo statistiškai reikšmingas, $\chi^2 = 10,02$; p = 0, 023; deviacijos ir laisvės laipsnių santykis 0,99.*

7.2. Puasono regresinė analizė su SPSS

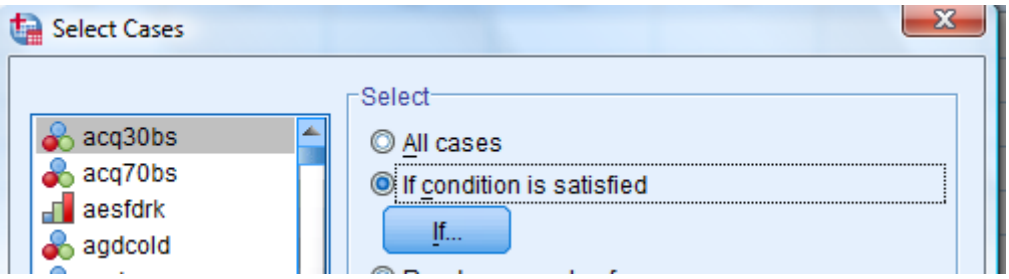
7.2.1. Duomenys

Kaip atlikti Puasono regresinę analizę parodysime, tirdami 2008 metų Europos Sąjungos socialinio tyrimo Prancūzijos duomenis **ESS4FR**. Tyrime naudosime tokius kintamuosius:

- *agea* – respondento amžius,
- *hhmmb* – namų ūkio pastovių narių skaičius,
- *imptrad* – svarbu laikytis tradicijų ir papročių (1 – visai, kaip aš,...., 6 – absoliučiai ne taip, kaip aš) ,
- *eduyrs* – prasimokytų metų skaičius,
- *cldcrsv* – paramos vaiko priežiūrai teikimas tėvams, turintiems teisę ją gauti (0 – itin blogas,....., 10 – labai geras).

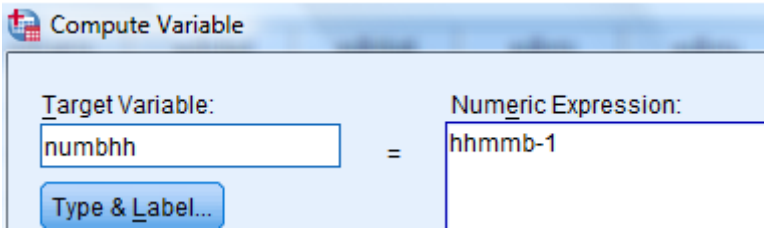
Modeliuosime kiek namų ūkyje yra gyventojų be paties respondento(ės). Regresoriai bus amžius (*agea*) ir požiūris į paramos vaiko priežiūrai prieinamumą (*cldrsv*). Tirsime stipriai tradicijų besilaikančius (*imptrad* ≤ 2) aukštojo mokslo neragavusius (*eduyrs* ≤ 10) respondentus(es).

Su komanda **Select Cases** atrenkame reikiamus respondentus. Tai daroma taip: meniu **Select Cases** pažymime **If condition is satisfied** ir spaudžiame klavišą **If**.



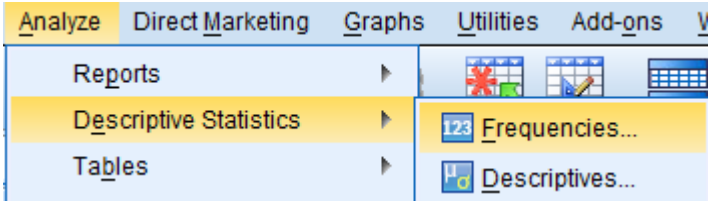
Naujajame meniu suformuluojame sąlygą *imptrad <= 2 & eduyrs <= 10*. Tada spaudžiame **Continue** ir **OK.**

Dar reikia sukurti priklausomą kintamąjį. Pavadinkime jį *numbhh*. Aišku, kad *numbhh=hhmmb-1*. Todėl reikiamą kintamąjį sukuriame **Transform → Compute Variable** komanda.

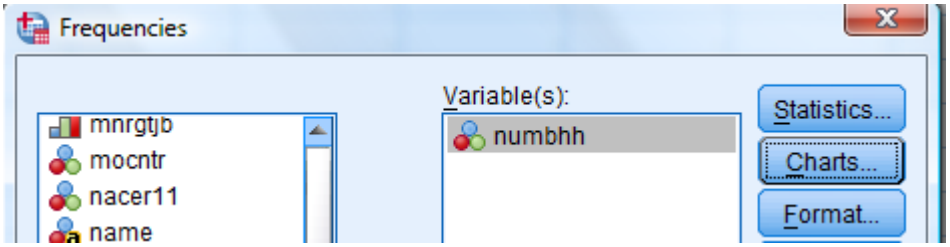


7.2.2. Preliminari analizė

Pradžioje pasižiūrėsime ar priklausomas kintamasis *numbhh* panašus į Puasono kintamąjį. Aišku, kad jis įgyja sveikas neneigiamas reikšmes. Taip pat labai tikėtina, kad respondentų iš daugiavaikių šeimų nėra daug. Nusibraižysime dažnių stulpelių grafiką ir suskaičiuosime empirinį vidurkį ir dispersiją. Spaudžiame **Analyze → Descriptive Statistics → Frequences.**



Atsivėrusiame meniu perkeliame kintamąjį *numbhh* į langelį **Variable(s)** ir spaudžiame **Statistics.** Atsivėrusiame meniu pažymime **Mean** ir **Variance.** Grįžę į ankstesnįjį meniu spaudžiame **Charts**. Pažymime **Bar chart.** Grįžtame į ankstesnį meniu ir spaudžiame **OK.**



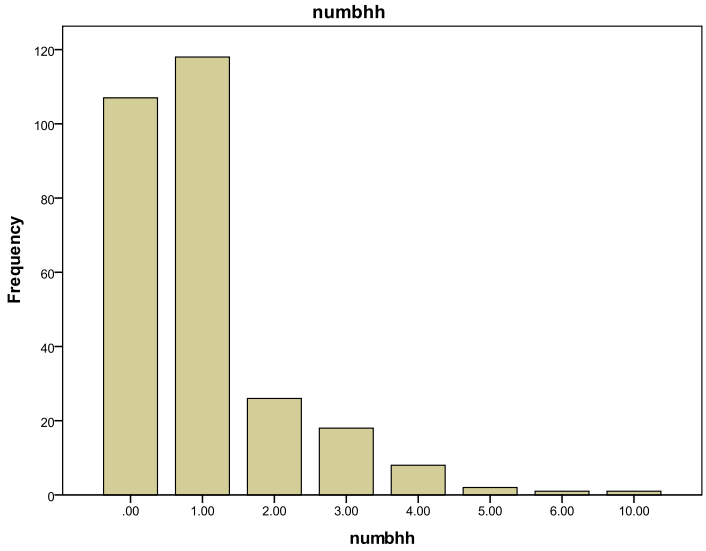
Rezultatų išklotinėje matyti, kad *numbhh* vidurkis (1,0036) nedaug skiriasi nuo dispersijos (1,482). Tai reiškia, kad *numbhh* tenkina vieną svarbiausių Puasono kintamojo savybių.

Statistics

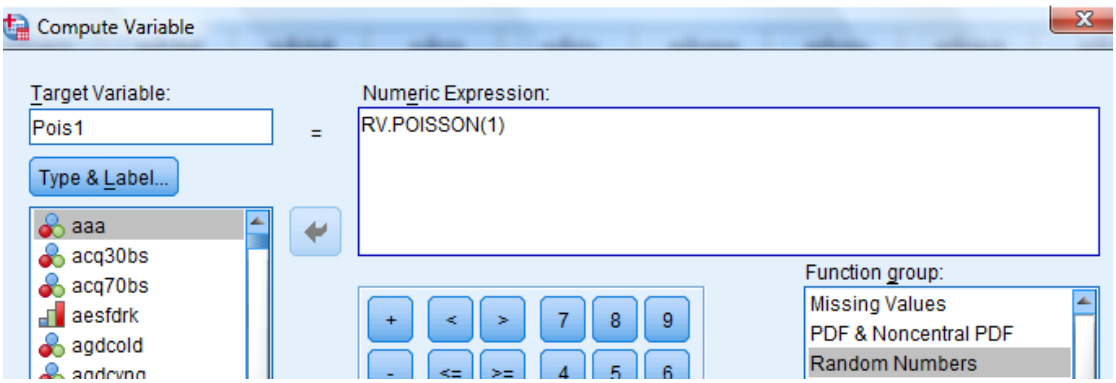
numbhh

N	Valid	281
	Missing	0
Mean		1.0036
Variance		1.482

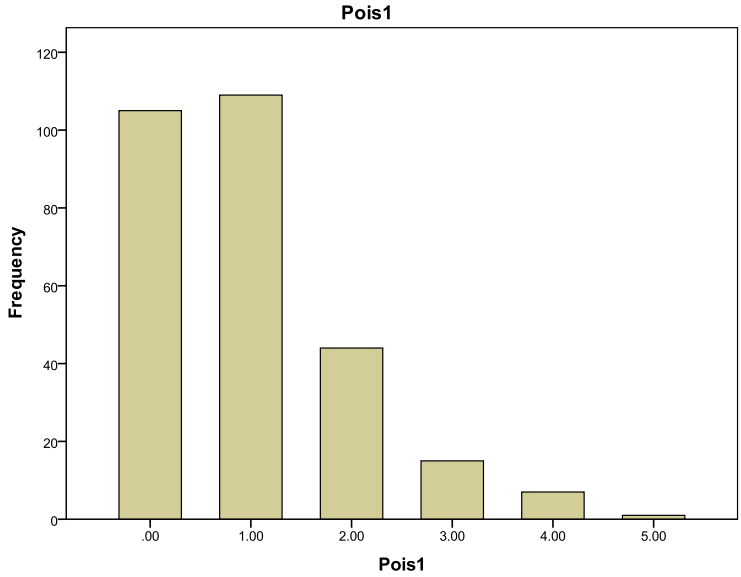
Panašumą į Puasono skirstinį rodo ir dažnių stulpelių diagrama. Skirstinys asimetriškas su daug reikšmių kairėje pusėje.



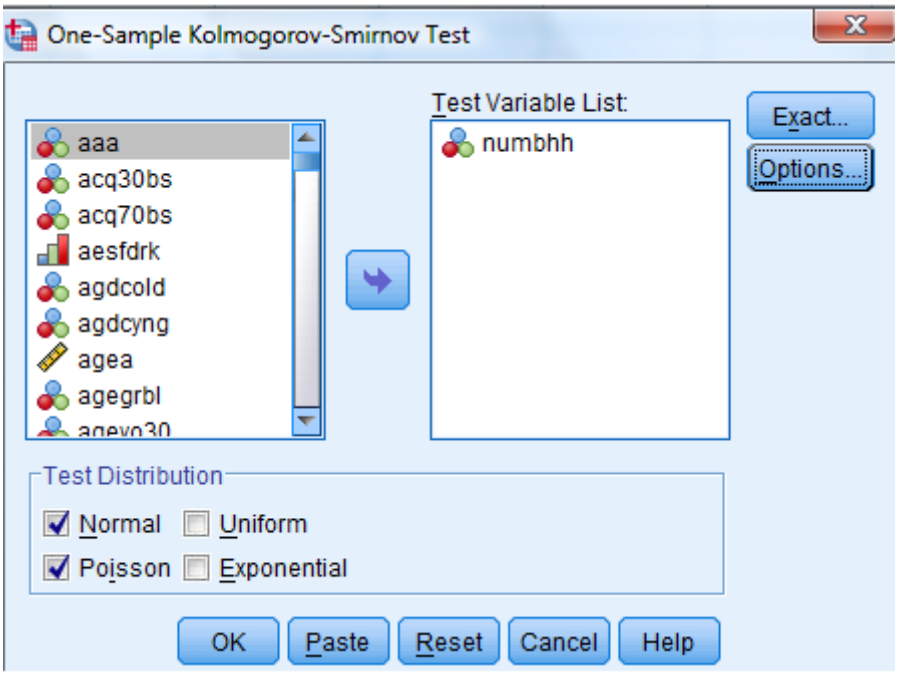
Dažniausiai tyrimuose ir apsiribojama vidurkio bei dispersijos reikšmėmis ir dažnių grafiku. Parodysime, ką dar galima padaryti preliminarioje kintamųjų analizėje. Norint įsivaizduoti, kaip gali atrodyti Puasono skirstinys su vidurkiu lygiu vienetui, galima atsitiktinai sugeneruoti atitinkamą Puasono skirstinį ir nubraižyti jo dažnių grafiką. Parodysime, kaip generuojamas Puasono dydis. Spaudžiame **Transform → Compute Variable.** Atidariusiame meniu dar gali būti likę kintmojo *numbhh* kūrimo pėdsakų. Viską išvalome, paspaudę **Reset**. Langelyje **Target Variable** įrašome naujo kintamojo vardą *Pois1*. Langelyje **Function Group** pasirenkame **Random Numbers**. Langelyje **Functions and Special Variables** pasirenkams **Rv.Poisson**. Paspaudę perkeliame šį užrašą į langelį **Numeric Expression**. Perkeltoje išraiškoje dar reikia įrašyti generuojamo dydžio vidurkį (Puasono skirstinio parametrą). Kadangi tiriamo *numbhh* vidurkis buvo praktiškai 1, tai įrašome 1. Paspaudžiame **OK**.



Nubraižę gautojo kintamojo dažnių grafiką įsitikiname, kad jis gana panašus į kintamojo *numbhh* grafiką.



Galima kintamojo puasoniškumą tikrinti ir **Kolmogorovo-Smirnovo** kriterijumi. Tik reikia turėti omenyje, kad šis kriterijus nėra galingas. Todėl neretai aukščiau aprašytas vertinimas „iš akies“ yra naudingesnis. Vis dėlto parodysime, kaip taikomas Kolmogorovo-Smirnovo kriterijus. Spaudžiame **Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S.** Kintamąjį *numbhh* perkeliame į laukelį **Test Variable List.** Pažymime **Poisson** ir **Normal**. Pastarąją opciją pasirinkome tam, kad galėtume įsitikinti jog *numbhh* nėra normalusis kintamasis.



Kintamojo skirstinys statistiškai reikšmingai skiriasi pasirinktojo skirstinio, jeigu *p* reikšmė < 0,05. Matome, kad pagal Kolmogorovo – Smirnovo kriterijų *numbhh* statistiškai reikšmingai nesiskiria nuo Puasono kintamojo ($p = 0{,}169$), bet skiriasi nuo normaliojo ($p = 0{,}000$).

One-Sample Kolmogorov-Smirnov Test

		numbhh
N		281
Normal Parameters[a,b]	Mean	1.0036
	Std. Deviation	1.21743
Most Extreme Differences	Absolute	.302
	Positive	.302
	Negative	-.205
Kolmogorov-Smirnov Z		5.060
Asymp. Sig. (2-tailed)		.000

a. Test distribution is Normal.

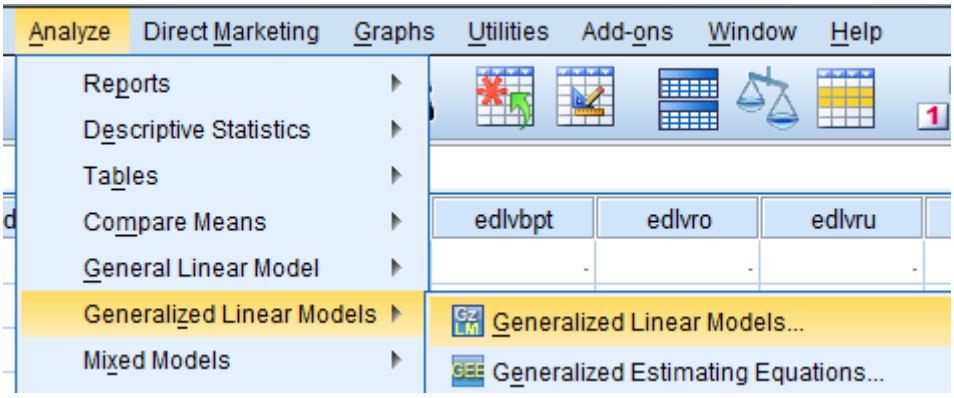
One-Sample Kolmogorov-Smirnov Test 2

		numbhh
N		281
Poisson Parameter[a,b]	Mean	1.0036
Most Extreme Differences	Absolute	.066
	Positive	.066
	Negative	-.026
Kolmogorov-Smirnov Z		1.111
Asymp. Sig. (2-tailed)		.169

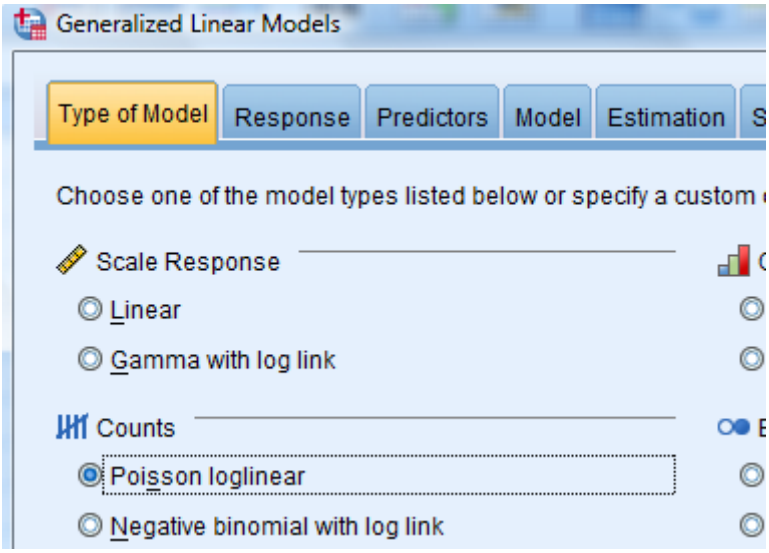
a. Test distribution is Poisson.

7.2.3. SPSS parinktys Puasono regresijos modeliui

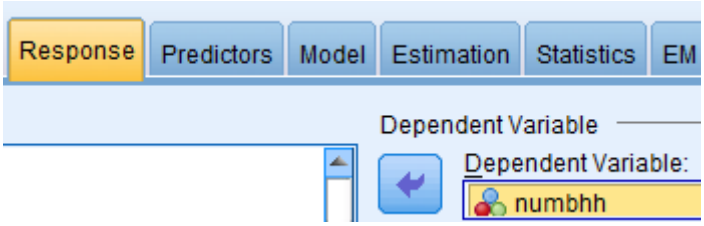
Patikrinsime, kaip atrinktiesiems respondentams tinka Puasono regresijos modelis. Primename, kad tikriname modelį *numbhh = f(agea, cldcrsv)*. Spaudžiame **Analyze → Generalized Linear Models → Generalized Linear Models.**



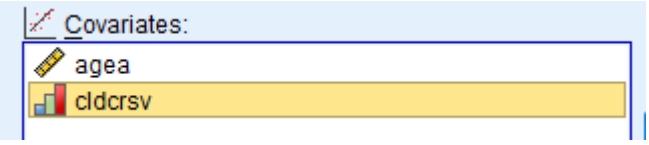
Atsidariusiame meniu spaudžiame **Type of Model** ir pažymime **Poisson loglinear**.



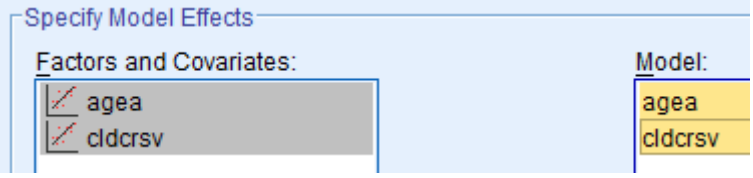
Po to spaudžiame **Response** ir perkeliame *numbhh* į laukelį **Dependent Vriable**.



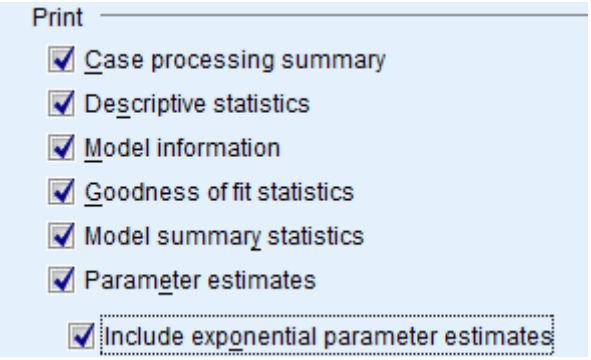
Paspaudę **Predictors** abudu kintamuosius *agea* ir *cldcsrv* perkeliame į langelį **Covariates**. Jeigu būtų kategorinių kintamųjų, juos keltume į **Factors**.



Pasirinkę **Model**, abudu kintamuosius perkeliame į langą **Model.**



Pasirinkę **Statistics** papildomai pažymime **Include exponential parameter estimates.** Jau pažymėtus langelius paliekame ramybėje. Spaudžiame **OK**.



7.2.4. Rezultatai

Rezultatų išklotinė prasideda nuo informacijos apie tolydžiuosius kintamuosius. Atkreipiame dėmesį, kad, lyginant su preliminariu tyrimu, kintamojo *numbhh* vidurkis ir standartinis nuokrypis kiek pasikeitė. Taip atsitiko todėl, kad 27 atrinktieji respondentai neatsakė į *cldcrsv* klausimą, todėl iš tyrimo buvo automatiškai pašalinti. Vis dėlto stebėjimų liko pakankamai daug – 254 .

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Numbhh	254	.00	10.00	1.0472	1.24385
Covariate	agea	254	15	93	63.10	16.153
	cldcrsv	254	0	10	5.26	2.170

Lentelėje **Goodness of Fit** yra surašytos deviacijos reikšmė, Pirsono chi kvadrato statistika ir t.t. Normuotos (angl. scaled) deviacijos ir jos laisvės laipsnių santykis tinkamo Puasono regresijos modelio atveju turi būti artimas vienetui. Matome, kad šis santykis yra 0,919. Gavome patvirtinimą, kad apskritai modelis tinka duomenims. Dar liko nustatyti regresorių svarbą. Beje Pirsono chi kvadrato statistika, padalinta iš savo laisvės laipsnių, irgi turi būti arti vieneto. Vis dėlto, manoma, kad ši modelio tikimo charakteristika yra mažiau svarbi, nei deviacija. Lentelėje surašytos informacinių kriterijų reikšmės (AIC, BIC ir pan.) naudojamos kelių modelių palyginimui. Geresnis būtų tas modelis, kurio informacinių kriterijų reikšmės mažesnės.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	230.635	251	.919
Scaled Deviance	230.635	251	
Pearson Chi-Square	188.314	251	.750
Scaled Pearson Chi-Square	188.314	251	
Log Likelihood[a]	-301.040		
Akaike's Information Criterion (AIC)	608.080		
Finite Sample Corrected AIC (AICC)	608.176		
Bayesian Information Criterion (BIC)	618.692		
Consistent AIC (CAIC)	621.692		

Lentelėje **Omnibus Test** yra didžiausio tikėtinumo santykio chi kvadrato statistika ir jos p reikšmė. Gerai, kai $p < 0,05$, nes tada modelyje yra bent vienas reikalingas (statistiškai reikšmingas) regresorius. Kadangi $p = 0,000...$, tai galime teigti, kad regresijos modelio tinkamumą patvirtina ir didžiausio tikėtinumo chi kvadrato kriterijus.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
112.919	2	.000

Lentelėje Tests of Model Effects yra informacija apie Voldo kriterijų. Kadangi visos p = 0,000 < 0,05, tai darome išvadą, kad abudu kintamieji (*agea* ir *cldcrsv*) yra statistiškai reikšmingi ir modelyje reikalingi.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	41.188	1	.000
agea	105.703	1	.000
cldcrsv	14.395	1	.000

Lentelėje **Parameter Estimates** pasikartoja informacija apie Voldo kriterijų. Taip pat surašyti parametrų įverčiai. Parametrų įverčių ženklai ir **Exp(B)** suteikia informaciją apie tai, kaip keisis vidutinis kitų namų ūkio gyventojų skaičius, keičiantis kintamiesiems *agea* ir *cldcrsv*.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	1.535	.2392	1.066	2.004	41.188	1	.000	4.642	2.905	7.419
agea	-.035	.0034	-.042	-.028	105.703	1	.000	.966	.959	.972
cldcrsv	.099	.0261	.048	.150	14.395	1	.000	1.104	1.049	1.162
(Scale)	1[a]									

Kintamojo *agea* koeficientas lygus -0,035 < 0. Tai reiškia, kad didėjant atsakinėjusio respondento amžiui, vidutinis kitų namų ūkio narių skaičius mažėja. Stulpelyje **Exp(B)** užrašyta reikšmė 0,966 rodo, kad vieneri papildomi respondento amžiaus metai sumažina tikėtiną kitų namų ūkio narių skaičių 0,966 **karto** (narių skaičių reikia dauginti iš 0,966). Stulpeliuose **Lower** ir **Upper** nurodyti **Exp(B)** pasikliautinųjų intervalų rėžiai. Pasikliautiniai intervalai naudojami, norint patikslinti informaciją apie koeficiento įvertį. Pavyzdžiui, pasikliautinis intervalas kintamojo *agea* daugiklio eksponentei **Exp(B)** yra [0,959; 0,972] ir interpretuotinas taip: labai tikėtina, kad vieneri papildomi respondento amžiaus metai sumažina tikėtiną kitų namų ūkio narių skaičių nuo 0,959 iki 0,972 karto.

Matematinis modelio užrašas atrodo taip:

$$\hat{\mu} = \exp\{1{,}535 - 0{,}035\, agea + 0{,}099 cldcrsv\}.$$

Čia $\hat{\mu}$ žymi vidutinį kitų namų ūkio narių skaičių.

Atrodytų viskas modelyje gerai. Vis dėlto, nors abudu regresoriai statistiškai reikšmingi, Voldo statistikos reikšmės (105,7 ir 14,39) skiriasi septynis kartus. Kuo Voldo statistika didesnė,

tuo koeficientas statistiškai reikšmingesnis. Todėl ramybės dėlei išbandome Puasono regresiją be regresoriaus *cldcrsv*. Nebepateiksime visų lentelių tik konstatuosime, kad

- deviacijos ir laisvės laipsnių santykis yra 0,958;
- didžiausio tikėtinumo chi kvadrato p = 0,000...< 0,05;
- Voldo kriterijaus p = 0,000...< 0,05;
- modelis $\hat{\mu} = \exp\{2{,}02 - 0{,}034\, agea\}$.

Darome išvadą, kad naujasis modelis tinkamesnis, nei pradinis. Parodysime, kaip šį modelį taikyti konkrečioms prognozėms. Tarkime, kad respondentui 50 metų. Kiek dar yra narių jo namų ūkyje? Įstatę kintamųjų reikšmes į aukščiau pateiktą modelio lygtį, gauname

$$\hat{\mu} = \exp\{2{,}02 - 0{,}035 \cdot 50\} = \exp\{0{,}27\} = 1{,}3...$$

Darome prognozę, kad penkiasdešimtmečių namų ūkiuose be pačių respondentų dar turi būti vidutiniškai[17] 2,17 narys.

7.2.5. SPSS komandų sintaksė Puasono regresinei analizei

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam Puasono regresijos pavyzdžiui. Jas galima naudoti sintaksės lange, pasirenkant opciją **Run**. Duomenų atranka:

```
USE ALL.
COMPUTE filter_$=(imptrad <= 2 & eduyrs <= 10).
VARIABLE LABELS filter_$ 'imptrad <= 2 & eduyrs <= 10 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

Naujo kintamojo kūrimas, jo vidurkis, dispersija ir dažnių grafikas:

```
COMPUTE numbhh=hhmmb-1.
EXECUTE.
FREQUENCIES VARIABLES=numbhh
  /STATISTICS=VARIANCE MEAN
  /BARCHART FREQ  /ORDER=ANALYSIS.
```

[17] Čia visų penkiasdešimtmečių vidurkis! Su vienais gyvena daugiau, su kitais – mažiau žmonių, o ne kiekvienoje šeimoje po invalidą.

Puasono kintamojo generavimas:

```
COMPUTE Pois1=RV.POISSON(1). EXECUTE.
```

Kolmogorovo – Smirnovo testas:

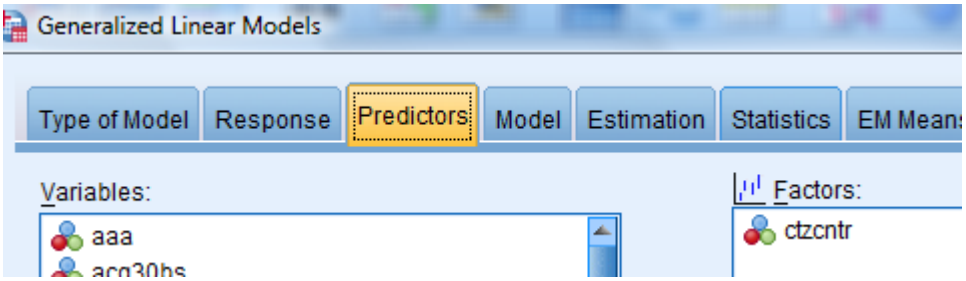
```
NPAR TESTS
  /K-S(NORMAL)=numbhh
  /K-S(POISSON)=numbhh
/MISSING ANALYSIS.
```

Pats Puasono regresijos modelis:

```
GENLIN numbhh WITH agea cldcrsv
 /MODEL agea cldcrsv INTERCEPT=YES
 DISTRIBUTION=POISSON LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL
MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012
ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT
SUMMARY SOLUTION (EXPONENTIATED).
```

7.2.6. Kategorinis regresorius

Jeigu į modelį reikia įtraukti kategorinį regresorių (pvz., kintamąjį *ctzntr* – ar yra pilietis), tai meniu **Generalized Linear Models - Predictors** jį įkeliame į langą **Factors.**



Nepamirštame *ctzntr* perkelti **Model** meniu į **Model** langą. Tolesnis modelio tyrimas parodo, kad kintamasis *ctzntr* nėra statistiškai reikšmingas. Vis dėlto parodysime, kaip reikėtų užrašyti modelį su šiuo kintamuoju. Pateikiame lentelės **Parameter Estimates** fragmentą:

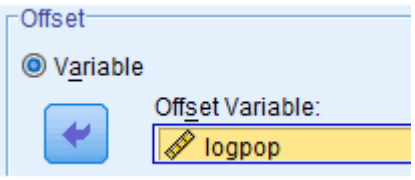
Parameter	B	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
(Intercept)	1.239	.3115	.629	1.850
agea	-.036	.0034	-.043	-.029
[ctzcntr=1]	.352	.2319	-.103	.806
[ctzcntr=2]	0[a]	.	.	.
cldcrsv	.104	.0263	.053	.156

Matematinė lygtis atrodys taip:

$$\ln \hat{\mu} = 1{,}239 - 0{,}036agea + 0{,}104cldcrsv + \begin{cases} 0{,}352, & \text{jei } ctzcntr = 1 \\ 0, & \text{jei } ctzcntr = 2 \end{cases}.$$

7.2.7. Puasono regresinė analizė dažniui

Tarkime, kad norime sumodeliuoti skirtingoms šalims AIDS sergamumą, atsižvelgdami į įvairius socialinius ir ekonominius faktorius. Žinome bendrąjį sergančiųjų AIDS skaičių *aids* ir gyventojų skaičių *pop*. Norime modelio santykiui *aids/pop*. Tada su komanda **Compute Variable** sukuriame kintamąjį *logpop* lygų *pop* logaritmui. Tada renkamės visas tas pačias opcijas, kaip ir standartinėje Puasono regresijoje (**Dependent Variable** *aids*), tiktai **Predictors** lange į lauką **Offset Variable** įkeliame *logpop*.



Visi rezultatai interpretuojami, kaip ir ankstesniu atveju. Tiktai dabar modelis sudarytas dažnio *aids/pop* logaritmui, o ne kintamojo *aids* logaritmui. Beje lentelėje **Parameter Estimates** atsiranda kintamasis **(Scale)** su koeficientu 1.

7.3. Puasono regresinė analizė su STATA

7.3.1. Duomenys

Pakartosime Puasono regresinę analizę jau atliktą su SPSS. Tiriame 2008 metų Prancūzijos duomenis **ESS4FR**. Tyrime naudosime tokius kintamuosius: *agea* – respondento amžius, *hhmmb* – namų ūkio pastovių narių skaičius, *cldcrsv* – paramos vaiko priežiūrai teikimas tėvams, turintiems teisę ją gauti (0 – itin blogas,....., 10 – labai geras), *imptrad* – svarbu laikytis tradicijų ir papročių (1 – visai, kaip aš,...., 6 – absoliučiai ne taip, kaip aš) , *eduyrs* – prasimokytų metų skaičius.

Tirsime stipriai tradicijų besilaikančius (*imptrad* ≤ 2) aukštojo mokslo neragavusius (*eduyrs* ≤ 10) respondentus(es):

```
keep if imptrad <= 2 & eduyrs <= 10
```

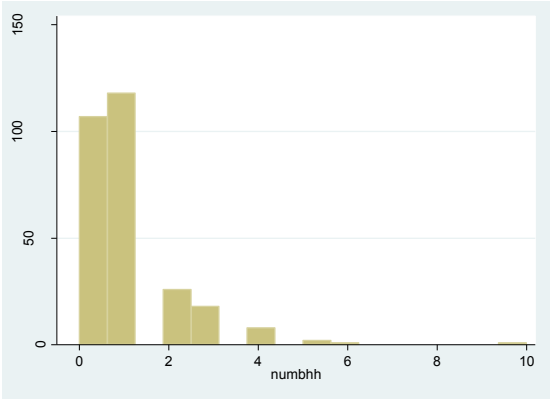
Modeliuosime kiek namų ūkyje yra gyventojų be paties respondento(ės):

```
gen numbhh = hhmmb - 1
```

Pasižiūrėsime ar priklausomas kintamasis *numbhh* panašus į Puasono kintamąjį. Aišku, kad jis įgyja sveikas neneigiamas reikšmes. Įvykdome: (ės):

```
summarize numbhh, detail
tab numbhh
hist  numbhh, freq
```

```
. summarize numbhh, detail

                          numbhh
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 281
25%            0              0       Sum of Wgt.         281

50%            1                      Mean           1.003559
                        Largest       Std. Dev.      1.217428
75%            1              5
90%            3              5       Variance        1.48213
95%            3              6       Skewness        2.49157
99%            5             10       Kurtosis       14.27513


      numbhh |      Freq.     Percent        Cum.
------------+-----------------------------------
           0 |        107       38.08       38.08
           1 |        118       41.99       80.07
           2 |         26        9.25       89.32
           3 |         18        6.41       95.73
           4 |          8        2.85       98.58
           5 |          2        0.71       99.29
           6 |          1        0.36       99.64
          10 |          1        0.36      100.00
------------+-----------------------------------
       Total |        281      100.00
```



Dispersija (1,48) nedaug skiriasi nuo vidurkio (1). Tai reiškia, kad *numbhh* tenkina vieną svarbiausių Puasono kintamojo savybių. Panašumą į Puasono skirstinį rodo ir dažnių stulpelių diagrama. Skirstinys asimetriškas su daug reikšmių kairėje pusėje.

7.3.2. Puasono modelio taikymas

Ištirsime modelį *numbhh = f(agea, cldcrsv)*. Primename, kad mes nebandome komentuoti nustatytų ryšių, nes tai – jau specialisto sociologo duona. (Žinoma, dvi minutes palaužę mažąjį pirštelį, nesunkiai išlaužėme tokią teoriją: respondentas tradicionalistas – kontracepcijos nepripažįsta, o gyvena jau ilgai. Mano, kad valstybės teikiama parama vaikams auginti pakankamai didelė. Tai, ko čia stebėtis, kad septyniolika vaikų). Įvykdome:

```
poisson numbhh agea cldcrsv
```

```
Poisson regression                                      Number of obs   =        254
                                                        LR chi2(2)      =     112.92
                                                        Prob > chi2     =     0.0000
Log likelihood = -301.04013                             Pseudo R2       =     0.1579

------------------------------------------------------------------------------
      numbhh |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agea |   -.03498   .0034023   -10.28   0.000    -.0416484   -.0283116
     cldcrsv |  .0988642   .0260573     3.79   0.000     .0477928    .1499355
       _cons |  1.535153   .2392022     6.42   0.000     1.066326    2.003981
------------------------------------------------------------------------------
```

Virš lentelės pateikta didžiausio tikėtinumo chi kvadrato statistikos reikšmė (112,92) ir p = 0,00. Kadangi p reikšmė mažesnė už 0,05, tai gavome vieną iš įrodymų, jog modelis gali tikti. Pačioje lentelėje įrašyti koeficientų įverčiai (su 95% pasikliautiniais intervalais) ir Voldo testų reikšmės. Regresoriai *agea* ir *cldcrsv* statistiškai reikšmingi (p < 0,05). Konstantai p reikšmės netikriname.

Matematinis modelio užrašas atrodo taip:

$$\text{vidutinis kitų namų ūkio narių skaičius} = \exp\{1{,}535 - 0{,}035\, agea - 0{,}099 cldcrsv\}.$$

Regresoriaus *agea* koeficientas neigiamas – kuo vyresnis respondentas, tuo iš mažesnio namų ūkio. Taigi iš piršto laužtoji teorija nepasitvirtino.

Norėdami gauti koeficientų įverčių eksponentes, įvykdome :

```
poisson  numbhh agea cldcrsv, irr
```

numbhh	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agea	.9656247	.0032854	-10.28	0.000	.9592069	.9720854
cldcrsv	1.103916	.0287651	3.79	0.000	1.048953	1.161759

Amžiui padidėjus 1 metais, namų ūkio narių skaičius sumažėja vidutiniškai 0,966 karto.

Deviacijos STATA programa Puasono regresijai neskaičuoja. Kaip galimas bendrojo modelio tikimo indikatorius pateikiamas Makfadeno determinacijos koeficientas R^2 = 0,144, kuris nėra didelis (ir yra labai nepatikimas). Žinoma, galima būtų su komanda *fitstat* rasti ir kitus pseudo-koeficientus (ir paaiškėtų, kad Nagelkerkės R^2 = 0,35). Vis dėlto, protingesnė alternatyva yra *Pirsono chi kvadrato statistika.* Ją suskaičiuojame, įvykdę :

```
estat gof
```

```
. estat gof

         Goodness-of-fit chi2  =  230.6347
         Prob > chi2(251)      =    0.8172
```

STATA nepateikia statistikos ir jos laisvės laipsnių santykio, bet pateikia *p* reikšmę. Jeigu *p* reikšmė ≥ 0,05, tai rodo gerą modelio tikimą. Patys surandame statistikos ir jos laisvės laipsnių santykį (laisvės laipsniai yra 251): 230,6347/251 = 0,9188... Šis santykis artimas vienetui ir taip pat rodo pakankamai gerą modelio tikimą.

Pabandome modelį be regresoriaus *cldcrsv*. Paeiliui įvykdome komandas

```
poisson  numbhh agea
estat gof
```

```
         Goodness-of-fit chi2  =  244.7568
         Prob > chi2(252)      =    0.6164
```

Matome, kad $p = 0{,}6164 > 0{,}05$, o Pirsono chi kvadrato statistikos ir jos laisvės laipsnių santykis 0,96 net geresnis, nei anksčiau. Galima puikiausiai apsieiti modeliu *numbhh = f(agea).* Tik jo tiriamoji vertė abejotina.

7.3.3. Prognozavimas

Pirminiam modeliui parodysime, kaip atlikti visutinio namų ūkio narių skaičiaus prognozę 50 mečiui respondentui. Įvykdome

```
adjust agea=50, exp
```

```
------------------------
     All |    exp(xb)
---------+--------------
         |    1.35873
------------------------
    Key:  exp(xb)  =  exp(xb)
```

Gavome, kad be pačių penkiasdešimtmečių, namų ūkiuose vidutiniškai gyvena dar po 1,346 žmogaus.

Beje, galima paprognozuoti ir pačias tikimybes, kad kartu negyvens nieks, gyvens vienas žmogus, gyvens dar du žmonės ir t.t. Įvykdome

```
prvalue, x(agea = 50)
```

```
Confidence intervals by delta method

                                 95% Conf. Interval
  Rate:                1.3587   [ 1.1902,    1.5272]
  Pr(y=0|x):           0.2570   [ 0.2137,    0.3003]
  Pr(y=1|x):           0.3492   [ 0.3336,    0.3647]
  Pr(y=2|x):           0.2372   [ 0.2184,    0.2561]
  Pr(y=3|x):           0.1074   [ 0.0856,    0.1293]
  Pr(y=4|x):           0.0365   [ 0.0245,    0.0484]
  Pr(y=5|x):           0.0099   [ 0.0054,    0.0144]
  Pr(y=6|x):           0.0022   [ 0.0010,    0.0035]
  Pr(y=7|x):           0.0004   [ 0.0001,    0.0007]
  Pr(y=8|x):           0.0001   [ 0.0000,    0.0001]
  Pr(y=9|x):           0.0000   [ 0.0000,    0.0000]

         agea    cldcrsv
x=         50  5.2637795
```

Jau matytas vidurkis 1,3461 pateikiamas lentelės viršuje. Tikimybė, kad kartu negyvens jokių žmonių lygi 0,26; kad vienas – 0,35; kad du – 0,23. Visos kitos tikimybės daug mažesnės.

7.3.4. Kategorinis regresorius

Jeigu regresorius įgyja tik 0 arba 1, tai jį galima į modelį įtraukti be jokių pakeitimų. Su STATA programa perkodavimas vyksta automatiškai. Reikia prieš *poisson* parašyti *xi:,* o prieš kategorinius kintamuosius *i.* Pagrindinis regresijos modelis sudaromas mažiausiai kategorinio regresoriaus reikšmei. Visoms kitoms reikšmėms sukuriami nauji dvireikšmiai regresoriumi (1 – atitinka tą senąją reikšmę, o 0 – visas kitas reikšmes). \

Įtraukime į anksčiau nagrinėtą modelį lytį *gndr*.

```
xi: poisson   numbhh agea i.gndr
```

```
Poisson regression                                Number of obs   =        254
                                                  LR chi2(2)      =      98.91
                                                  Prob > chi2     =     0.0000
Log likelihood = -308.04404                       Pseudo R2       =     0.1383

------------------------------------------------------------------------------
      numbhh |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agea |   -.034575   .0033749   -10.24   0.000    -.0411896   -.0279604
    _Igndr_2 |   .0418111   .1237499     0.34   0.735    -.2007342    .2843565
       _cons |   2.033162   .1945051    10.45   0.000     1.651939    2.414385
------------------------------------------------------------------------------
```

Buvo sukurtas naujas dvireikšmis kintamasis *_Igdnr*_2, atitinkantis antrąją *gndr* kategoriją (moteris). Nors jis ir nėra statistiškai reikšmingas, kurį laiką į tai nekreipsime dėmesio ir pasiaiškinsime, kaip toks modelis traktuojamas. Realiai jis išskaidomas į dalinius modelius kiekvienai kategorinio regresoriaus reikšmei. Užrašome bendrąjį formalų modelį:

$\hat{\mu} = 2{,}03 - 0{,}034\, agea + 0{,}042\_Igndr\_2.$

Įstatę į jį *_Igndr_2* = 0, gauname modelį vyrams:

$\hat{\mu} = 2{,}00 - 0{,}034\, agea.$

Įstatę *_Igndr_2* = 1, gauname modelį moterims:

$\hat{\mu} = 2{,}072 - 0{,}034\, agea.$

Analogiškai sudaromi visi daliniai modeliai, kai kategorinis regresorius įgyja daugiau reikšmių.

Uždaviniai

1. 2008 metų Europos Sąjungos socialinio tyrimo Izraelio duomenų failas **ESS4IL.** Kintamieji:

- *agea* – respondento amžius,
- *freehms* – gėjai ir lesbietės gali gyventi, kaip nori (matuojamas skale nuo 1 – pilnai sutinku, iki 5 – visiškai nesutinku),

- *hhmmb* – namų ūkio pastovių narių skaičius,
- *blgetmg* – priklausymas etninėms mažumoms (1 – taip, 2 – ne),
- *gndr* – lytis (1 – vyr., 2 – mot.),
- *imptrad* – svarbu laikytis tradicijų ir papročių (1 – visai, kaip aš,...., 6 – absoliučiai ne taip, kaip aš) ,
- *pray* – meldžiamasi (1 – kasdien, 2 – dažniau, nei kartą per savaitę, 3 – kartą per savaitę, ..., 7 – niekada).

Pritaikyti Puasono regresijos modelį namų ūkio pastovių gyventojų skaičiui (*hhmmb*). Regresoriai amžius (*agea*) ir požiūris į seksualines mažumas (*freehms*). Tiriami labai tradicijų besilaikantys (*imptrad* = 1), religingi (*pray* ≤ 3) vyrai (*gndr* = 1), priklausantys nacionalinei daugumai (*blgtmg* = 2).

2. Ar tiks Puasono regresijos modelis, jeigu ankstesnio uždavinio sąlygoje pakeisime repondentus vyrus moterimis?

Puasono regresijos pavyzdys?

Tai - katastrofų modelis....

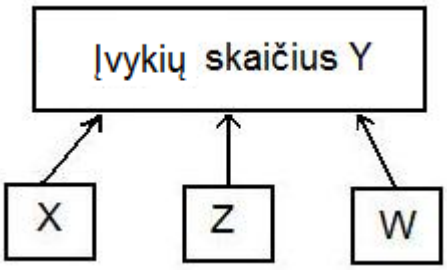
Tinka statistikos egzaminui

8. NEIGIAMA BINOMINĖ REGRESIJA

Tipinės situacijos, kai taikoma neigiama binominė regresinė analizė yra šios:

a) Norime ištirti, ar praleistų paskaitų skaičius priklauso nuo ankstenės sesijos statistikos egzamino pažymio.
b) Bandome išsiaiškinti, kaip peršalimų skaičius susijęs su maitinimosi įpročiais.
c) Modeliuojame loterijos laimėtojų įvairiuose regionuose skaičių, atsižvelgdami į gyventojų skaičių ir parduotų bilietų skaičių.
d) Tiriame nuo kokių draudžiamojo objekto savybių priklauso draudiminių įvykių skaičius.

Modelio schema:



Skaičiuojame kažkokius įvykius ir manome, kad tų įvykių skaičius (pažymėkime jį Y) nėra panašus į normalųjį kintamąjį ir priklauso nuo kitų stebimų kintamųjų X, Z, W. Kintamasis Y vadinamas *priklausomu* (arba regresuojamu) *kintamuoju*, kintamieji X, Z, W vadinami *nepriklausomais kintamaisiais* arba *regresoriais*. Faktiškai tokį pat modelį nagrinėjome Puasono regresijos atveju. Iš tikrųjų neigiama binominė regresija yra alternatyva Puasono regresijai, kai kintamojo Y dispersija daug didesnė už jo vidurkį (angl. overdispersion). Beveik viskas, kas pasakyta apie Puasono regresijos modelį tinka ir neigiamos binominės regresijos atveju. Pats modelio pavidalas ir prognozavimas atliekamas visiškai analogiškai. Vis dėlto, kad būtų patogiau naudotis šiais konspektais, pagrindines idėjas trumpai pakartosime.

8.1. Neigiamos binominės regresijos modelis

8.1.1. Modelio lygtis ir jos taikymas prognozavimui

Matematinis modelis sudaromas ne pačiam priklausomam kintamajam Y, o jo vidurkiui, kurį pažymėsime simboliu μ. Modelis atrodo taip:

$$\mu = \mathrm{e}^{z},$$

$$z = C + b_1 X + b_2 Z + b_3 W.$$

Naudojant imties duomenis įvertinami modelio koeficientai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$. Jų ženklai padeda suprasti regresorių poveikį *Y*.

Jeigu koeficientas teigiamas, tai regresoriui didėjant, vidutinė *Y* reikšmė didėja. Jeigu regresoriaus neigiamas, tai regresoriui didėjant, vidutinė *Y* reikšmė mažėja.

Į *modelio koeficientų eksponentes* atsižvelgiama, kai norima įvertinti *kiek kartų* padidės (sumažės) kintamojo *Y* vidurkis μ, atitinkamam kintamajam padidėjus vienetu.

Kai *X* padidėja vienetu, kintamojo *Y* vidurkis μ padidėja (sumažėja) $e^{\beta 1}$ karto.

Vidutinio įvykių skaičiaus prognozavimas konkrečioms *X, Z, W* reikšmėms atliekamas taip: į formulę $z = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W$ įstatome tas *X*, *Z* ir *W* reikšmes, kurioms norime gauti prognozę. Vidurkio μ įvertis gaunamas pagal pagal formulę: $\hat{\mu} = e^z$. Čia e = 2,718… Tarkime, kad konkrečioms *X, Z, W* reikšmėms apskaičiavome $z = 2$. Tada $\hat{\mu} = e^2 = 7{,}3890$ ...

Modeliuojant dažnį (įvykių skaičių per laiko vienetą), reikia duomenys turėti tyrimo trukmę rodantį kintamąjį (tarkime *s*). Pavyzdžiui, pateikiant kiekvieno traumatologinio punkto pacientų skaičiaus (*Y*) statistiką, dar reikia nurodyti, kelių dienų informaciją pateikiame (kintamasis *s*). Šiuo atveju naudojamas modifikuotas neigiamos binominės regresijos modelis:

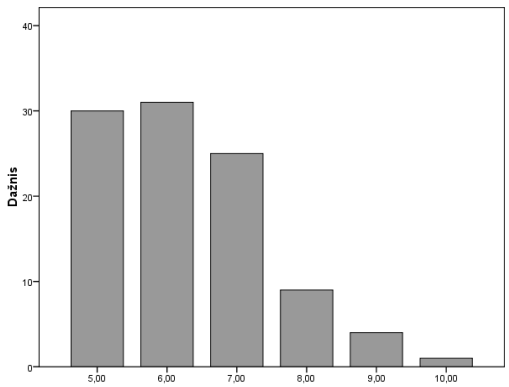
$$\ln(\mu/s) = C + b_1 X + b_2 Z + b_3 W.$$

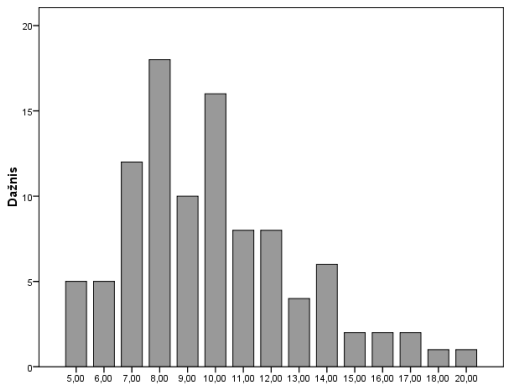
Konstantos *C*, b_1, b_2, b_3 nėra žinomos. Jų įverčiai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$ gaunami, panaudojus imties duomenis. Taikant šį modelį, reikia nepamiršti susikurti naują kintamąjį lygų kintamojo *s* logaritmui. Prognozavimas niekuo nesiskiria nuo ankstesnio atvejo.

8.1.2. Duomenys

Duomenys turi tenkinti tokius reikalavimus:

- Kintamasis *Y* turi neigiamą binominį skirstinį. Dispersija didesnė už vidurkį. Žemiau pateikiami neigiamų binominių skirstinių pavydžiai.
- Dalis regresorių gali būti intervaliniai, dalis dvireikšmiai arba kategoriniai (jie keičiami pseudokintamaisiais). Modelį lengviau interpretuoti, kai regresoriai intervaliniai.





8.1.3. Modelio tinkamumas

Modelio tinkamumą duomenims parodančios charakteristikos:

a) *Deviacija.* Tariama, kad modelis pakankamai geras, jeigu deviacija, padalinta iš savo laisvės laipsnių, nedaug skiriasi nuo vieneto.

b) *Pirsono chi kvadrato statistika* (Pearson Chi – Square) yra deviacijos analogas. Modelis tuo geresnis, kuo ši statistika mažesnė.

c) *Tikėtinumo santykio Chi kvadrato statistika* (angl. Likelihood Ratio Chi-Square). Parodo, ar modelyje yra bent vienas reikalingas regresorius. Jei*gu* p reikšmė didesnė už 0,05, tai regresijos modelio tinkamumas labai abejotinas.

d) *Akaikės iformacinis indeksas AIC.* Naudojamas, lyginant du modelius. Geresnis tas modelis, kurio AIC mažesnis. Primename, kad lyginamųjų modelių parametrų įverčiai turi būti suskaičiuoti, naudojant tuos pačius duomenis. Be to, vieno modelio regresoriai turi būti kito modelio regresorių aibės dalis.

e) *Voldo testai atskiriems regresoriams*. Jeigu $p < 0{,}05$, tai sakome, kad *regresorius yra statistiškai reikšmingas* ir dažniausiai jį modelyje paliekame. Jeigu $p \geq 0{,}05$, tai kintamasis yra statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais. Dažniausiai modelio konstanta C paliekama net ir tada, kai ji statistiškai nereikšminga.

8.1.4. Neigiamos binominės regresinės analizės etapai

Regresinė analizė susideda iš preliminaraus kintamųjų tinkamumo tyrimo ir modelio tinkamumo duomenims tikrinimo. Išsami analizė susideda iš tokių etapų:

1) Patikriname, ar Y dispersija didesnė už vidurkį. Jeigu mažesnė, tai neigiama binominė regresinė analizė netaikytina.

2) Sudarome neigiamos binominės regresijos modelį ir patikriname ar deviacija, padalinta iš savo laisvės laipsnių, nedaug skiriasi nuo vieneto. Alternatyviai galima tikrinti ar Pirsono chi kvadrato reikšmė padalinta iš laisvės laipsnių nedaug skiriasi nuo vieneto.
3) Patikriname ar tikėtinumo santykio chi kvadrato statistikos $p < 0,05$. Jeigu ne – modelis netinkamas.
4) Patikriname ar visi regresoriai statistiškai reikšmingi (visos Voldo kriterijaus $p < 0,05$). Jeigu ne – modelis taisytinas. Konstantai *p* reikšmės nežiūrime.
5) Jeigu viskas gerai – modelį aprašome. Jeigu ne – tobuliname.

Labai gerai duomenims tinkančiame modelyje:

- Deviacija (Pirsono chi kvadrato statistika) padalinta iš laisvės laipsnių nedaug skiriasi nuo vieneto.
- Didžiausio tikėtinumo chi kvadrato kriterijaus $p < 0,05$.
- Visų regresorių Voldo kriterijaus $p < 0,05$.

8.1.5. Modelio tobulinimas

Modelio tobulinimas reiškia kintamųjų šalinimą arba jų transformacijas.

Pašalinus bent vieną kintamąjį, keičiasi visos p reikšmės.

Modelis tobulinamas, jeigu yra problemų:

- *Normuotosios deviacijos ir jos laisvės laipsnių santykis* labai skiriasi nuo 1. Dažniausiai tai reiškia arba netinkamus regresorius arba, kad priklausomas kintamasis visiškai nepanašus į neigiamą binominį kintamąjį (pavyzdžiui, jo dispersija mažesnė už vidurkį). Bandome keisti modelį.
- *Yra statistiškai nereikšmingas kintamasis* (Voldo kriterijaus *p* reikšmė didesnė už 0,05). Kintamasis pašalinamas iš modelio ir neigiama binominė regresinė analizė pakartojama be šio kintamojo. Viskas gerai, jeigu deviacija, padalinta iš laisvės laipsnių, liko panaši į 1, o Akaikės informacinis indeksas AIC sumažėjo.
- *Per daug nulių* (angl. zero – inflated model). Kintamojo *Y* grafike nulio dažnis daug didesnis, nei turėtų įgyti neigiamai binominis atsitiktinis dydis. Kai kuriuose statistiniuose paketuose numatyta pataisyta neigiamai binominė regresinė analizė ir tokiai situacijai.

8.1.6. Pastabos apie modelio taikymą

- Neigiamas binominis skirstinys turi du parametrus *alpha* ir *p*. Parametras *p* yra tam tikra tikimybė ir visada įvertinamas, panaudojus imties duomenis. Dažnai statistinėse programose, skirtose neigiamai binominei regresijai, tariama, kad parametras *alpha* = 1. Tuo tarpu, jeigu kintamasis turi neigiamai binominį skirstinį su vidurkiu μ ir dispersija σ^2, tai $alpha = \mu^2/(\sigma^2 - \mu)$. O ši išraiška gali net labai skirtis nuo 1. Todėl vertėtų parametrą *alpha* pasirinkti patiems arba įvertinti jį naudojantis duomenimis (abu atvejai aptariami kituose skyreliuose). Beje, Lagranžo daugiklių testu, kuris realizuotas SPSS, galima patikrinti, ar parametras *alpha* nuo pasirinkto skaičiaus skiriasi statistiškai reikšmingai.
- Dar kartą primename, kad nors stebime sveikus skaičius įgyjantį dydį, prognozuojame vidutinę reikšmę. Vidurkis visai neturi būti sveikas skaičius.
- Reikia nemaišyti tikėtinumo santykio chi kvadrato (Likelihood Ratio Chi-Square) su Pirsono chi kvadratu (Pearson Chi-Square). Pastaroji statistika yra deviacijos analogas ir rodo geresnį modelio tikimą, kai yra *maža* (statistiškai nereikšminga). Tikėtinumo santykio chi kvadrato statistika lygina tiriamą modelį su neturinčiu jokių regresorių ir rodo geresnį modelio tikimą, kai yra *didelė* (statistiškai reikšminga).
- Nereikia suabsoliutinti normuotosios deviacijos ir kitų charakteristikų vaidmens. Nesunku sukonstruoti pavyzdį visai be regresorių, kada normuotoji deviacija maža, o konstanta C statistiškai reikšminga. Modelis turi būti dar ir logiškas.

8.1.7. Standartiniai neigiamos binominės regresijos modelio aprašymai

Visada aprašomas galutinis, po visų tobulinimų gautasis, modelis. Pateiksime keletą galimų aprašymų fragmentų.

- *Pritaikytas neigiamos binominės regresijos modelis gerai tiko duomenims, $\chi^2 = 12,27$; $p < 0,05$. Deviacijos santykis su laisvės laipsniais lygus 1,01. Visi modelio kintamieji statistiškai reikšmingi. Kintamojo Y vidurkio regresinis modelis atrodo taip:* $\ln \hat{\mu} = -2 + 0,46X + 5Z$. *Kai X padidėja 1, kintamojo Y vidurkis išauga 1,58 karto.....*
- *Iš modelio buvo pašalintas statistiškai nereikšmingas kintamasis X. Galutinis modelis buvo statistiškai reikšmingas, $\chi^2 = 10,02$; $p = 0, 023$ .....*
- *Taikėme neigiamos regresijos modelį, kuriame abudu parametrai buvo įvertinti didžiausio tikėtinumo metodu.....*

8.2. Neigiama binominė regresinė analizė su SPSS

8.2.1. Duomenys

Kaip atlikti neigiamai binominę regresinę analizę parodysime, tirdami 2008 metų Europos Sąjungos socialinio tyrimo Švedijos duomenis **ESS4SE**. Tyrime naudosime tokius kintamuosius:

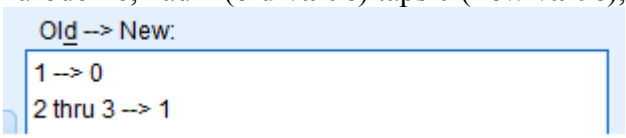
- *emplno* – turimų darbuotojų skaičius,
- *emplnof* – tėvo turimų darbuotojų skaičius (1 – neturi, 2 – turi 1 – 24 darbuotojus, 3 – turi 25 ir daugiau darbuotojų),
- *brmwmny* – pasiskolinti pinigų išgyvenimui (1 – labai sunku, ..., 5 – labai lengva) ,
- *eduyrs* – prasimokytų metų skaičius.

Modeliuosime *emplno* priklausomybę nuo *emplnof, brwmny, eduyrs*.Visų pirma įsitikiname, ar visi kintamieji turi pakankamai daug reikšmių. Atsižvelgdami į tai, kad modeliuosime *emplno*, dažnių lenteles tiriame tik preliminariai **Select Cases** pasirinkę *emplno* ≥ 0. Ši sąlyga užtikrina, kad gausime informaciją tik apie respondentus, atsakiusius į klausimą apie turimą darbuotojų skaičių. Renkamės **Analyze → Descriptive Statistics → Frequences.** Į langą **Variable(s)** įkeliame kintamuosius *emplno, emplnof, brwmny, eduyrs* ir spaudžiame **O.K.** Visi kintamieji turi pakankamai daug įvairių reikšmių, išskyrus kintamąjį *emplnof.* Yra tik vienas respondentas, kurio tėvo darbuotojų skaičius buvo ne mažesnis už 25.

emplnof Number of employees father had

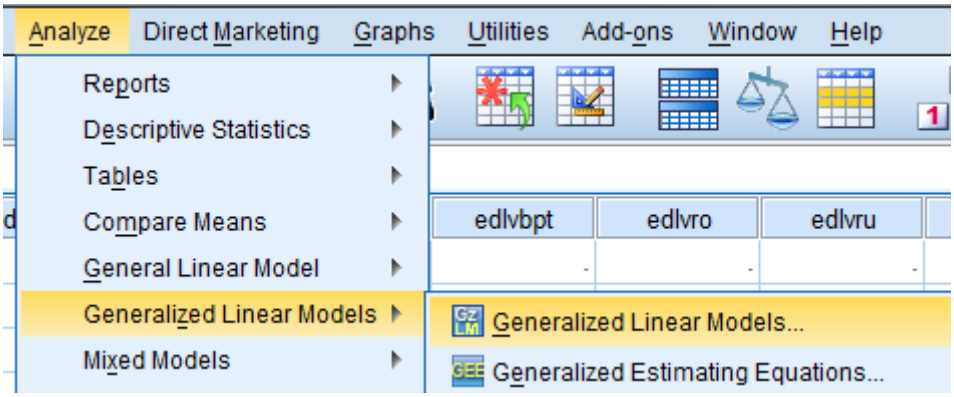
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 None	34	19.3	50.7	50.7
	2 1 to 24	32	18.2	47.8	98.5
	3 25 or more	1	.6	1.5	100.0
	Total	67	38.1	100.0	

Pakeisime kintamąjį *emplnof* nauju kintamuoju *emplnof2*, kuris įgis tik dvi reikšmes: 0 – kai respondento tėvas darbuotojų neturėjo ir 1 – kai respondento tėvas turėjo bent vieną darbuotoją. Renkamės **Transform → Recode into Different Variables**. Į langą perkeliame kintamąjį *emplnof* . Langelyje **Name** įrašome *emplnof2* ir paspaudžiame **Change.** Paspaudę **Old and New Values** nurodome, kad 1 (**old value**) taps 0 (**new value**), o 2 bei 3 taps 1. Pasirenkame **Continue** ir **O.K.**

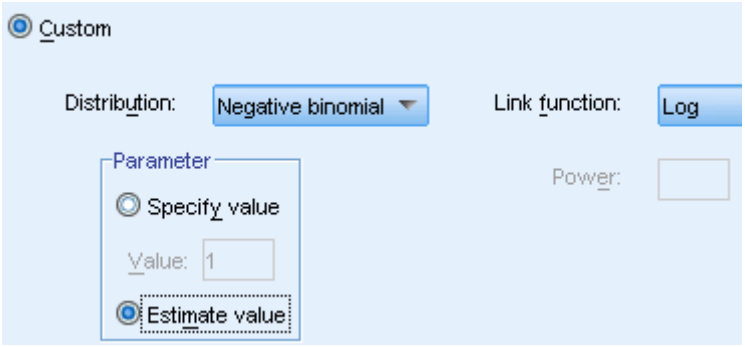


8.2.2. SPSS parinktys neigiamos binominės regresijos modeliui

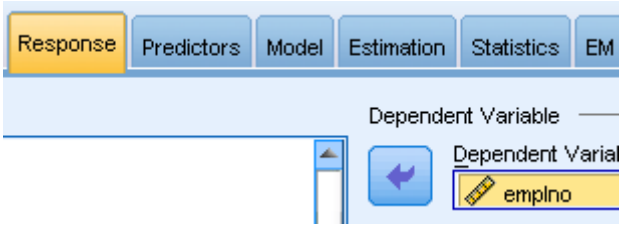
Patikrinsime, kaip atrinktiesiems respondentams tinka neigiama binominė regresija. **Analyze → Generalized Linear Models → Generalized Linear Models.**



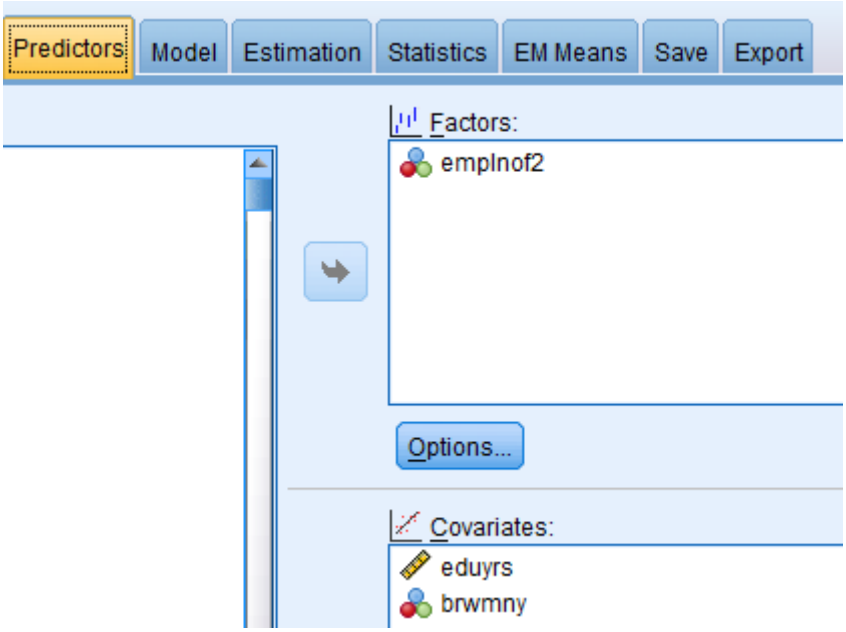
Atsidariusiame meniu spaudžiame **Type of Model.** Nors atsidariusiame meniu yra opcija **Negative binomial with log link**, tos opcijos **nesirenkame** (kodėl pakomentuota viename iš skyrelių žemiau). Pažymime **Custom.** Parinktyse **Distribution** pasirenkame **Negative binomial, Link function – Log , Parameter – Estimate value.**



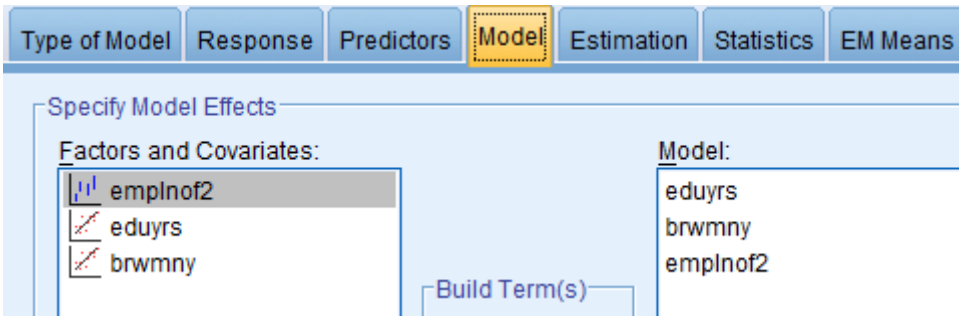
Po to spaudžiame **Response** ir perkeliame *emplno* į laukelį **Dependent Variable**.



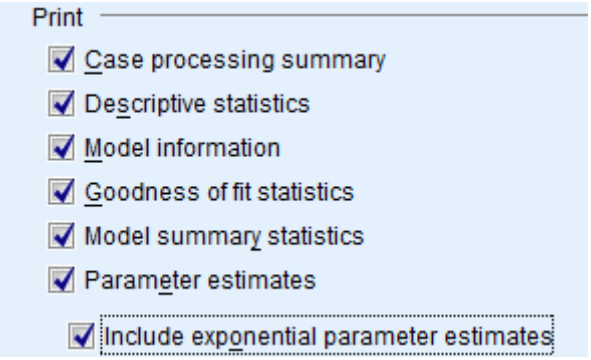
Paspaudę **Predictors** abudu kintamuosius *eduyrs* ir *brwmny* perkeliame į langelį **Covariates**. Kategorinį kintamąjį *emplnof2* keliame į **Factors**.



Pasirinkę **Model**, visus kintamuosius perkeliame į langą **Model.**



Pasirinkę **Statistics** papildomai pažymime **Include exponential parameter estimates.** Jau pažymėtus langelius paliekame ramybėje. Spaudžiame **OK**.



8.2.3. Rezultatų aptarimas

Rezultatų išklotinė prasideda nuo informacijos apie kintamuosius. Matome, kad pusės respondentų tėvai darbuotojų neturėjo. Atkreipiame dėmesį, kad, *emplno* standartinis nuokrypis (o juo labiau dispersija) daug didesnis už vidurkį. Todėl paprastas Puasono regresijos modelis šiems duomenims netiktų.

Categorical Variable Information

			N	Percent
Factor	emplnof2	.00	33	50.0%
		1.00	33	50.0%
		Total	66	100.0%

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	emplno Number of employees respondent has/had	66	0	763	14.73	93.831
Covariate	eduyrs Years of full-time education completed	66	5	23	11.71	3.732
	brwmny Borrow money to make ends meet, difficult or easy	66	1	5	3.68	1.069

Lentelėje **Goodness of Fit** yra surašytos deviacijos reikšmė, Pirsono chi kvadrato statistika ir t.t. Deviacijos ir jos laisvės laipsnių santykis tinkamo neigiamos binominės regresijos atveju turi būti artimas vienetui. Matome, kad šis santykis yra 0,901. Pirsono chi kvadrato statistika, padalinta iš savo laisvės laipsnių, irgi turi būti arti vieneto. Šis santykis (0,968) irgi labai artimas vienetui. Gavome patvirtinimą, kad apskritai modelis tinka duomenims. Informacinių kriterijų reikšmės (AIC, BIC ir pan.) naudojamos kelių modelių palyginimui. Jeigu papildytume modelį naujais kintamaisiais (pašalintume dalį kintamųjų) ir lygintume su anksčiau buvusiu, tai geresnis būtų tas modelis, kurio informacinių kriterijų reikšmės mažesnės.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	54.989	61	.901
Scaled Deviance	54.989	61	
Pearson Chi-Square	59.032	61	.968
Scaled Pearson Chi-Square	59.032	61	
Log Likelihood[a]	-133.952		
Akaike's Information Criterion (AIC)	277.905		
Finite Sample Corrected AIC (AICC)	278.905		
Bayesian Information Criterion (BIC)	288.853		
Consistent AIC (CAIC)	293.853		

Dependent Variable: Number of employees respondent has/had

Model: (Intercept), emplnof2, eduyrs, brwmny

Lentelėje **Omnibus Test** yra didžiausio tikėtinumo santykio chi kvadrato statistika ir jos *p* reikšmė. Gerai, kai $p < 0{,}05$, nes tada modelyje yra bent vienas reikalingas (statistiškai reikšmingas) regresorius. Kadangi $p = 0{,}000...$, tai galime teigti, kad neigiamos binominės regresijos modelio tinkamumą patvirtina ir didžiausio tikėtinumo chi kvadrato kriterijus.

Lentelėje **Tests of Model Effects** yra informacija apie Voldo kriterijų. Kintamieji *emplnof2, eduyrs, brwmny* yra statistiškai reikšmingi (jų *p* reikšmės $< 0{,}05$). Konstantai *p* reikšmės nežiūrime, nes jos iš modelio nežadame šalinti. Taigi visi regresoriai yra statistiškai reikšmingi ir modelyje reikalingi.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.151	1	.698
emplnof2	6.298	1	.012
eduyrs	4.959	1	.026
brwmny	7.399	1	.007

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
23.777	3	.000

Lentelėje **Parameter Estimates** pasikartoja informacija apie Voldo kriterijų. Taip pat surašyti parametrų įverčiai. Parametrų įverčių ženklai ir **Exp(B)** suteikia informaciją apie tai, kaip keisis vidutinis darbuotojų skaičius *emplno*, keičiantis kintamiesiems *emplnof2, eduyrs, brwmny.*

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	1.590	2.1316	-2.588	5.768	.556	1	.456	4.904	.075	319.831
[emplnof2=.00]	-1.629	.6493	-2.902	-.357	6.298	1	.012	.196	.055	.700
[emplnof2=1.00]	0[a]	.	.	.	.	.	.	1	.	.
eduyrs	.286	.1286	.034	.539	4.959	1	.026	1.332	1.035	1.714
brwmny	-.753	.2768	-1.295	-.210	7.399	1	.007	.471	.274	.810
(Scale)	1[b]									
(Negative binomial)	5.327	1.2084	3.415	8.310						

Kintamojo *eduyrs* koeficientas lygus 0,286 > 0. Tai reiškia, kad didėjant atsakinėjusio respondento išsilavinimui, vidutinis turimų darbuotojų skaičius didėja. Stulpelyje **Exp(B)** užrašyta reikšmė 1,332 rodo, kad vieneri papildomi respondento prasimokyti metai padidina tikėtiną turimų darbuotojų skaičių 1,332 **karto**. Stulpeliuose **Lower** ir **Upper** nurodyti pačio koeficiento ir **Exp(B)** pasikliautinųjų intervalų rėžiai. Analogiškai galima interpretuoti kintamojo *brwmny* įtaką. Kategorinis kintamasis *emplnof2* aprašytas dviem lygtimis. Matome, kad kad tėvo darbuotojų nebuvimas vidutinį respondento darbuotojų skaičių veikia neigiamai, nes modelio konstanta sumažėja per - 1.629.

Matematinis modelio užrašas atrodo taip:

$$\ln \hat{\mu} = 1{,}590 + 0{,}286\, eduyrs - 0{,}753\, brwmny + \begin{cases} 0, & \text{jei } emplnof2 = 1 \\ -1{,}629, & \text{jei } emplnof2 = 0 \end{cases}.$$

Čia $\hat{\mu}$ žymi vidutinį respondento turimų darbuotojų skaičių. Parodysime, kaip šį modelį taikyti konkrečioms prognozėms. Tarkime, kad respondentas mokėsi 10 metų, mano, kad pasiskolinti pinigų labai lengva (*brwmny* = 5), o jo tėvas dabuotojų neturėjo (*emplnof2* = 0). Kiek vidutiniškai tokie respondentai turi dabuotojų? Įstatę regresorių reikšmes į aukščiau pateiktą lygtį, gauname

$$\ln \hat{\mu} = 1{,}590 + 0{,}286 \cdot 10 - 0{,}753 \cdot 5 - 1{,}629 = -0{,}944.$$

Pačią narių skaičiaus prognozę gauname, pakėlę gautoju skaičiumi e = 2,718.... Todėl

$$\hat{\mu} = e^{-0{,}944} = 0{,}389.$$

Darome prognozę, kad menkai išsilavinę optimistai, kurių tėvai darbuotojų neturėjo, ir patys nelabai jų turės (vidutiniškai 0,389 darbuotojo).

8.2.4. Rezultatų aprašymas

Pradžioje aprašome, ką tyrėme, kokius kintamuosius naudojome ir kaip juos perkodavome. Tada kompaktiškai surašome visa tai, ką išsamiai aptarėme ankstesniame skyrelyje. Pateikiame vieną iš daugybės galimų tokio aprašymo variantų.

Neigiamos binominės regresijos modelis gerai tiko duomenims, didžiausio tikėtinumo santykio chi kvadrato statistika $\chi^2 = 23{,}777$, $p < 0{,}01$. Deviacijos ir jos laisvės laipsnių santykis yra 0,901. Voldo kriterijus parodė, kad visi modelio regresoriai, išskyrus konstantą, yra statistiškai reikšmingi. Modelio matematinė išraiška

$$\ln \hat{\mu} = 1{,}590 + 0{,}286\, eduyrs - 0{,}753\, brwmny + \begin{cases} 0, & \text{jei } emplnof2 = 1 \\ -1{,}629, & \text{jei } emplnof2 = 0 \end{cases}$$

rodo, kad vidutinis turimų darbuotojų skaičius didėja, jeigu augant respondento išsilavinimui ir esant didesniam pesimizmui, vertinant paskolos galimybes. Didesnį darbuotojų skaičių turi tie respondentai, kurių tėvai irgi turėjo darbuotojų. Koeficientų ekponenčių analizė parodė, kad

vieneriais metais padidėjus respondento prasimokytų metų skaičiui, vidutinis darbuotojų skaičius išauga 1,332 karto (95% pasikliautinis intervalas [1,035; 1,714]). Vienu balu didesnis optimizmas, vertinant paskolos galimybes, sumažina galimą respondento turimų darbuotojų skaičių 0,471 karto (95% pasikliautinis intervalas [0,274; 0,810]). Respondentai, kurių tėvai neturėjo darbuotojų, vidutiniškai turės 0,196 karto mažiau darbuotojų, nei nei respondentai, kurių tėvai darbuotojų turėjo. Taigi praktiškai penkis kartus mažiau. Vis dėlto šio rodiklio 95% pasikliautinis intervalas [0,055; 0,700] yra pakankamai didelis. Todėl teisingiau būtų konstatuoti, kad tikėtina, jog tokių respondentų turimas darbuotojų skaičius sudarys nuo 5 iki 70 procentų darbuotojų skaičiaus tų respondentų, kurių tėvai irgi turėjo darbuotojų.

8.2.5. SPSS komandų sintaksė

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam regresijos pavyzdžiui. Kintamojo *emplnof2* sukūrimas:

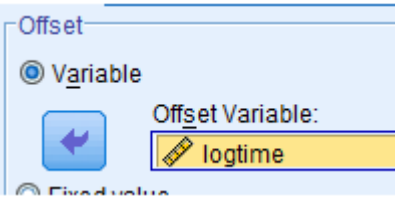
```
RECODE emplnof (1=0) (2 thru 3=1) INTO emplnof2.
EXECUTE.
```

Neigiamos binominės regresijos sintaksė:

```
GENLIN emplno BY emplnof2 (ORDER=ASCENDING) WITH eduyrs brwmny
  /MODEL emplnof2 eduyrs brwmny INTERCEPT=YES
 DISTRIBUTION=NEGBIN(MLE) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL
MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012
 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO
FIT SUMMARY SOLUTION (EXPONENTIATED).
```

8.2.6. Neigiama binominė regresinė analizė dažniui

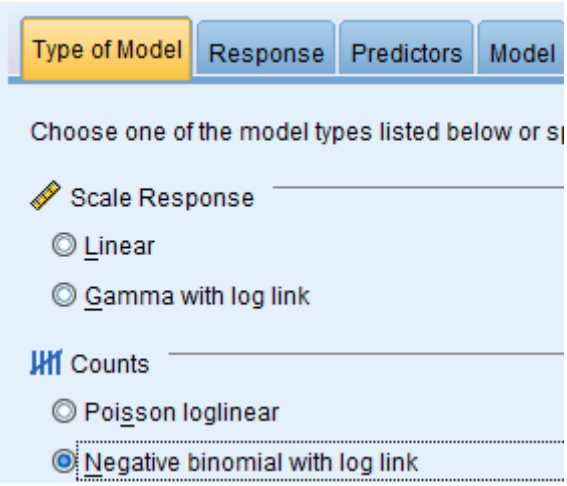
Pateiksime teorinį pavyzdį, parodantį, kas tada keisis tyrime. Tarkime, kad turime informaciją apie respondento patirtų epilepsijos priepuolių skaičių (*epnumb*). Kiekvienas respondentas stebėtas skirtingą laiko tarpą (kintamasis *time*). Aišku, kad šiuo atveju natūraliau modeliuoti priepuolių dažnumą, o ne patį jų skaičių. Komanda **Compute Variable** sukuriame kintamąjį *lotime* lygų *time* logaritmui. Po to renkamės visas tas pačias opcijas, kaip ir standartinėje neigiamoje binominėje regresijoje, tiktai **Predictors** lange į lauką **Offset Variable** įkeliame *logtime*.



Visi rezultatai interpretuojami, kaip ir ankstesniu atveju. Tiktai dabar modelis sudarytas dažnio *epnumb/time* logaritmui, o ne kintamojo *epnumb* logaritmui.

8.2.7. Neigiamos binominės regresinės analizės standartinė opcija SPSS

SPSS parinkčių lange **Type of Model** yra nurodyta neigiama binominė regresija **Negative binomial with log link**:



Kuo ši parinktis skiriasi nuo aukščiau naudotos? Neigiamas binominis skirstinys turi du parametrus. Mes juos ir išnaudojome, pasistengdami, kad modelio vidurkis ir dispersija sutaptų su modeliuojamojo *emplno* imties vidurkiu ir dispersija. Tuo tarpu **Negative binomial with log link** parinktis numato, kad vienas iš skirstinio savybes lemiančių parametrų bus lygus vienetui. Todėl negalima tikėtis, kad modelio dispersija sutaps su *emplno* dispersija. Pasekmė – blogas modelio

tikimas. Iš tikro, pritaikę nagrinėtajam modeliui **Negative binomial with log link** parinktį, gauname, kad deviacijos ir laisvės laipsnių santykis yra beveik trys (2,980), o Pirsono chi kvadrato ir laisvės laipsnių santykis yra 4. Tai rodo labai blogą modelio tikimą duomenims.

Goodness of Fit[b]			
	Value	df	Value/df
Deviance	184.745	62	2.980
Scaled Deviance	184.745	62	
Pearson Chi-Square	248.708	62	4.011

8.3. Neigiama binominė regresinė analizė su STATA

8.3.1. Duomenys

Atliksime tą pačią analizę, kaip ir su SPSS programa. Tiriame Švedijos duomenis **ESS4SE**. Modeliuosime *emplno* (turimų darbuotojų skaičius) priklausomybę nuo regresorių

- *brwmny* (ar sunku pasiskolinti pinigų išgyvenimui, 1 – labai sunku, ..., 5 – labai lengva),
- *eduyrs* (prasimokytų metų skaičius),
- *emplnof2* (tėvo turimų darbuotojų skaičius, 0 – neturi, 1 – turi).

SPSS duomenis išsaugojome *dta* formatu. Atidarius failą su STATA dar reikia užtikrinti, kad kodai, kurie SPSS žymėjo praleistas reikšmes, nebūtų traktuojami, kaip regresorių reikšmės. Todėl įvykdėme:

```
keep if emplno < 900 & emplnof < 4 & eduyrs < 88 &brwmny < 8
```

Kintamasis *emplnof2* gautas tinkamai perkodavus *emplnof* (recode emplnof (1 = 0) (2/3 = 1), gen(emplnof2)). Pradžioje įsitikiname, kad kintamojo *emplno* dispersija daug didesnė už vidurkį. Įvykdome:

```
summarize emplno, detail
```

```
50%          0                    Mean          14.72727
                      Largest     Std. Dev.     93.83071
75%          2            24
90%         15            30      Variance      8804.201
95%         24            42      Skewness      7.859733
99%        763           763      Kurtosis      63.18019
```

Dispersija lygi 8804, o vidurkis tik 14,727. Šiuo atveju Puasono regresija visiškai netinka, o neigiama binominė regresija gali ir tikti.

8.3.2. Modelio parametrų analizė

Regresorius *emplnof2* yra kategorinis, bet įgyja tik dvi reikšmes (0 ir 1). Todėl jokių specialių komandų nereikia. Neigiama binominė regresija atliekama, įvykdant :

```
nbreg emplno brwmny eduyrs emplnof2
```

```
Negative binomial regression                      Number of obs   =         66
                                                  LR chi2(3)      =      23.78
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -133.95237                       Pseudo R2       =     0.0815

------------------------------------------------------------------------------
      emplno |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      brwmny |   -.752886   .2767894    -2.72   0.007    -1.295383   -.2103887
      eduyrs |   .2864844   .1286464     2.23   0.026     .0343422    .5386267
    emplnof2 |   1.629398   .6492644     2.51   0.012     .3568631    2.901933
       _cons |  -.0394117   1.910606    -0.02   0.984    -3.784131    3.705307
-------------+----------------------------------------------------------------
    /lnalpha |    1.67284   .2268809                      1.228161    2.117518
-------------+----------------------------------------------------------------
       alpha |   5.327274   1.208657                      3.414944    8.310486
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 2444.86 Prob>=chibar2 = 0.000
```

Lentelės viršuje dešinėje yra didžiausio tikėtinumo santykio chi kvadrato statistika ir jos *p* reikšmė. Kadangi $p = 0{,}000 < 0{,}05$, tai galime teigti, kad neigiamos binominės regresijos modelio tinkamumą didžiausio tikėtinumo chi kvadrato kriterijus patvirtina. Makfadeno pseudo $R^2 = 0{,}08$ yra mažas (< 0,20), bet neigiamai binominei regresijai tai nėra svarbus rodiklis. Aprašant modelio rezultatus, prie R^2 dažniausiai net neužsimenama.

Po lentele yra informacija, apie vieną neigiamos binominės regresijos modelio parametrą. Kai šis parametras lygus nuliui, tai neigiama binominė regresija virsta Puasono regresija. STATA programoje jis žymimas *alpha.* Matome, kad *alpha* = 5,327 ir nuo nulio skiriasi statistiškai reikšmingai (Prob >= chibar2 yra didžiausio tikėtinumo santykio testo *p* reikšmė. Ji lygi 0,000.. < 0,05). Gavome statistiškai reikšmingą patvirtinimą, kad Puasono regresija nagrinėjamam pavyzdžiui netinka.

Pačioje lentelėje pateikta informacija apie Voldo kriterijų. Regresoriai *emplnof2, eduyrs, brwmny* yra statistiškai reikšmingi (jų *p* reikšmės < 0,05). Konstantai *p* reikšmės nežiūrime, nes jos iš modelio nežadame šalinti. Taigi visi regresoriai yra statistiškai reikšmingi ir modelyje reikalingi. Modelį galima užrašyti taip:

Vidutinis darbuotojų skaičius = exp{-0,04 – 0,75 *brwmny* + 0,29 *eduyrs* + 1,63 *emplnof2*}.

Regresorių ženklai atitinka logiką: neigiamas *brwmny* koeficientas reiškia, kad respondentas labiau įsitikinęs paskolos gavimo lengvumu turi mažiau darbuotojų. Analogiškai darome išvadą, kad daugiau metų prasimokę respondentai turi ir daugiau darbuotojų. Daugiau darbuotojų turi ir žmonės, kurių tėvai jų irgi turi. Norėdami skaitiškai įvertinti regresorių įtaką, įvykdome:

```
listcoef, help
```

```
nbreg (N=66): Factor Change in Expected Count

 Observed SD: 93.830706

------------------------------------------------------------------------------
      emplno |      b         z     P>|z|     e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------------
      eduyrs |   0.28648    2.227   0.026   1.3317   2.9133     3.7325
      brwmny |  -0.75289   -2.720   0.007   0.4710   0.4471     1.0693
    emplnof2 |   1.62940    2.510   0.012   5.1008   2.2726     0.5038
-------------+----------------------------------------------------------------
    ln alpha |   1.67284
       alpha |   5.32727   SE(alpha) = 1.20866
------------------------------------------------------------------------------
 LR test of alpha=0: 2444.86  Prob>=LRX2 = 0.000
------------------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in expected count for unit increase in X
 e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
   SDofX = standard deviation of X
```

Lentelėje pakartoti koeficientų įverčiai ir jų eksponentės – t.y. daugikliai, rodantys kiek kartų padidės (sumažės) modeliuojamo kintamojo vidurkis, regresoriui padidėjus vienetu. Pavyzdžiui, matome, kad respondentai, kurių tėvai turi samdomų darbuotojų, vidutiniškai ir patys jų turi 5,1 karto daugiau, lyginant su respondentais, kurių tėvai samdomų darbuotojų neturi.Norėdami gauti analogišką informaciją, surašytą procentais, įvykdome:

```
listcoef, help percent
```

```
------------------------------------------------------------------------------
      emplno |      b         z     P>|z|       %      %StdX      SDofX
-------------+----------------------------------------------------------------
      brwmny |  -0.75289   -2.720   0.007    -52.9    -55.3     1.0693
      eduyrs |   0.28648    2.227   0.026     33.2    191.3     3.7325
    emplnof2 |   1.62940    2.510   0.012    410.1    127.3     0.5038
-------------+----------------------------------------------------------------
    ln alpha |   1.67284
       alpha |   5.32727   SE(alpha) = 1.20866
------------------------------------------------------------------------------
 LR test of alpha=0: 2444.86  Prob>=LRX2 = 0.000
```

Iš tikro, jeigu tėvo samdomų darbuotojų turėjimas padidina galimą vidutinį respondento darbuotojų skaičių 5,1 karto, tai darbuotojų skaičius padidėja 410 procentų. Tai ir užfiksuota gautojoje lentelėje.

Nors ir visi regresoriai statistiškai reikšmingi, ir modelio didžiausio tikėtinumo chi kvadrato statistikos *p* reikšmė labai maža, dar reikėtų pažvelgti į modelio deviaciją. Ją gausime paeiliui įvykdę:

```
local v = e(alpha)
glm emplno brwmny eduyrs emplnof2, family(nb `v') nolog
```

```
Generalized linear models                          No. of obs      =        66
Optimization     : ML                              Residual df     =        62
                                                   Scale parameter =         1
Deviance         =  54.98876078                    (1/df) Deviance =  .8869155
Pearson          =  59.03185108                    (1/df) Pearson  =  .9521266

Variance function: V(u) = u+(5.327273896968735)u^2 [Neg. Binomial]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  4.180375
Log likelihood   = -133.9523705                    BIC             = -204.7698
```

Ieškomoji modelio deviacija yra 54,988. Deviacia padalinta iš laisvės laipsnių lygi 0,887. Nėra taip toli nuo vieneto. Dar arčiau vieneto yra liekamųjų paklaidų Pirsono chi kvadrato ir laisvės laipsnių santykis 0,952. Darome išvadą, kad neigiamos binominės regresijos modelis duomenims gerai tinka.

8.3.3. Prognozavimas

Prognozavimas niekuo nesiskiria nuo Puasono regresijos atvejo. Parodysime, kaip atlikti vidutinio darbuotojų skaičiaus prognozę respondentui, prasimokusiam 15 metų ir manančiam, kad pinigų pasiskolinti labai sunku. Be to žinome, kad jo tėvas turi samdomų darbuotojų. (Jeigu tikrinome deviaciją su glm, tai vėl įvykdome nbreg emplno emplnof2 eduyrs brwmny). Įvykdome :

```
adjust eduyrs=15 brwmny = 1 emplnof2 = 1, exp
```

All	exp(xb)
	169.757

Key: exp(xb) = exp(xb)

Gavome, kad tokie respondentai vidutiniškai turėtų 169,75 darbuotoją. Galima prognozę išskaidyti į dvi, atsižvelgiant į tėvo darbuotojus: respondentai, kurių tėvai samdomų darbuotojų neturi turi vidutiniškai 33,28 darbuotojus.

```
adjust  eduyrs=15 brwmny = 1, by(emplnof2) exp
```

```
had)        |    exp(xb)
------------+-----------
         0  |    33.2804
         1  |    169.757
------------+-----------
     Key:  exp(xb)  =  exp(xb)
```

Galima prognozuoti kiekvieno darbuotojų skaičiaus tikimybes. Tarkime, kad tiriame respondentus, kurie apie paskolų gavimą turi optimistinę nuomonę: *brwmny* = 4. Įvykdome:

```
prvalue , x(brwmny = 4, emplnof2 = 0)
```

```
                                  95% Conf. Interval
  Rate:                 1.3558   [ .09312,    2.6185]
  Pr(y=0|x):            0.6733   [ 0.5699,    0.7767]
  Pr(y=1|x):            0.1110   [ 0.1066,    0.1155]
  Pr(y=2|x):            0.0579   [ 0.0537,    0.0621]
  Pr(y=3|x):            0.0371   [ 0.0302,    0.0440]
  Pr(y=4|x):            0.0260   [ 0.0182,    0.0337]
  Pr(y=5|x):            0.0191   [ 0.0112,    0.0270]
  Pr(y=6|x):            0.0145   [ 0.0069,    0.0221]
  Pr(y=7|x):            0.0113   [ 0.0041,    0.0185]
  Pr(y=8|x):            0.0089   [ 0.0022,    0.0156]
  Pr(y=9|x):            0.0071   [ 0.0010,    0.0133]

       brwmny     eduyrs   emplnof2
x=          4  11.712121          0
```

STATA pateikia pirmųjų dešimties reikšmių tikimybes. Darydami prognozę, nurodėme tik dviejų regresorių reikšmes. Regresoriaus *eduyrs* reikšmės nenurodėme. Darant prognozę, į modelį automatiškai buvo įtrauktas jos vidurkis *eduyrs* = 11,71.

8.3.4. Kategorinis kintamasis

Kategorinis kintamasis automatiškai perkoduojamas į keletą dvireikšmių kintamųjų. Išsamiai tokį atvejį aprašėme, nagrinėdami Puasono regresiją. Neigiamos binominės regresijos atveju viskas atliekama analogiškai. Todėl apsiribosime tik vienu pavyzdžiu, kai *emplnof2* aprašysime, nurodydami jo kategorinę skalę. Įvykdome:

```
xi: nbreg emplno brwmny eduyrs i.emplnof2
```

Regresorius *emplnof2* buvo pakeistas į *_Iemplnof2_1* , kuris faktiškai sutapo su pačiu *emplnof2*, nes šis ir taip buvo dvireikšmis. Nenuostabu, kad ir koeficientų įverčiai, lyginant su jau

daryta regresija, nepasikeitė. Jeigu *emplnof2* būtų trikategorinis kintamasis, jis būtų pakeistas dviem dvireikšmiais regresoriais. Jų reikšmės(0,0), (1,0) ir (0,1) atitiktų tas tris pradines kintamojo reikšmes.

emplno	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
brwmny	-.752886	.2767894	-2.72	0.007	-1.295383	-.2103887
eduyrs	.2864844	.1286464	2.23	0.026	.0343422	.5386267
_Iemplnof2_1	1.629398	.6492644	2.51	0.012	.3568631	2.901933
_cons	-.0394117	1.910606	-0.02	0.984	-3.784131	3.705307
/lnalpha	1.67284	.2268809			1.228161	2.117518
alpha	5.327274	1.208657			3.414944	8.310486

Likelihood-ratio test of alpha=0: chibar2(01) = 2444.86 Prob>=chibar2 = 0.000

8.3.5. GLM alternatyva

Norėdami gauti deviaciją, pasinaudojome glm komanda. Kaip viena iš opcijų, joje numatyta ir neigiama binominė regresija. Deja, regresijos parametras parenkamas lygiu vienetui (*alpha* = 1). Pasekmė – blogai tinkantis modelis.

```
glm:  emplno brwmny eduyrs emplnof2, family(nb)
```

```
Generalized linear models                          No. of obs      =        66
Optimization     : ML                              Residual df     =        62
                                                   Scale parameter =         1
Deviance         =  184.7447012                    (1/df) Deviance =  2.979753
Pearson          =  248.7084221                    (1/df) Pearson  =  4.011426

Variance function: V(u) = u+(1)u^2                 [Neg. Binomial]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  5.282873
Log likelihood   = -170.3348194                    BIC             = -75.01389
```

emplno	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
brwmny	-.7773024	.1339439	-5.80	0.000	-1.039828	-.5147772
eduyrs	.2753865	.0577262	4.77	0.000	.1622452	.3885278
emplnof2	1.731883	.3054939	5.67	0.000	1.133126	2.33064
_cons	.1134292	.8967958	0.13	0.899	-1.644258	1.871117

Deviacijos ir laisvės laipsnių santykis labai viršija vienetą (2.97). Dar didesnis yra Pirsono liekamųjų paklaidų chi kvadrato statistikos ir laisvės laipsnių santykis – 4,01. Situacija aiškiai bloga ir analogiška aptartajai darbo su SPSS septintajame skyrelyje.

Uždaviniai

1. Tiriame 2008 metų Prancūzijos duomenis **ESS4FR**. Tyrime naudosime tokius kintamuosius: *agea* – respondento amžius, *hhmmb* – namų ūkio pastovių narių skaičius, *cldcrsv* – paramos vaiko priežiūrai teikimas tėvams, turintiems teisę ją gauti (0 – itin blogas,....., 10 – labai geras), *imptrad* – svarbu laikytis tradicijų ir papročių (1 – visai, kaip aš,...., 6 – absoliučiai ne taip, kaip aš) , *eduyrs* – prasimokytų metų skaičius. Tirsime stipriai tradicijų besilaikančius (*imptrad* $\leq$ 2) aukštojo mokslo neragavusius (*eduyrs* $\leq$ 10) respondentus. Modeliuosime kiek namų ūkyje yra gyventojų: *hhmmb = f(cldcrsv, agea).* Pritaikyti neigiamą binominę regresiją. Ar rezultatai skiriasi nuo gautųjų, aptartų Puasono regresijai skirtame skyriuje? Kaip atrodys neigiama binominė regresija, atlikta su standartinėmis *glm* opcijomis?

2. Tiriami2008 metų Europos Sąjungos socialinio tyrimo Izraelio duomenų failas **ESS4IL**. Kintamieji:

- *agea* – respondento amžius,
- *freehms* – gėjai ir lesbietės gali gyventi, kaip nori (matuojamas skale nuo 1 – pilnai sutinku, iki 5 – visiškai nesutinku),
- *hhmmb* – namų ūkio pastovių narių skaičius,
- *blgetmg* – priklausymas etninėms mažumoms (1 – taip, 2 – ne),
- *gndr* – lytis (1 – vyr., 2 – mot.),
- *imptrad* – svarbu laikytis tradicijų ir papročių (1 – visai, kaip aš,...., 6 – absoliučiai ne taip, kaip aš) ,
- *pray* – meldžiamasi (1 – kasdien, 2 – dažniau, nei kartą per savaitę, 3 – kartą per savaitę, ..., 7 – niekada).

Pritaikyti neigiamos regresijos modelį namų ūkio pastovių gyventojų skaičiui (*hhmmb*). Regresoriai amžius (*agea*) ir požiūris į seksualines mažumas (*freehms*). Tiriami labai tradicijų besilaikantys (*imptrad* = 1), religingi (*pray* $\leq$ 3) vyrai (*gndr* = 1), priklausantys nacionalinei daugumai (*blgtmg* = 2).

9. MODIFIKUOTI SKAIČIUOJAMIEJI MODELIAI

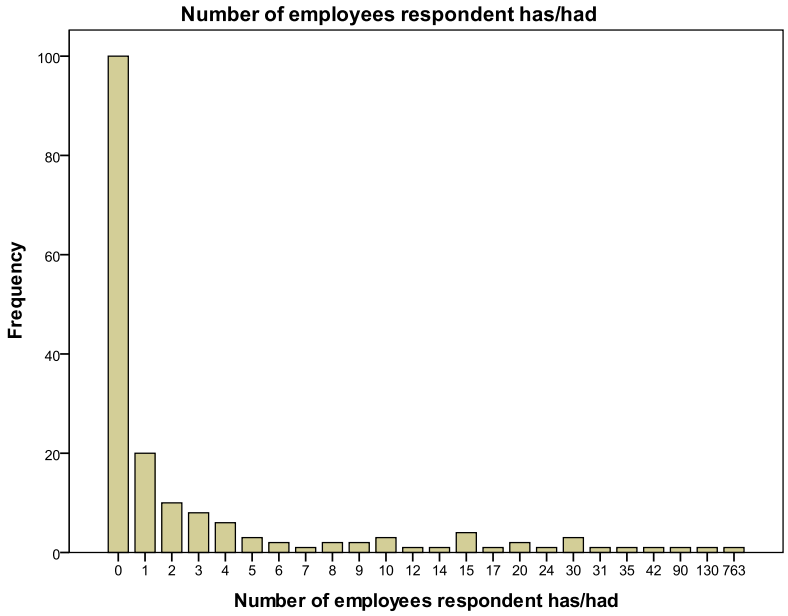
Kai modeliuojame kažkokių reiškinių skaičių (neišlaikytų egzaminų skaičių, žemės drebėjimų skaičių, laimėjimų loterijoje skaičių) ir duomenys nėra normalūs, tai dažniausiai taikome Puasono arba neigiamą binominę regresijas. Šiame skyrelyje aptarsime dvi šių regresijos modelių modifikacijas tais atvejais, kai:

- duomenyse per daug nulių;
- duomenyse nėra nulių.

9.1. Perteklinių nulių Puasono regresijos modelis

9.1.1. Modelio paskirtis

Kai modeliuojame retus įvykius, gana dažnai nulių būna daug daugiau, nei tiktų Puasono regresijai. Pavyzdžiui, turimų samdomų darbuotojų skaičiaus Švedijoje dažnių grafikas aiškiai rodo, kad dauguma respondentų samdomų darbuotojų neturi.



Perteklinių nulių Puasono regresijos (angl. zero-inflated Poisson regression) idėja glūdi samprotavime, kad nulis nuliui nelygus. Galbūt dalis respondentų yra tokie, kad jie iš principo negali turėti samdomų darbuotojų (pavyzdžiui, dirba valstybinėje tarnyboje). Todėl juos reikėtų atskirti nuo galinčių turėti samdomų darbuotojų (pavyzdžiui, užsiimančių privačia veikla) respondentų. Idėjiniame lygmenyje šis reikalavimas nusako kokiems duomenims skirtas modelis – visi respondentai skirstytini į dvi grupes: tuos, kuriems kintamojo reikšmė visada nulis ir tuos,

kuriems kintamasis dar gali įgyti ir kitokias reikšmes. Jeigu hareme klausinėjame apie vaikų skaičių, tai reikia skirti eunuchus ir žmonas.

Problema ta, kad realybėje nežinome, kurie respondentai kokiai grupei priklauso. Matematiškai tai reiškia, kad bendrasis modelis sudaromas, apjungus dvinarę logistinę regresiją su standartine Puasono regresija. Dvinarė logistinė regresija turi padėti rasti respondentus, kuriems priklausomas kintamasis visada lygus nuliui (nulinukų grupę). Tiems, kurie į šią grupę nepateko, taikomas Puasono regresijos modelis.

Regresoriai naudojami viename modelyje gali skirtis nuo kito modelio regresorių.

Kitais žodžiais tariant, visai nebūtina (bet leistina) į abudu modelius įtraukti tuos pačius regresorius. Formaliai kiekvieną modelio tikimybę galime užrašyti, kaip tikimybių mišinį:

P(Y = 0) = P(priklauso nulinukų grupei) + P(nepriklauso nulinukų grupei)×P(Y= 0, o respondentas yra iš antrosios grupės),

P(Y=k) = P(nepriklauso nulinukų grupei) ×P(Y=k, o respondentas yra iš antrosios grupės).

Nustatant modelio tikimą reikia:

- Pažiūrėti, ar statistiškai reikšminga didžiauso tikėtinumo chi kvadrato reikšmė. Geruose modeliuose $p < 0,05$.
- Su Voldo testu patikrinti, ar visi kintamieji statistiškai reikšmingi Puasono regresijos modelyje, skirtame ne nulinukams. Geruose modeliuose visiems regresoriams $p < 0,05$. Konstantai p reikšmės nežiūrime.
- Su Voldo testu patikrinti, ar visi kintamieji statistiškai reikšmingi logistiniame modelyje, atskiriančiame nulinukus. Geruose modeliuose visiems regresoriams $p < 0,05$. Konstantai p reikšmės nežiūrime.
- Vuongo testu patikrinti, ar perteklinių nulių Puasono regresijos modelis statistiškai reikšmingai skiriasi nuo paprasto Puasono regresijos modelio. Skiriasi, jeigu $p < 0,05$.

Perteklinių nulių Puasono regresijos nerekomenduojama taikyti labai mažoms imtims (n< 100).

9.1.2. Perteklinių nulių Puasono regresija su STATA

Atliksime tą pačią analizę, kaip ir nagrinėdami neigiamą binominę regresiją. Tiriame Švedijos duomenis **ESS4SE**. Modeliuojame *emplno* (turimų darbuotojų skaičius) priklausomybę nuo regresorių

- *brwmny* (ar sunku pasiskolinti pinigų išgyvenimui, 1 – labai sunku, ..., 5 – labai lengva),
- *eduyrs* (prasimokytų metų skaičius),
- *emplnof2* (tėvo turimų darbuotojų skaičius, 0 – neturi, 1 – turi).

Atidarius failą su STATA dar įvykdome: keep if emplno < 900 & emplnof < 4 & eduyrs < 88 & brwmny < 8.

Puasono modeliui pasirinksime tokius regresorius: *brwmny, eduyrs, emplnof2*. Logistiniame modelyje apsiribosime vienu regresoriumi – *emplnof2*. Regresorių parinkimas modeliams remiasi teoriniais samprotavimais ir praktiniais bandymais. Teoriškai bandome išmąstyti, nuo ko priklauso – turės respondentas darbuotojų, ar neturės (logistinis modelis) ir, nuo ko priklauso darbuotojų skaičius, jeigu jų respondentas turi (Puasono modelis). Praktiniai bandymai reiškia, kad išbandomi keli variantai, ir renkamasis geriausias. Perteklinių nulių Puasono regresija atliekama, įvykdant:

```
zip emplno eduyrs brwmny emplnof2 , inflate(  emplnof2) vuong
```

```
Zero-inflated Poisson regression                  Number of obs   =         66
                                                  Nonzero obs     =         30
                                                  Zero obs        =         36

Inflation model = logit                           LR chi2(3)      =    3477.94
Log likelihood  = -598.8196                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
emplno       |
      eduyrs |   .1851223   .0190059     9.74   0.000     .1478715     .222373
      brwmny |  -1.687293   .0484736   -34.81   0.000      -1.7823   -1.592286
    emplnof2 |   4.018727   .1661605    24.19   0.000     3.693058    4.344395
       _cons |   2.908324   .2662471    10.92   0.000     2.386489    3.430158
-------------+----------------------------------------------------------------
inflate      |
    emplnof2 |   2.165891   1.013478     2.14   0.033     .1795099    4.152272
       _cons |  -2.048418   .9481365    -2.16   0.031    -3.906732   -.1901049
------------------------------------------------------------------------------
Vuong test of zip vs. standard Poisson:            z =     2.12  Pr>z = 0.0169
```

Didžiausio tikėtinumo chi kvadrato statistikos (3477,94) *p* reikšmė (0,000) daug mažesnė už 0,05. Tai rodo gerą modelio bendrąjį tinkamumą.

Pirmoji lentelės dalis skirta Puasono regresijai. Visos interpretacijos išlieka, kaip ir Puasono regresijos atveju. Pavyzdžiui, teigiama *emplnof2* regresoriaus koeficiento reikšmė rodo, kad daugiau darbuotojų turi tie respondentai, kurių tėvai irgi turi samdomų darbuotojų. Matome, kad pagal Voldo testą visi regresoriai statistiškai reikšmingi. Antroji lentelės dalis (**inflate**) yra skirta

logistiniam modeliui, padedančiam atskirti nulinukus nuo likusiųjų respondentų. Matome, kad regresorius *inflate* yra statistiškai reikšmingas. Lentelės pabaigoje pateikiamas Vuong testas, palyginantis nagrinėtąjį modelį su paprastu Puasono regresijos modeliu. Testo *p* reikšmė lygi 0,0169 < 0,05. Todėl hipotezę, kad modeliai nesiskiria atmetame.

Norėdami gauti išsamesnę informaciją apie koeficientų eksponentes ir kitas charakteristikas, galime įvykdyti:

```
listcoef
```

emplno	b	z	P>|z|	e^b	e^bStdX	SDofX
eduyrs	0.18512	9.740	0.000	1.2034	1.9956	3.7325
brwmny	-1.68729	-34.808	0.000	0.1850	0.1646	1.0693
emplnof2	4.01873	24.186	0.000	55.6302	7.5743	0.5038

b = raw coefficient

Pateikėme tik trumpą rezultatų fragmentą, skirtą ne nulinukų grupei. Kiekvieneri papildomai prasimokyti metai didina vidutinį turimų darbuotojų skaičių 1,203 karto.

Reikšmės prognozuojamos analogiškai Puasono regresijai. Pavyzdžiui, padarysime prognozes visoms (realiai pimoms dešimčiai) tikimybių, kai *emplnof2* = 1. Kitų regresorių reikšmės bus paimtos lygios jų vidurkiui. Įvykdome:

```
prvalue, x(emplnof2 = 1)
```

```
zip: Predictions for emplno

  Expected y:          8.4091
  Pr(Always0|z):       0.5293
  Pr(y=0|x,z):         0.5293
  Pr(y=1|x):           0.0000
  Pr(y=2|x):           0.0000
  Pr(y=3|x):           0.0000
  Pr(y=4|x):           0.0000
  Pr(y=5|x):           0.0001
  Pr(y=6|x):           0.0004
  Pr(y=7|x):           0.0009
  Pr(y=8|x):           0.0021
  Pr(y=9|x):           0.0042

x values for count equation

       eduyrs     brwmny   emplnof2
x=  11.712121  3.6818182          1

z values for binary equation

    emplnof2
z=         1
```

Visos tikimybės įvertintos tam atvejui, kai *emplnof2* = 1, *eduyrs* = 11,71 (*eduyrs* vidurkis yra lygus 11,71), *brwmny* = 3.68. Vidutinis tikėtinas darbuotojų skaičius – 8,409. Tikimybė, kad

respondentas yra iš nulinukų grupės, lygi 0,529. Tikimybė, kad respondentas turi aštuonis darbuotojus lygi 0,0021 ir t.t.

9.2. Perteklinių nulių neigiama binominė regresija

9.2.1. Modelio paskirtis

Viskas, ką galima pasakyti apie perteklinių nulių Puasono regresiją, tinka ir perteklinių nulių neigiamai binominei regresijai. Trumpai priminsime svarbiausius momentus.

- Modeliuojame kažkokių reiškinių skaičių. Duomenys – sveiki neneigiami skaičiai, juose daug nulių.
- Duomenys (pvz., histograma) visiškai nepanašūs į normaliuosius.
- Spėjama, kad visi respondentai skaidytini į tuos, kuriems priklausomas kintamasis neišvengiamai turi įgyti nulines reikšmes (nulinukų grupė), ir tuos, kuriems nulis yra tik viena iš galimybių.
- Perteklinių nulių neigiamos binominės regresijos modelis – tai dviejų modelių mišinys. Pirmasis iš jų – logistinės regresijos modelis, kuriuo nustatinėjama respondento priklausomybė nulinukų grupei. Antrasis modelis – neigiamos binominės regresijos modelis, skirtas ne nulinukų grupei.
- Regresoriai naudojami viename modelyje gali skirtis nuo kito modelio regresorių.

Nustatant modelio tikimą reikia:

1) Pažiūrėti, ar statistiškai reikšminga didžiauso tikėtinumo chi kvadrato reikšmė. Geruose modeliuose $p < 0,05$.
2) Su Voldo testu patikrinti, ar visi kintamieji statistiškai reikšmingi neigiamos binominės regresijos modelyje, skirtame ne nulinukams. Geruose modeliuose visiems regresoriams $p < 0,05$. Konstantai p reikšmės nežiūrime.
3) Patikrinti, ar neigiamo binominio skirstinio parametras *alpha* skiriasi nuo nulio. Skirtumas nuo nulio statistiškai reikšmingas, kai atitinkama p reikšmė $< 0,05$. Jeigu *alpha* nuo nulio statistiškai reikšmingai nesiskiria, tai labai tikėtina, kad taikytina Puasono regresija.
4) Su Voldo testu patikrinti, ar visi kintamieji statistiškai reikšmingi logistiniame modelyje, atskiriančiame nulinukus. Geruose modeliuose visiems regresoriams $p < 0,05$. Konstantai p reikšmės nežiūrime.
5) Vuongo testu patikrinti, ar perteklinių nulių neigiamos binominės regresijos modelis statistiškai reikšmingai skiriasi nuo paprasto neigiamos binominės regresijos modelio. Skiriasi, jeigu $p < 0,05$.

9.2.2. Perteklinių nulių neigiama binominė regresija su STATA

Atliksime tą pačią analizę, kaip ir nagrinėdami neigiamą binominę regresiją. Tiriame Švedijos duomenis **ESS4SE**. Modeliuojame *emplno* (turimų darbuotojų skaičius) priklausomybę nuo regresorių

- *brwmny* (ar sunku pasiskolinti pinigų išgyvenimui, 1 – labai sunku, ..., 5 – labai lengva),
- *eduyrs* (prasimokytų metų skaičius),
- *emplnof2* (tėvo turimų darbuotojų skaičius, 0 – neturi, 1 – turi).

Atidarius failą su STATA dar įvykdome: keep if emplno < 900 & emplnof < 4 & eduyrs < 88 & brwmny < 8.

Neigiamai binominei regresijai imsime tokius regresorius: *brwmny, eduyrs, emplnof2*. Logistiniame modelyje apsiribosime vienu regresoriumi – *emplnof2*. Įvykdome:

```
zinb emplno eduyrs brwmny emplnof2 , inflate(  emplnof2) vuong
```

```
Zero-inflated negative binomial regression        Number of obs   =         66
                                                  Nonzero obs     =         30
                                                  Zero obs        =         36

Inflation model = logit                           LR chi2(3)      =      25.62
Log likelihood  = -133.0299                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
emplno       |
      eduyrs |   .2411996    .119541     2.02   0.044     .0069036    .4754956
      brwmny |  -.8253969   .2681687    -3.08   0.002    -1.350998   -.2997959
    emplnof2 |   2.080372   .6475327     3.21   0.001      .811231    3.349513
       _cons |   .7150847   1.809166     0.40   0.693    -2.830816    4.260985
-------------+----------------------------------------------------------------
inflate      |
    emplnof2 |    13.6117   852.2105     0.02   0.987     -1656.69    1683.914
       _cons |  -14.46382   852.2107    -0.02   0.986    -1684.766    1655.839
-------------+----------------------------------------------------------------
    /lnalpha |   1.310606   .3326411     3.94   0.000     .6586413     1.96257
-------------+----------------------------------------------------------------
       alpha |    3.70842   1.233573                      1.932165    7.117599
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =     0.61  Pr>z = 0.2710
```

Visi rezultatai aprašomi ir interpretuojami analogiškai perteklinių nulių Puasono regresijos modeliui. Statistiškai reikšmingą bendrąjį modelio tikimą rodo maža (0,000 < 0,05) didžiausio tikėtinumo chi kvadrato *p* reikšmė (Prob > chi2). Sprendžiant pagal Voldo kriterijus, visi regresoriai neigiamos binominės regresijos modelyje yra statistiškai reikšmingi (visos $p < 0,05$, išskyrus konstantos *p* reikšmę).

Logistinės regresijos modelis blogai atskiria nulinukus nuo likusiųjų, nes regresorius *emplnof2* nėra statistiškai reikšmingas ($p = 0,987 > 0,05$).

Neigiamos binominės regresijos parametras *alpha* = 3,708 statistiškai reikšmingas (p =0,00). Todėl tradicinė Puasono regresija šiems duomenims netinka.

Vuongo testo p = 0,27 > 0,05. Negalime atmesti hipotezės, kad perteklinių nulių modelis nesiskiria nuo įprasto neigiamos binominės regresijos modelio.

Galutinė išvada tokia – rezultatai nerodo, kad perteklinių nulių neigiamos binominės regresijos modelis būtų geresnis už standartinį neigiamos binominės regresijos modelį. Tai gana pamokoma išvada – nereikia tikėtis, kad kuo modelis sudėtingesnis, tuo daugiau šansų, jog jis tiks.

Jeigu reikėtų prognozuoti tikimybes, naudotina prvalue komanda, analogiška ankstesniems skyreliams.

9.3. Skaičiuojamų regresijos modelių palyginimas su STATA

Visi skaičiuojamieji regresijos modeliai (Puasono, perteklinių nulių Puasono, neigiamos binominės ir perteklinių nulių neigiamos binominės regresijos) turi labai nedaug tikimo duomenims rodiklių. STATA programoje yra numatyta galimybė palyginti skirtingus modelius. Tam skirta komanda *countfit*.

Jau ne kartą modeliavome respondento darbuotojų skaičių (*emplno*), atsižvelgdami į jo prasimokytų metų skaičių (*eduyrs*), požiūrį į paskolas (*brwmny,* ar sunku pasiskolinti pinigų išgyvenimui, 1 – labai sunku, ..., 5 – labai lengva) ir tėvo turimų darbuotojų skaičių (*emplnof2* (tėvo turimų darbuotojų skaičius, 0 – neturi, 1 – turi). Modelis: *emplno = f(eduyrs, brwmny, brwmny).*

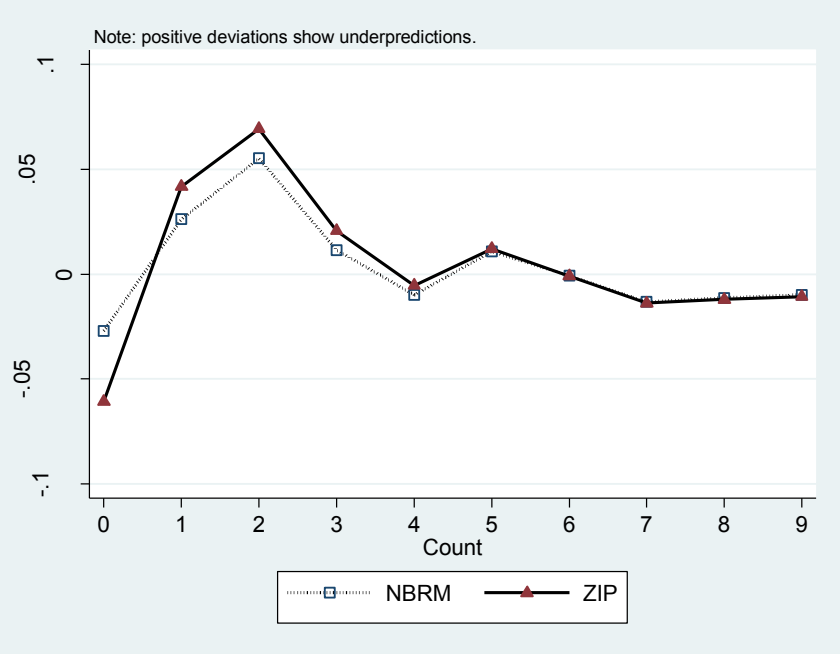
Kokius modelius norime lyginti, nurodome užrašydami:

- prm – Puasono regresijai,
- nbreg – neigiama binominė regresijai
- zip – perteklinių nulių Puasono regresijai
- zinb – perteklinių nulių neigiama binominė regresijai.

Atitinkamuose skyreliuose įsitikinome, kad šiems duomenims tinka ir perteklinių nulių Puasono regresijos ir neigiamos binominės regresijos modeliai. Palyginsime tuos du modelius:

```
countfit emplno eduyrs brwmny emplnof2 , inflate( emplnof2) zip nbreg
```

Visų pirma, aptarsime liekamųjų paklaidų grafikus. Kuo liekamosios paklaidos arčiau nulio (x-ų ašies, tuo modelis geresnis). Matome, kad abudu modeliai šiuo atžvilgiu pakankamai geri ir elgiasi panašiai. Nulius ir vienetus truputį geriau modeliuoja neigiama binominė regresija (NBRM).



Rezultatų išklotinės pradžioje pateikiama abiejų modelių koeficientų eksponentės (exp{b_j}) ir jų t – reikšmės. Pavyzdžiui, regresoriaus *eduyrs* toks eksponentinis koeficientas neigiamos binominės regresijos modelyje yra 1,332, o perteklinių nulių Puasono modelyje (ZIP) 1,203. Kai kurie koeficientai skiriasi radikaliai (5,1 ir 55,6). Kai kurios Voldo kriterijaus t reikšmės irgi skiriasi keletą kartų (-2,72 ir -34,81).

Variable	NBRM	ZIP
emplno		
Years of full-time education~m	1.332	1.203
	2.23	9.74
Borrow money to make ends me~	0.471	0.185
	-2.72	-34.81
	5.101	55.630
	2.51	24.19
Constant	0.961	18.326
	-0.02	10.92
lnalpha		
Constant	5.327	
	7.37	
inflate		
		8.722
		2.14
Constant		0.129
		-2.16
Statistics		
alpha	5.327	
N	66	66
ll	-133.952	-598.820
bic	288.853	1222.777
aic	277.905	1209.639

Neigiamos binominės regresijos (NBRM) atveju pateiktas parametro *alpha* logaritmo įvertis (5.327). Perteklinių nulių modeliui pateikti rezultatai apie logistinį (nulinukų) modelį. Įdomiausi rezultatai yra grafoje **Statistics**. Čia yra abiejų modelių didžiausio tikėtinumo funkcijos logaritmų išraiškos: -133,95 ir -598,82. Abi jos pakankamai didelės. Kuo didžiausio tikėtinumo funkcijos logaritmo reikšmė absoliutiniu didumu didesnė, tuo geriau modelis tinka. Pagal šį rodiklį geriau tiktų ZIP . Galų gale nurodytos informacinių indeksų BIC ir AIC reikšmės. Kuo šios reikšmės mažesnės, tuo modeliai geriau tinka. Pagal šį rodiklį žymiai geresnis būtų NBRM modelis.

Kitoje lentelėje išrašytos abiejų modelių didžiausios liekamosios paklaidos, kartu nurodant kur jos buvo. Šalia užrašytos liekamųjų paklaidų vidurkio absoliutinės reikšmės. Faktiškai ši lentelė pateikia tą pačią informaciją, kaip ir grafikas. Matyti, kad didžiausias neatitikimas buvo modeliuojant reikšmę 2. Vis dėlto abiejų modelių neatitikimai labai nedideli, mažos ir vidutinės paklaidos. Pagal šiuos rodiklius truputį geresnis atrodo NBRM modelis.

Comparison of Mean Observed and Predicted Count

Model	Maximum Difference	At Value	Mean |Diff|
NBRM	0.055	2	0.018
ZIP	0.069	2	0.025

NBRM: Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.545	0.572	0.027	0.084
1	0.121	0.095	0.026	0.480
2	0.106	0.051	0.055	3.966
3	0.045	0.034	0.012	0.261
4	0.015	0.025	0.010	0.255
5	0.030	0.020	0.011	0.394
6	0.015	0.016	0.001	0.002
7	0.000	0.013	0.013	0.872
8	0.000	0.011	0.011	0.743
9	0.000	0.010	0.010	0.644
Sum	0.879	0.847	0.176	7.702

ZIP: Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.545	0.606	0.061	0.403
1	0.121	0.079	0.042	1.455
2	0.106	0.037	0.069	8.639
3	0.045	0.025	0.021	1.151
4	0.015	0.020	0.005	0.091
5	0.030	0.018	0.012	0.539
6	0.015	0.016	0.001	0.003
7	0.000	0.014	0.014	0.914
8	0.000	0.012	0.012	0.793
9	0.000	0.011	0.011	0.703
Sum	0.879	0.838	0.248	14.691

Dar vienoje lentelėje išrašytos stebimos 0 – 9 tikimybės (t.y. kurią dalį stebėjimų sudarė atitinkamos reikšmės) ir prognozuojamos, taikant vieną ir kitą metodus (regresorius, laikant lygiais jų vidutinėms reikšmėms). Kuo mažiau skiriasi tikimybės, tuo modelis geresnis. Vis dėlto, įžiūrėti tuos skirtumus sudėtinga, todėl jų absoliutiniai didumai pateikti stulpelyje |**Diff**|. Stulpelyje **Perason** išrašyta tų skirtumų indėliai į Pirsono statistikos reikšmę. Atmetant detales, galima pasakyti, kad kuo šio reikšmė didesnė, tuo blogesnis tikimas. Užmetę akį į paskutinius du stulpelius, nustatome, kad NBRM modelis tikslesnis už ZIP modelį.

Paskutinėje lentelėje pateikiama reziume apie modelius, atsižvelgiant į informacinius indeksus. Matome, kad NBRM modelis atrodo daug geriau, nei ZIP.

```
NBRM          BIC=    12.336  AIC=     4.211  Prefer  Over  Evidence
-----------------------------------------------------------------------
  vs ZIP      BIC=   946.260  dif=  -933.924  NBRM    ZIP   Very strong
              AIC=    18.328  dif=   -14.117  NBRM    ZIP
-----------------------------------------------------------------------
ZIP           BIC=   946.260  AIC=    18.328  Prefer  Over  Evidence
```

Pastabos. Kartojant *countfit* komandą su kitais modeliais, reikia komandos gale pridėti *replace*. Priešingu atveju gausime pranešimą apie klaidą. Jeigu nenurodome, kuriuos modelius norime lyginti, tai palyginami visi keturi modeliai.

Galutinė lyginamosios analizės išvada – neigiamos binominės regresijos modelis tiko duomenims daug geriau, nei perteklinių nulių Puasono modelis.

9.4. Benulė Puasono regresija

9.4.1. Modelio paskirtis

Būna, kad kintamasis pagal savo prigimtį negali įgyti nulių, nors visais kitais atžvilgiais labai panašus į Puasono (ir visai nepanašus į normalųjį) kintamąjį. Pavyzdžiui, minimalus praleistų ligoninėje dienų skaičius yra 1. Minimalus namų ūkio narių skaičius – vienas. Puasono regresijos modelyje *visada* tariama, kad gali būti ir nulinės reikšmės ir surandamos jų įgijimo tikimybės. Todėl tokiems duomenims tiesiai taikyti Puasono regresijos modelio negalima. Vienas iš variantų – transformuoti duomenis, pavyzdžiui, atimant vienetą. Būtent taip mes transformavome namų ūkio narių skaičių į kitų namų ūkio subjektų skaičių (*numbhh* = *hhmmb* – 1). Alternatyva būtų benulis Puasono regresijos modelis.

Pastumtas Puasono modelis (tiksliau Puasono modelis pastumtiems duomenims) skiriasi nuo benulio Puasono modelio. Pastumtojo Puasono modelio atveju, Puasono kintamojo nulio tikimybė priskiriama vienetui, vieneto tikimybė priskiriama dvejetui ir t.t. Benulio Puasono modelio idėja yra kitokia, tikimybes laikome sąlyginėmis Puasono tikimybėmis. Maždaug taip – stebime tai Puasono kintamąjį, tik jis nulių neįgyja. Tikimybė tokiam Puasono dydžiui įgyti reikšmę *k*:

$$P(Y = k|y > 0) = \frac{P(Y = y)}{1 - P(Y = 0)}.$$

Svarbiausia benulio Puasono regresijos modelio bėda – minimalus statistinių rodiklių rinkinys. Labai sunku nuspręsti, ar šis modelis tinka duomenims. Faktiškai apsiribojama:

- Didžiausio tikėtinumo chi kvadrato statistika (gerai, kai jos $p < 0{,}05$.
- Patikrinama, ar pagal Voldo kriterijų regresoriai statistiškai reikšmingi.
- Palyginimui su pastumtuoju Puasono regresijos modeliu galima panaudoti informacinius indeksus AIC, BIC.

Analogiškai benuliam Puasono regresijos modeliui galim ištirti ir benulį neigiamos binominės regresijos modelį.

9.4.2. Benulė Puasono regresija su STATA

Pakartosime benulę Puasono regresinę analizę duomenims, kuriems jau taikėme Puasono regresiją. Tiriame 2008 metų Prancūzijos duomenis **ESS4FR**. Tyrime naudosime tokius kintamuosius: *agea* – respondento amžius, *hhmmb* – namų ūkio pastovių narių skaičius, *cldcrsv* – paramos vaiko priežiūrai teikimas tėvams, turintiems teisę ją gauti (0 – itin blogas,....., 10 – labai geras), *imptrad* – svarbu laikytis tradicijų ir papročių (1 – visai, kaip aš,...., 6 – absoliučiai ne taip, kaip aš) , *eduyrs* – prasimokytų metų skaičius. Tiriame stipriai tradicijų besilaikančius (*imptrad* ≤ 2) aukštojo mokslo neragavusius (*eduyrs* ≤ 10) respondentus(es). Su *tab* komanda peržiūrime kintamųjų reikšmes ir išmetame tas, kurios SPSS programoje buvo praleistų reikšmių kodai. Benulė Puasono regresija atliekama taip:

```
ztp hhmmb agea cldcrsv
```

```
Zero-truncated Poisson regression                 Number of obs   =        254
                                                  LR chi2(2)      =     109.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -307.33139                       Pseudo R2       =     0.1517
```

hhmmb	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agea	-.0292351	.0031247	-9.36	0.000	-.0353594	-.0231109
cldcrsv	.0829182	.0238419	3.48	0.001	.036189	.1296475
_cons	1.737643	.2178941	7.97	0.000	1.310579	2.164708

Didžiausio tikėtinumo chi kvadrato statistika pakankamai didelė (109,89) ir statistiškai reikšminga ($p = 0{,}00 < 0{,}05$). Visi regresoriai statistiškai reikšmingi (visiems jiems Voldo testo $p < 0{,}05$). Galima su *pvalue* surasti pirmųjų dešimties reikšmių skaičiaus tikimybių prognozes:

```
prvalue
```

```
                        Uncond     Cond
  Rate:                   1.39    1.851
  Pr(y=0|x):            0.2491        .
  Pr(y=1|x):            0.3462   0.4611
  Pr(y=2|x):            0.2406   0.3204
  Pr(y=3|x):            0.1115   0.1485
  Pr(y=4|x):            0.0387   0.0516
  Pr(y=5|x):            0.0108   0.0143
  Pr(y=6|x):            0.0025   0.0033
  Pr(y=7|x):            0.0005   0.0007
  Pr(y=8|x):            0.0001   0.0001
  Pr(y=9|x):            0.0000   0.0000

          agea    cldcrsv
x=  63.102362  5.2637795
```

Tikimybės buvo suskaičiuotos vidutinėms regresorių *agea* ir *cldcrsv* reikšmėms. Šios reikšmės parašytos lentelės apačioje (*agea* = 63,1; *cldcrsv* = 5,26). Pateiktos dviejų rūšių tikimybės. Stulpelyje **Uncond** yra tikimybės to Puasono skirstinio, kurį atsieit stebime su visomis nulio tikimybėmis. Stulpelyje **Cond** yra to paties Puasono kintamojo tikimybės, *kai žinome,* kad nulių imtyje nėra.

Norėdami sužinoti, kuris modelis – benulis ar pastumtasis – geriau tinka duomenims, surandame šio modelio informacinį indeksą AIC . Įvykdome

```
fitstat
```

```
BIC:                    -775.208   BIC':                    -98.820
BIC used by Stata:       631.275   AIC used by Stata:       620.663
```

Analogišką statistiką surandame Puasono regresijos modeliui su priklausomu kintamuoju *numbhh* = *hhmmb* – 1. Gauname:

```
BIC:                    -787.791   BIC':                   -101.844
BIC used by Stata:       618.692   AIC used by Stata:       608.080
```

Matome, kad AIC skiriasi nedaug (AIC = 620,66 ir AIC = 608,08), nors standartinei Puasono regresijai šis rodiklis ir mažesnis. Griežtai kalbant, toks AIC lyginimas, skirtingais būdais apskaičiuotiems modeliams nėra iki galo korektiškas. Vis dėlto jis duoda bendrą supratimą apie modelių gerumą. Reikia turėti omenyje, kad nedidelis AIC skirtumas, dar neleidžia teigti, kad kažkuris modelis yra geresnis. Koks AIC skirtumas jau yra pakankamai didelis, kad tą galėtume teigti? Šiuo klausimu nėra kažkokios nusistovėjusios vieningos nuomonės. Dažnai siūloma svarbiu laikyti AIC skirtumą viršijantį 10. Bet, žinoma, tai nėra privaloma. Tarkime, mūsų nagrinėjamu atveju skirtumas yra maždaug 12. Dešimtį viršija visai nedaug. Todėl mūsų nuomone prioritetas

turėtų būti teikiamas įprastai Puasono regresijai (po kintamojo „pastūmimo“, kad įgytų nulį), nes tiesiog šis modelis turi daugiau tinkamumo charakteristikų (pvz., deviaciją). Beje, nebūtina atsižvelgti būtent į AIC. Galima naudoti ir BIC. Kai kurie statistikai labiau vertina būtent šį indeksą. Išvados daromos, kaip ir AIC atveju. Matome, kad jos nepasikeičia.

Uždaviniai

1. Visiems Puasono regresijos ir neigiamos binominės regresijos uždaviniams ir pavyzdžiams pritaikyti perteklinių nulių modelius. Palyginti, kurie modeliai geriausiai tinka duomenims.

Biudžeto išlaidos visada būna perteklinių nulių modelis

10. PROBIT REGRESIJA

Tokiu visai nelietuvišku terminu vadinama viena dvinarės logistinės regresijos alternatyva. Ją galima taikyti dvejopai: a) tose situacijose, kai taikoma ir logistinė regresija, b) specifiniams dozavimo (dažnių) modeliams nagrinėti. Tipiniai šių dviejų situacijų pavyzdžiai::

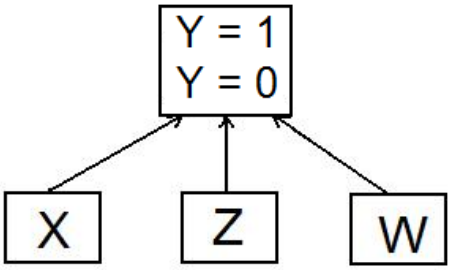
- Norime išsiaiškinti, ar respondento lytis ir socialinis statusas daro įtaką jo požiūriui į mirties bausmės panaikinimą.
- Tiriame, kaip finansiniai stimulai veikia respondentų norą tarnauti profesionalioje kariuomenėje. Kokio dydžio turi būti atlyginimas, kad 30% abiturientų rinktųsi Karo akademiją?

Abi situacijos aptartos atskirai.

10.1. Probit regresijos modelis

10.1.1. Modelio lygtis

Modeliuojamas *dvireikšmis* kintamasis Y priklauso nuo X, Z, W. Modelio schema:



Kintamasis *Y* vadinamas *priklausomu* (arba *regresuojamu*) *kintamuoju*, kintamieji *X, Z, W* vadinami *nepriklausomais kintamaisiais* arba *regresoriais*. Tradiciškai manoma, kad *Y* įgyjamos reikšmės yra **kodai 0 ir 1**. Ką šie kodai reiškia, priklauso nuo konkretaus tyrimo. Matematinis modelis sudaromas ne pačiam priklausomam kintamajam, o jo tikimybei P(*Y* = 0):

$$\boxed{\mathrm{P}(Y = 0) = \Phi(C + b_1 X + b_2 Z + b_3 W).}$$

Čia $\Phi(\cdot)$ žymi standartinio normaliojo dydžio pasiskirstymo funkciją. Šią funkciją galima užrašyti tik kaip gana bjaurų integralą. Ekvivalentus modelio užrašas yra

$$\Phi^{-1}\big(\mathrm{P}(Y = 0)\big) = C + b_1 X + b_2 Z + b_3 W.$$

Čia $\Phi^{-1}(\cdot)$ žymi atvirkštinę funkciją, kuri dar vadinama *probit* funkcija. Iš čia ir kilo pačio modelio pavadinimas. Koeficientų *C*, b_1, b_2, b_3 reikšmės nėra žinomos. Jų įverčiai $\hat{C}, \hat{b}_1, \hat{b}_2, \hat{b}_3$ gaunami, panaudojus imties duomenis. Atsižvelgdami į koeficientų reikšmes, nustatysime, kurie

regresoriai svarbesni. Jeigu koeficientas prie kažkurio kintamojo teigiamas, tai šiam kintamajam didėjant, tikimybė Y įgyti nulį (o kokią situaciją ta modelio lygybė $Y = 0$ atitinka, turime žinoti iš sąlygos) didėja. Jeigu koeficientas neigiamas, tai atitinkamam kintamajam didėjant, tikimybė Y įgyti nulį mažėja (didėja tikimybė, kad Y įgis 1).

Jeigu $b_1 > 0$, tai X didėjant, didėja ir tikimybė P(Y= 0).

Jeigu $b_1 < 0$, tai X didėjant, didėja ir tikimybė P(Y= 1).

Priešingai dvinarei logistinei regresijai, jokie galimybių santykiai probit regresijoje nėra skaičiuojami. Prognozė, kam bus lygi tikimybė P(Y = 0), atliekama tik konkrečioms X, Z, W reikšmėms. Prognozę lengviausia atlikti dviem etapais. Pradžioje į formulę

$$z = \hat{C} + \hat{b}_1 X + \hat{b}_2 Z + \hat{b}_3 W$$

įstatome tas konkrečias X, Z ir W reikšmes, kurioms reikia prognozės. Tada tikimybių įverčiai apskaičiuojamas pagal formules:

$$\widehat{\mathrm{P}}(Y = 0) = \Phi(z), \quad \widehat{\mathrm{P}}(Y = 1) = 1 - \Phi(z)\,.$$

Funkcijos $\boldsymbol{\Phi(z)}$ reikšmių lentelės yra kiekviename statistikos vadovėlyje.

Pastaba. Gali būti ir modelis tikimybei P(Y = 1). Tada visose išvadose reikia sukeisti tikimybes P(Y =1) ir P(Y = 0) vietomis.

10.1.2. Reikalavimai duomenims

Duomenys turi tenkinti tokius reikalavimus:

b) *Priklausomas kintamasis Y yra dvireikšmis*. Formaliai kalbant, visi kiti kintamieji gali būti arba intervaliniai, arba dvireikšmiai (įgyti reikšmes 0 arba 1). Geriausia, kai dauguma regresorių yra intervaliniai, o dvireikšmių kintamųjų nedaug. Kai kada nurodoma, kad priklausomas kintamasis *savo prigimtimi* turi būti diskretizuotas normalusis kintamasis.

d) *Duomenyse negali vyrauti viena iš Y reikšmių*. Dažniausiai reikalaujama, kad tarp stebimų Y reikšmių vienetų (nulių) būtų ne mažiau penktadalio.

e) Jeigu modelyje yra daug kategorinių kintamųjų, tai *kiekvienam tų kategorijų deriniui turi būti pakankamai daug respondentų*. Minimalus reikalavimas – bent 5 respondentai.

f) Regresoriai neturi labai stipriai tarpusavyje koreliuoti (neturi būti regresorių multi-kolinearumo).

10.1.3. Modelio tinkamumas

Dozavimo modelių tinkamumo charakteristikas aptarsime atskirai. Bendruoju atveju modelio tikimą duomenims parodo tokie rodikliai:

1) *Klasifikacinė lentelė.* Konkretiems stebėjimams prognozuojama *Y* reikšmė, ir žiūrima, ar spėjimas sutapo su tikrąja *Y* reikšme. Kuo daugiau sutapimų, tuo modelis geresnis. Kiekvienos kategorijos stebėjimų turi būtu suklasifikuota daugiau, nei gautume atsitiktinai spėdami.
2) *Didžiausio tikėtinumo chi kvadrato statistika*. Parodo, ar modelyje yra bent vienas reikalingas regresorius. Jeigu statistikos *p* reikšmė didesnė už 0,05, tai regresijos modelio tinkamumas labai abejotinas.
3) *Voldo testai regresoriams*. Padeda nuspręsti ar kintamasis šalintinas iš modelio. Jeigu konkretaus testo *p* reikšmė < 0,05, tai sakome, kad *regresorius yra statistiškai reikšmingas* ir dažniausiai jį modelyje paliekame. Jeigu *p* reikšmė ≥ 0,05, tai regresorius yra statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais. Dažniausiai modelio konstanta *C* paliekama net ir tada, kai ji statistiškai nereikšminga.
4) *Determinacijos (pseudo) koeficientai* (angl. R square). Rodo bendrąjį modelio tikimą duomenims. Determinacijos koeficientas įgyja reikšmes iš intervalo [0, 1]. Kuo koeficiento reikšmė didesnė, tuo modelis geriau tinka duomenims. Ne itin gerai, kai R^2 yra labai mažas (R^2 < 0,2). Determinacijos koeficientų yra net keli ir visi jie yra pseudokoeficientai. Todėl į juos kreipti daug dėmesio neverta. Galima naudoti bet kurį. Tiesa, privalu nurodyti tikslų determinacijos koeficiento pavadinimą.
5) *Kuko matas.* Parodo ar duomenų imtyje yra išskirčių. Išskirčiai Kuko matas viršija 1. Skaičiuojamas kiekvienam respondentui. Modelis su išskirtimis – nepatikimas.
6) *Deviacija.* Probit regresijos modelis yra apibendrintojo tiesinio modelio atskiras atvejis. Kaip ir kitiems tos rūšies modeliams galima patikrinti ar deviacijos ir jos laisvės laipsnių santykis yra artimas vienetui. Kuo artimesnis, tuo modelis geriau tinka. Vis dėlto, probit regresijos atveju žymiai patikimesnis modelio gerumo rodiklis yra klasifikacinė lentelė ir didžiausio tikėtinumo Chi kvadrato statistika.

10.1.4. Bendrosios probit regresinės analizės etapai

Išsamios modelio analizės etapai:

1) Peržiūrime klasifikacinę lentelę. Tikriname, ar kiekvienai *Y* kategorijai teisingai klasifikuota bent tiek atvejų, kiek gautume atsitiktinai spėliodami.

2) Patikriname, ar chi kvadrato statistikos p reikšmė < 0,05. Jeigu ne – modelis netinkamas.
3) Patikriname, ar visi regresoriai statistiškai reikšmingi (visos Voldo p reikšmės < 0,05). Jeigu ne – modelis taisytinas. Konstantai p reikšmės nežiūrime.
4) Patikriname, ar modelyje nėra išskirčių (nėra stebėjimų, kuriems Kuko matas > 1).
5) Žiūrime, ar koeficientų ženklai atitinka spėjamą regresorių įtaką (neįtariame multikolinearumo).
6) Pasirinktasis determinacijos pseudokoeficientas yra pakankamai didelis (> 0,20).
7) Jeigu viskas gerai – modelį aprašome. Jeigu ne – tobuliname.

Labai gerai duomenims tinkančiame modelyje:

- Didžiausio tikėtinumo chi kvadrato p < 0,05.
- Visiems regresoriams Voldo kriterijaus p < 0,05.
- Koeficientų ženklai atrodo logiški.
- Teisingai klasifikuojama daug atvejų, kai $Y = 1$ ir kai $Y = 0$.
- Visų duomenų Kuko matai ≤ 1.
- Pasirinktasis determinacijos koeficientas ≥ 0,20.

10.1.5. Modelio tobulinimas

Modelio tobulinimas reiškia regresorių šalinimą arba jų transformaciją. Transformacija (logaritmavimas ir pan.) retai kada padeda, bet pabandyti visada galima. Modelis tobulinamas, jeigu yra problemų:

1) *Yra statistiškai nereikšmingas regresorius* (jam Voldo kriterijaus arba didžiausio tikėtinumo statistikos $p \geq 0{,}05$). Regresorius pašalinamas iš modelio ir probit regresinė analizė pakartojama be jo. Viskas gerai, jeigu klasifikacinėje lentelėje teisingai klasifikuotų atvejų sumažėjo nežymiai. Jeigu teisingai klasifikuotų atvejų sumažėja radikaliai, tai regresorių į modelį gražiname. Aprašydami tada konstatuosime, kad „regresorius buvo statistiškai nereikšmingas. Vis dėlto jis paliktas modelyje, nes be jo labai sumažėja teisingų klasifikavimų procentas“.
2) *Yra multikolinearumo problema.* Modelyje kai kurių koeficientų ženklai neatitinka sveiko proto. Tada pasitikrinama, ar intervaliniai regresoriai stipriai koreliuoja. Jeigu taip – tai dalį regresorių pašaliname. Kartais (jeigu toks veiksmas prasmingas) stipriai koreliuojančius regresorius keičiame jų vidurkiu.
3) *Yra išskirčių.* Kartais išskirtį galima pašalinti (pavyzdžiui, įtarus vedimo klaidą).

10.1.6. Dozavimo modelis

Vienas svarbiausių ir neabejotinai populiariausias probit regresijos taikymas yra vadinamasis dozavimo (dažnių) modelis. Kada jis taikomas? Lengviausia tą bus suprasti, aptariant konkretų pavyzdį. Tarkime, kad reikia nuspręsti kokios koncentracijos chemikalų tirpalą naudoti, naikinant amarus. Eksperimentas vykdomas taip: amarus suskirstome į grupes, kurias apipurškiame skirtingos koncentracijos tirpalu. Žinome, kiek kurioje grupėje buvo amarų ir kokios koncentracijos tirpalu jie buvo apipurkšti. Suskaičiuojame kiek amarų išsikraustė į aną pasaulį. Norime sužinoti, kaip amaro nugaišimo tikimybė priklauso nuo tirpalo koncentracijos. Be to, neretai reikia atsakyti į klausimą, kokios koncentracijos tirpalo reikia, jeigu norime išnaikinti 90% ar 80% amarų. Nes 100 % amarų išnaikinimas gali reikšti, kad drauge išnaikinome visus aplinkinių kaimų gyventojus ir nebėra kam uogas valgyti.

Tradiciniame dozavimo modelyje žinome:

- kiekvienos grupės respondentų skaičių,
- kiek respondentų kiekvienoje grupėje įgijo norimą būseną,
- kokią dozę kiekvienos grupės respondentai gavo,
- kiek respondentų, įgytų norimą būseną be dozės.

Dažniausiai tariama, kad be poveikio kintamojo (nulinė dozė) norimos būsenos neįgis nė vienas respondentas. Nenuodytas amaras išgyvens. Kartais taip nėra. Tiriant, kaip finansiniai stimulai veikia norą studijuoti karybą, stimulų nebuvimas dar nereiškia, kad niekas karybos nestudijuos. Gal respondentas visą vaikystę alaviniais kareivėliais žaidė, o dabar norėtų pabandyti pažaisti gyvais.

Dozavimo modeliuose yra palyginus nedaug modelio tikimo charakteristikų:

1) *Pirsono chi kvadratas.* Ji rodo gerą/blogą bendrąjį modelio tikimą duomenims. Gerai, kai p reikšmė $\geq 0{,}05$.
2) *Regresorių Z statistika.* Tai t-testo tiesinėje regresijoje ir Voldo testo logistinėje regresijoje analogas. Regresorius yra statistiškai reikšmingas (modelis geras), kai $p < 0{,}05$.
3) *Duomenų probit grafikas.* Probit regresijos modelyje tariama, kad priklausomybė tarp regresuojamo kintamojo Y reikšmių $\Phi^{-1}(Y)$ ir regresorių yra tiesinė. Grafike ir atidedamos šios reikšmės. Modelis tuo geresnis, kuo šis grafikas labiau primena tiesę. Jeigu reikia, duomenys transformuojami – logaritmuojami ir pan.

Nors klasikiniuose dozavimo modeliuose yra tik vienas dozuojamas poveikio kintamasis (regresorius), tokių regresorių gali būti ir daugiau.

Labai gerai duomenims tinkančiame dozavimo modelyje:

- Pirsono chi kvadrato $p \geq 0{,}05$.
- Visiems regresoriams Z kriterijaus $p < 0{,}05$.
- Duomenų probit grafikas primena tiesę.

10.1.7. Standartiniai regresijos modelių aprašymai

Visada aprašomas galutinis, po visų tobulinimų gautasis, modelis. Privaloma aprašyti modelio chi kvadrato statistikos reikšmingumą ir aptarti regresorių statistinius reikšmingumus. Pateiksime keletą galimų aprašymų fragmentų.

- *Probit regresijos modelis tiko, didžiausio tikėtinumo* $\chi^2 = 130{,}2$; $p < 0{,}01$. *Visi regresoriai buvo statistiškai reikšmingi. Teisingai klasifikuota 81,3% imties respondentų. Nagelgerkės determinacijos koeficientas lygus 0,64.*
- *Amarų skaičiaus priklausomybę nuo STPK_3000 koncentracijos purškiamame skystyje aprašo probit modelis:* $P(amarui\ amen) = \Phi(-1{,}8 + 0{,}11STPK\_3000)$. *Pirsono chi kvadrato statistika* $\chi^2 = 3{,}2$; $p = 0{,}65$. *Norint išnaikinti 80% amarų reikalinga koncentracija STPK_3000 = 30. Transformuotų duomenų grafikas primena tiesę.....*

10.2. Probit regresinė analizė su SPSS

10.2.1. Duomenys

Kaip atlikti standartinę probit regresinę analizę, parodysime, tirdami universitetų absolventų integraciją darbo rinkoje, 2009 m. lapkritis – 2010 m. sausis. Failas **LAMS**, kintamieji:

- *K2* – aukštoji mokykla
- *K33_2* – pasiekiant dabartinę padėtį darbo rinkoje studijų metu įgytos žinios (1 - visai ne-reikšmingos, ..., 5 – labai reikšmingos),
- *K35_1* – esamo darbo atitkimas bakalauro studijų krypčiai (1 - tikrai taip, 2 – greičiau taip, 3 – greičiau ne, 4 – tikrai ne),
- *K36_1* – studijų metu įgyto profesinių žinių panaudojimas darbe (1 - visiškai nenaudoju, ..., 5 – labai dažnai naudoju),
- *K37_1* – pasitenkinimas esamu darbu (1 – labai nepat. , ..., 5 – labai pat.).

Tirsime, kaip Klaipėdos universiteto bakalaurų profesinių žinių panaudojimo vertinimas susijęs su požiūriu į žinių svarbą darbo rinkoje ir pasitenkinimu esamu darbu. Papildomai atsižvelgsime į darbo atitikimą studijų krypčiai.

Visų pirma, pašaliname iš tyrimo respondentus, kurie į klausimus neatsakė (visiems kintamiesiems nurodome, kad *missing value all (99)*). Atsirenkame Klaipėdos universiteto absolventus (*Select cases if K2* = 3). Su *Frequences* komanda ištiriame *K36_1* reikšmes:

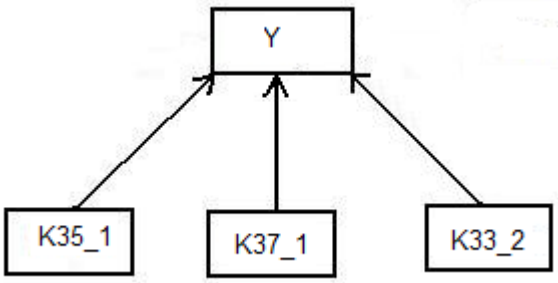
K36_1 Pirmųjų studijų metu įgytų profesinių žinių panaudojimo darbe įvertinimas

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Visiškai nenaudoju	10	2.6	3.1	3.1
	2	22	5.6	6.7	9.8
	3	69	17.6	21.1	30.9
	4	109	27.9	33.3	64.2
	5 Labai dažnai naudoju	117	29.9	35.8	100.0
	Total	327	83.6	100.0	
Missing	99 Nestudijavau / negaliu atsakyti	1	.3		
	System	63	16.1		
	Total	64	16.4		
Total		391	100.0		

Matome, kad labai nedaug respondentų rinkosi reikšmes 1 ir 2. Todėl su komanda *recode* sukuriame naują kintamąjį *Y*, kuris įgyja dvi reikšmes: 0 – jeigu respondentai nedažnai naudojasi studijų žiniomis (*K36_1* = 1, 2 arba 3) ir 1 – jeigu respondentai dažnai naudojasi studijų žiniomis (*K36_1* = 4 arba 5). Tirsime modelį:

$$Y = f(K35\_1, K33\_2, K37\_1).$$

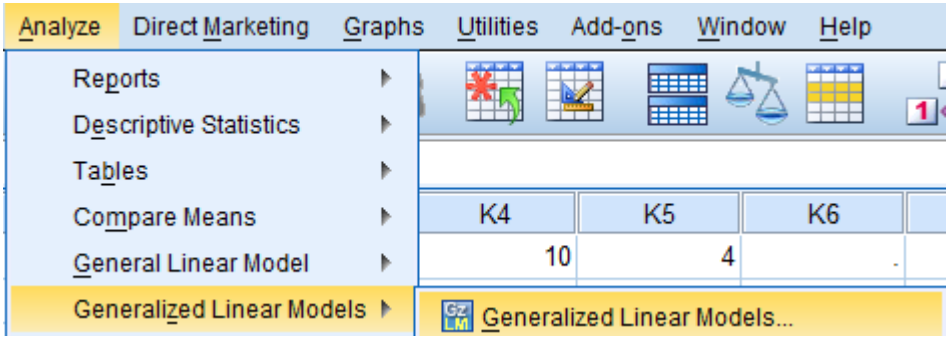
Grafiškai tiriamą modelį galima pavaizduoti taip:



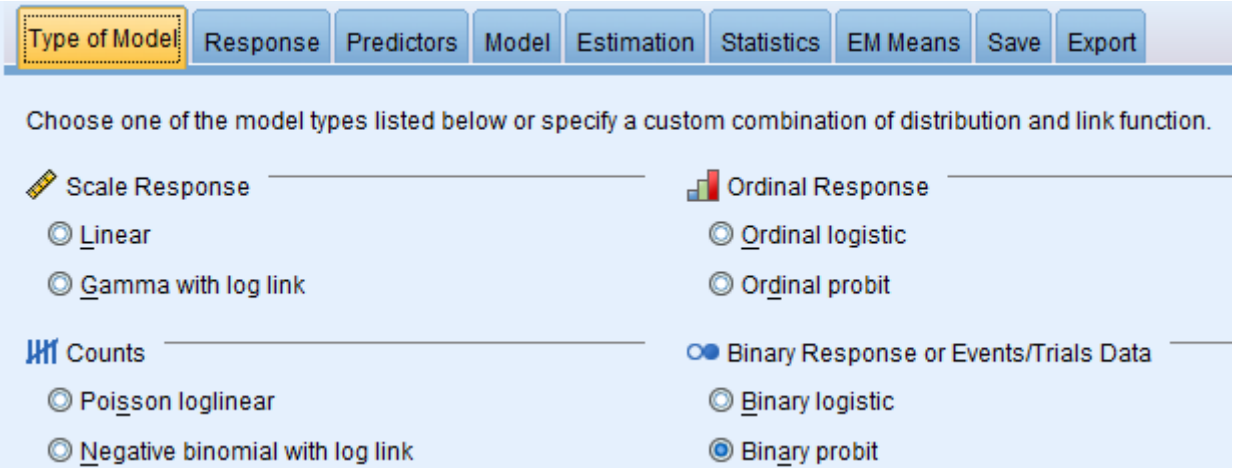
Mes atlikome pradinę modelio specifikaciją. Dabar patikrinsime, ar parinktasis modelis tinka duomenims.

10.2.2. SPSS parinktys pradiniam regresijos modeliui

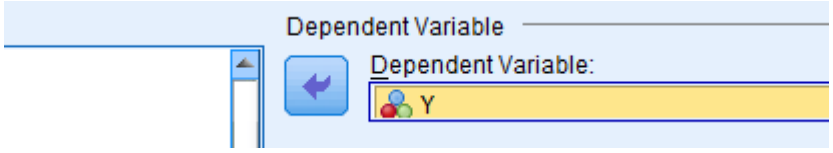
Tradicinis probit regresijos modelis tiriamas, kaip apibendrintasis tiesinis modelis. Atsidarome langelį **Analyze** ir renkamės **Generalized Linear Models → Generalized Linear Models.**



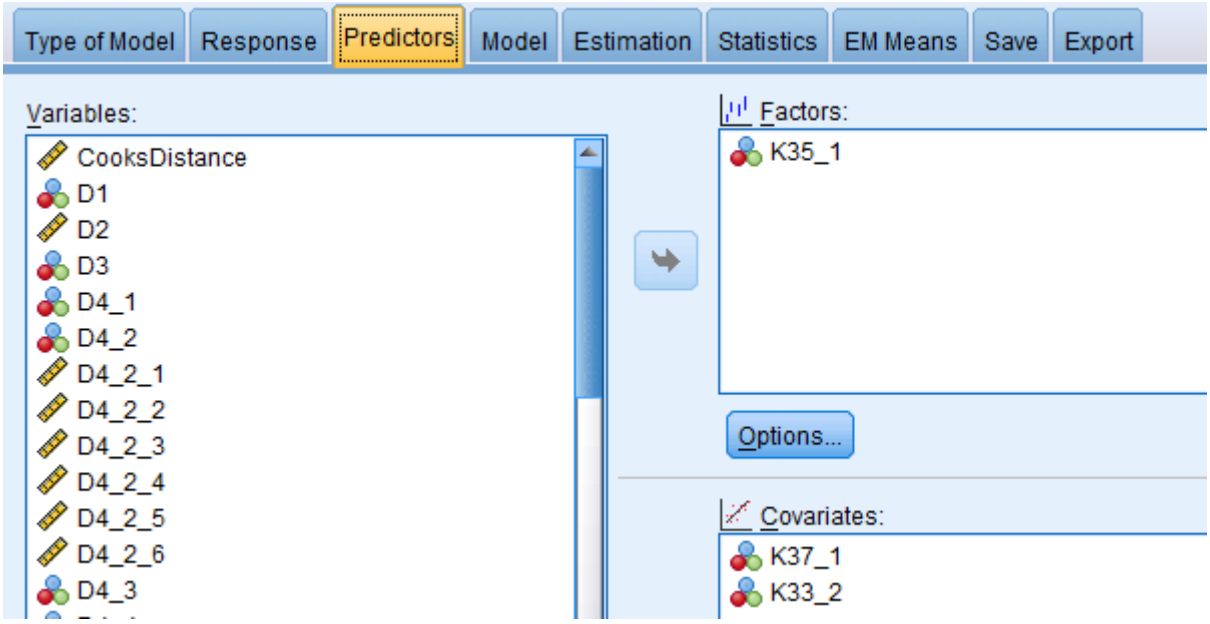
Atsidariusiame meniu renkamės langą **Type of Model** ir pažymime **Binary probit.**



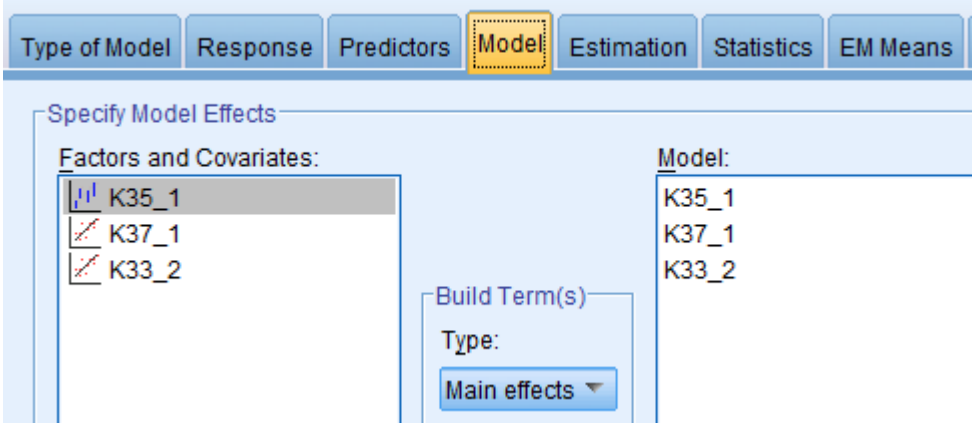
Atsidarome langą **Response** ir į laukelį **Dependent Variable** įkeliame kintamąjį *Y*.



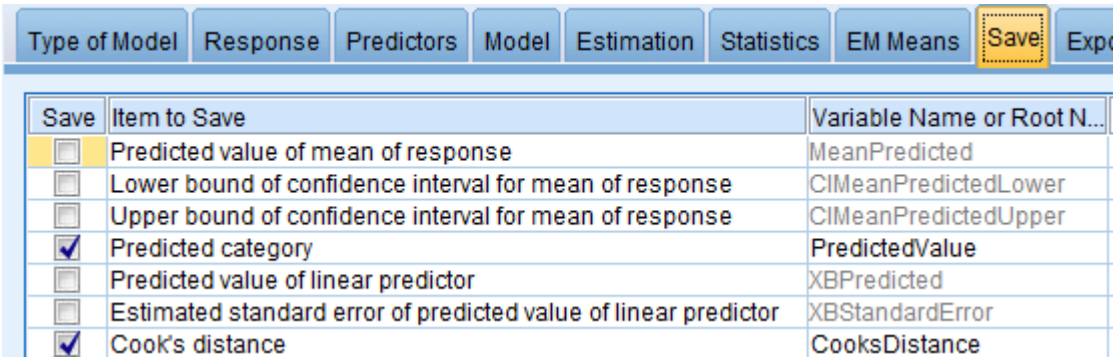
Atidarome langą **Predictors** ir į **Covariates** įkeliame *K37_1* ir *K33_2* (su tam tikromis išlygomis juos galima laikyti intervaliniais kintamaisiais. Regresorius *K35_1* įgyja tik 4 reikšmes, todėl traktuosime jį, kaip kategorinį kintamąjį. Įkeliame *K35_1* į langą **Factors.**



Atidarome langą **Model** ir visus regresorius perkeliame į dešinįjį langą.



Atidarome langą **Save** ir pažymime **Predicted category** ir **Cook's distance**. Pastarosios parinktys leis sudaryti klasifikacinę lentelę ir patikrinti, ar imtyje yra išskirčių.



Pasirenkame **OK**.

10.2.3. Rezultatai

Rezultatų išklotinė prasideda informacija, kad modelis sudaromas tikimybei P(Y = 0):

Model Information

Dependent Variable	Y[a]
Probability Distribution	Binomial
Link Function	Probit

a. The procedure models .00 as the response, treating 1.00 as the reference category.

Lentelėje **Categorical Variable Information** pateikiama informacija apie kategorinių kintamųjų reikšmes. Primename, kad duomenyse negali dominuoti nė viena ir priklausomo kintamojo kategorijų. Nedažnai (Y = 0) studijų žiniomis naudojasi 31,1% respondentų, dažnai (Y = 1) – 68,9% respondentų. Primename, kad minimalus kiekvienos kategorijos procentas turi būti bent 20%. Šis reikalavimas išpildytas. Šiuos procentus prisiminsime, kai tirsime klasifikacinę lentelę. Gerame modelyje, turi būti geresni teisingų klasifikavimų procentai.

Regresoriaus *K35_1* reikšmes verta peržiūrėti tuo pačiu tikslu – jeigu, kai kurių reikšmių nedaug, tai vertėtų regresoriaus kategorijas sustambinti. Matome, kad tokių atsakymų nėra. Regresoriaus transformuoti nereikia.

Categorical Variable Information

			N	Percent
Dependent Variable	Y	.00	100	31.1%
		1.00	222	68.9%
		Total	322	100.0%
Factor	K35_1 |Esamo darbo atitikimas bakalauro (vientisųjų) studijų krypčiai	1 Tikrai taip	153	47.5%
		2 Greičiau taip	95	29.5%
		3 Greičiau ne	36	11.2%
		4 Tikrai ne	38	11.8%
		Total	322	100.0%

Didžiausio tikėtinumo chi kvadrato statistika ir jos p reikšmė yra lentelėje **Omnibus Test**. Matome, p = 0,00...< 0,05. Taigi vienas svarbiausių rodiklių rodo gerą modelio tikimą duomenims.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
163.847	5	.000

Dependent Variable: Y

Model: (Intercept), K35_1, K37_1, K33_2

.

Lentelėje **Parameter Estimates** yra modelio koeficientų įverčiai, tų įverčių 95% pasikliautinieji intervalai ir Voldo kriterijaus reikšmės. Kategorinis kintamasis *K35_1* buvo pakeistas į keturis pseudokintamuosius. Vienas iš šių pseudokintamųjų, atitinkantis reikšmę *K35_1* = 3 yra statistiškai nereikšmingas. Vis dėlto, dėl vienos tokios reikšmės neverta modelį tobulinti.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	4.853	.7092	3.463	6.243	46.832	1	.000
[K35_1=1]	-1.577	.3272	-2.218	-.936	23.229	1	.000
[K35_1=2]	-1.018	.3226	-1.650	-.385	9.953	1	.002
[K35_1=3]	-.261	.3722	-.991	.468	.493	1	.482
[K35_1=4]	0[a]	.	.	.	.	.	.
K37_1	-.273	.1141	-.496	-.049	5.720	1	.017
K33_2	-.780	.1151	-1.005	-.554	45.859	1	.000
(Scale)	1[b]						

Dependent Variable: Y

Model: (Intercept), K35_1, K37_1, K33_2

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Gavome keturis modelius, kurie skiriasi tik konstanta. Juos galim užrašyti taip:

$$\mathrm{P}(Y = 0) = \mathrm{P}(\text{Retai naudoja studijų žinias darbe}) = \Phi(z),$$

$$z = 4{,}85 - 0{,}273\, K37\_1 - 0{,}78\, K33\_2 + \begin{cases} -1{,}57, \text{kai } K35\_1 = 1, \\ -1{,}02, \text{kai } K35\_1 = 2, \\ -0{,}26, \text{kai } K35\_1 = 3. \end{cases}$$

Įsitikinsime, kad modelio koeficientų ženklai atitinka sveiko proto reikalavimus. Koeficientas prie *K37_1* yra neigiamas. Jeigu respondentas renkasi didesnę *K37_1* reikšmę (yra labiau patenkintas darbu), tai mažiau tikėtina, kad jis retai naudoja studijų žinias darbe. Analogiškai darome išvadą, kad kuo labiau suvokia studijų žinių reikšmę, tuo didesnė tikimybė dažnai naudoti tas žinias darbe (mažiau tikėtina, kad tai įvyksta retai). Kuo *K35_1* įgyja didesnę reikšmę (darbo pobūdis labiau atitinka studijų kryptį), tuo labiau tikėtina, kad žinias darbe naudos dažnai. Trumpai kalbant, respondentai labiau vertinantys studijų žinias ir labiau patenkinti savo darbu, dažniau studijų žinias naudoja savo darbe. Ir ši tendencija stipresnė tiems, kurių darbo pobūdis labiau atitinka studijų kryptį.

Kadangi probit regresiją tyrėme, kaip apibendrintąjį tiesinį modelį, tai lentelėje **Goodness of Fit** galima peržiūrėti deviacijos ir laisvės laipsnių santykį. Gerai, kai jis arti vieneto. Matome, kad jis toks ir yra (1,156). Pažymėtina, kad probit regresijai svarbesnis rodiklis yra didžiausio tikėtinumo chi kvadrato maža *p* reikšmė. Jeigu deviacijos ir laisvės laipsnių santykis daug skiriasi nuo vieneto, bet visos kitos charakteristikos (įskaitant klasifikacinę lentelę) geros, tai modelis laikytinas geru.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	49.722	43	1.156
Scaled Deviance	49.722	43	
Pearson Chi-Square	48.218	43	1.121
Scaled Pearson Chi-Square	48.218	43	
Log Likelihood[a]	-47.932		
Akaike's Information Criterion (AIC)	107.865		
Finite Sample Corrected AIC (AICC)	108.131		
Bayesian Information Criterion (BIC)	130.512		
Consistent AIC (CAIC)	136.512		

Dependent Variable: Y

Model: (Intercept), K35_1, K37_1, K33_2

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Norint nustatyti, ar imtyje nėra išskirčių reikia patikrinti, ar duomenyse (ne rezultatų išklotinėje) atsiradusio stulpelio *CooksDistance* visos reikšmės mažesnės už vienetą. Panaudojame komandą **Analyze → Descriptive Statistics → Descriptives.** Perkeliame *CooksDistance* į **Variable(s)**. Pasirenkame **OK.**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CooksDistance Cook's Distance	322	.000	.039	.00324	.006749
Valid N (listwise)	322				

Matome, kad maksimali Kuko mato reikšmė 0,039 < 1. Todėl darome išvadą, kad duomenyse išskirčių nėra.

Norėdami gauti klasifikacinę lentelę, renkamės **Analyze** → **Descriptive Statistics** → **Crosstabs.** Į **Row(s)** perkeliame *Y,* į **Column(s)** – *PredictedValue.* Pasirenkame **Cells** ir pažymime **Row.** Renkamės **Continue** ir **OK.**

Y * PredictedValue Predicted Category Value Crosstabulation

			PredictedValue Predicted Category Value		
			.00	1.00	Total
Y	.00	Count	66	34	100
		% within Y	66.0%	34.0%	100.0%
	1.00	Count	17	205	222
		% within Y	7.7%	92.3%	100.0%
Total		Count	83	239	322
		% within Y	25.8%	74.2%	100.0%

Iš 100 respondentų, kurie retai naudojasi studijų metu gautomis žiniomis, teisingai klasifikuoti 66 (66%). Iš 222 respondentų, kurie dažnai naudojasi studijų metu gautomis žiniomis, teisingai klasifikuoti 205 (92,3%). Prisiminę lentelės **Categorical Variable Information** procentus (atitinkamai 31,1% ir 68,9%), įsitikiname, kad taikomas modelis geras.

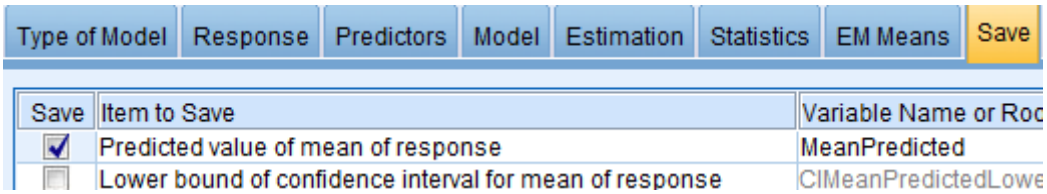
Galutinė išvada: probit regresijos modelis gerai tinka duomenim.

10.2.4. Prognozavimas

Vieną konkrečią reikšmę galima prognozuoti ir su SPSS. Tarkime norime ištirti, ką apie studijų naudojimą mąsto respondentai, kuriems *K33_2* = 4, *K35_1* = 1, *K37_1* = 4. Duomenyse įvedame papildomą eilutę: stulpelyje *K33_2* įrašome 4, stulpelyje *K35_1* įrašome 1, stulpelyje *K37_1* įrašome 4, stulpelyje *filter_$* įrašome 1, kitus stulpelius paliekame tuščius. Jeigu išmestume kitus kintamuosius, tai paskutinė duomenų eilutė atrodys taip:

K33_2	K35_1	K37_1	filter_$	Y	var
4	1	4	1	.	

Vykdydami probit analizę lange **Save** pažymime **Predicted value of mean of response.**



Duomenyse atsiranda papildomas stulpelis **MeanPredicted**, kuriame įrašytas tikimybės P(Y = 0) = P(Retai naudoja studijų žinias) įvertis. Gavome, kad ši tikimybė lygi 0,175. Taigi labai mažai tikėtina, kad toks respondentas retai tnaudojasi studijų metais įgytomis žiniomis. Daug tikėtiniau (tikimybė lygi 1 - 0,175 = 0,825), kad jis tomis žiniomis naudojasi dažnai.

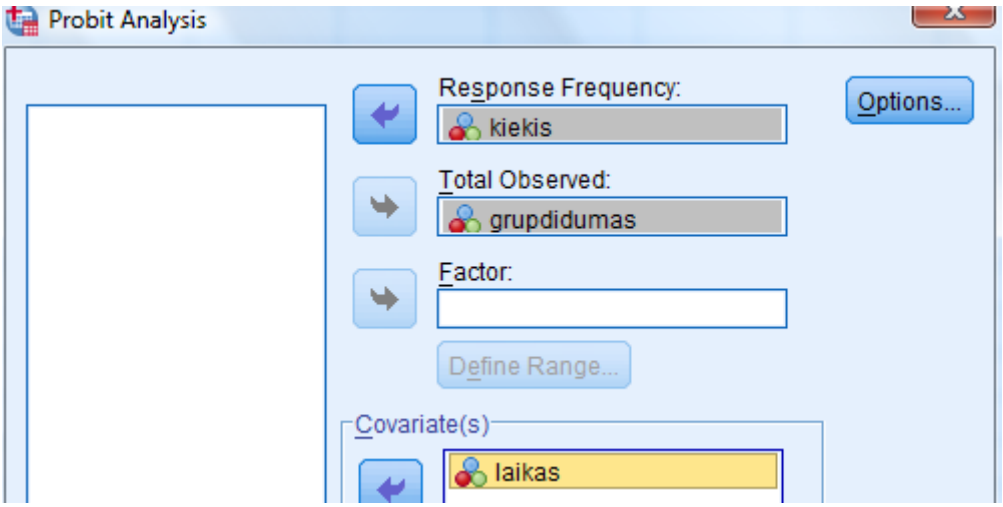
K33_2	K35_1	K37_1	filter_$	Y	MeanPredicted	PredictedValue
4	1	4	1	.	.175	1.00

10.2.5. Dozavimo modelio tyrimas

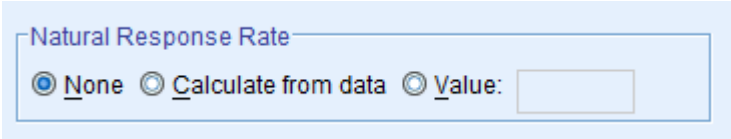
Tarkime, kad dešimt grupių po dešimt savanorių buvo skirtingą laiką agituojami vykti pakariauti į vieną Afrikos valstybę. Stulpelyje *grupdidumas* nurodyta, kiek kiekvienoje grupėje buvo savanorių, stulpelyje *laikas* – kiek laiko jiems buvo „plaunamos smegenys“, stulpelyje *kiekis* – kiek agitacijai pasidavė. Šiaip jau siekiama, kad visose grupėse būtų panašus resspondentų skaičius. Vis dėlto, grupių didumai nebūtinai turi sutapti. Pavyzdžiui, dešimtojoje grupėje buvo 11 respondentų, o likusiose – po dešimt.

	kiekis	grupdidumas	laikas
1	.00	10.00	2.00
2	.00	10.00	.00
3	1.00	10.00	8.00
4	2.00	10.00	14.00
5	4.00	10.00	17.00
6	5.00	10.00	20.00
7	6.00	10.00	22.00
8	7.00	10.00	25.00
9	8.00	10.00	30.00
10	9.00	11.00	36.00

Su probit regresija sumodeliuosime *kiekio* priklausomybę nuo *laiko*. Pasirenkame **Analyze → Regression → Probit.** Į langelį **Response Frequency** įkeliame *kiekis*, į **Total Observed** – *grupdidumas*, į **Covariates** – *laikas*. Spaudžiame **OK**.



Jeigu turėtume kategorinį regresorių (vieną), jį reikėtų įkelti į langelį **Factor**. Jeigu norėtume logaritmuoti duomenis, tai **Transform** pasirinktume tinkamą transformaciją. Duomenis rekomenduojama logaritmuoti, kai duomenų probit grafikas neprimena tiesės. Mes tarėme, kad visai neagituoti respondentai vykti į Afriką nenorės. Jeigu manytume kitaip, tai pasirinkę **Options** parinktyje **Natural Response Rate** nurodytume norimą skaičių (**Value**) arba, kad įvertintų iš imties duomenų (**Calculate from data**).



Norint išsiaiškinti, ar probit regresija tiko, rezultatų išklotinėje reikia surasti lentelę **Chi-Square Tests.** Joje yra Pirsono chi kvadrato statistikos reikšmė (2,775) ir jos *p* reikšmė (*p* = 0,948). Matome, kad *p* reikšmė daug didesnė už 0,05. Probit regresijos modelio atmesti, kaip netinkamo negalima.

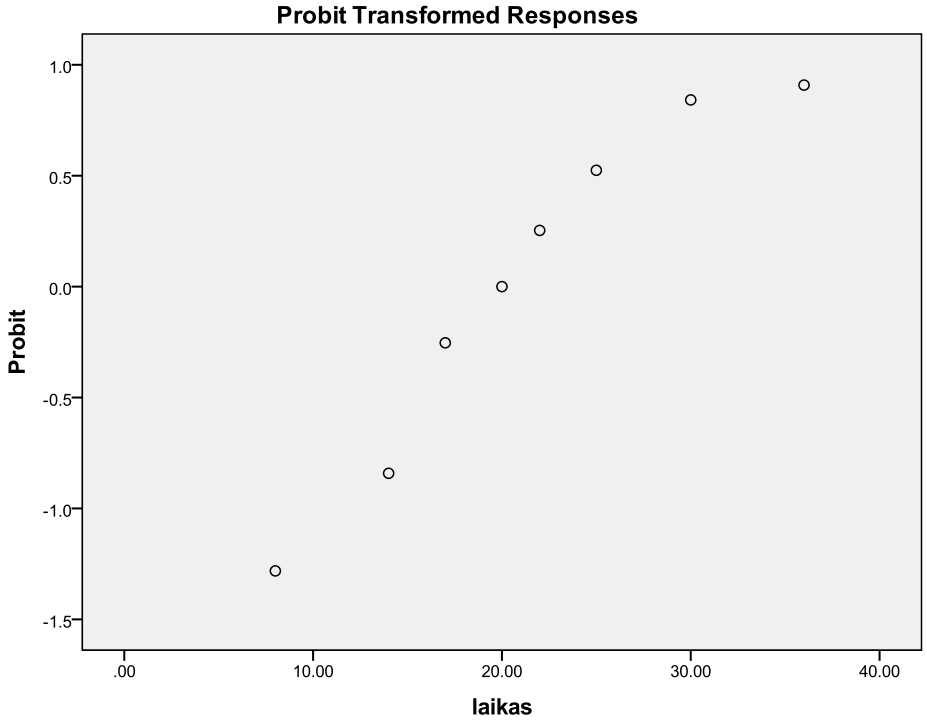
Chi-Square Tests

		Chi-Square	df[a]	Sig.
PROBIT	Pearson Goodness-of-Fit Test	2.775	8	.948[b]

a. Statistics based on individual cases differ from statistics based on aggregated cases.

b. Since the significance level is greater than .150, no heterogeneity factor is used in the calculation of confidence limits.

Duomenų probit grafikas pakankamai tiesinis. Tai irgi rodo gerą regresijos modelio tikimą. Darome išvadą, kad probit modelis tinkamai aprašo duomenis.



Modelio koeficientų įverčiai yra lentelėje **Parameter Estimates**.

Parameter Estimates

	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
PROBIT[a]	laikas	.105	.019	5.496	.000	.067	.142
	Intercept	-2.185	.413	-5.286	.000	-2.598	-1.772

a. PROBIT model: PROBIT(p) = Intercept + BX

Modelis atrodo taip:

$$\mathrm{P(norės\ į\ Afriką)} = \Phi(\mathrm{z}), \quad \mathrm{z} = -2{,}185 + 0{,}105 \times \mathit{laikas}.$$

Kuo ilgesnė įtikinėjimo trukmė, tuo didesnė tikimybė įtikinti.

Lentelėje **Confidence Limits** yra informacija apie *laiko* kiekį, reikalingą atitinkamam procentui respondentų suagituoti. Pavyzdžiui, jeigu norime suagituoti 15% respondentų, vidutinis agitavimo laikas yra 10,447 valandos. Greta dar pateikiamas 95% kintamojo reikšmės intervalas (nuo 3,737 iki 14,237 valandos). Matome, kad pasikliautinis intervalas gana ilgas. Tai nėra labai gerai, nes atsakymas – nuo 3,7 iki 14, 2 valandos – skamba nepraktiškai.

Confidence Limits

	Probability	95% Confidence Limits for laikas		
		Estimate	Lower Bound	Upper Bound
PROBIT	.010	-3.158	-16.832	3.597
	.020	-.283	-12.425	5.786
	.030	1.542	-9.640	7.185
	.040	2.914	-7.551	8.244
	.050	4.030	-5.857	9.110
	.060	4.980	-4.419	9.852
	.070	5.813	-3.162	10.506
	.080	6.559	-2.040	11.095
	.090	7.238	-1.023	11.634
	.100	7.862	-.090	12.133

Jeigu būtų kategorinis regresorius, tai atitinkama procentinė lentelė būtų sudaryta kiekvienai to regresoriaus kategorijai. Pavyzdžiui, jeigu būtume dar grupes skirstę pagal lytį, tai gautume dvi lenteles – vieną vyrams, kitą moterims. Jeigu modelyje būtų ne vienas intervalinis regresosius, tai šios lentelės rezultatų išklotinėje iš viso nebūtų.

Rezultatų išklotinėje dar pateikiama pagalbinė lentelė **Cell Counts and Residuals**. Joje yra informacija apie tai, kiek kiekvienoje grupėje buvo suagituotų savanorių ir kiek jų turėjo būti pagal modelį. Nedidelės prognozės ir stebėjimų skirtumų reikšmės grafoje **Residuals** rodo gerą modelio tikimą duomenims.

Cell Counts and Residuals

	Number	laikas	Number of Subjects	Observed Responses	Expected Responses	Residual	Probability
PROBIT	1	2.000	10	0	.331	-.331	.033
	2	.000	10	0	.213	-.213	.021
	3	8.000	10	1	1.023	-.023	.102
	4	14.000	10	2	2.421	-.421	.242
	5	17.000	10	4	3.390	.610	.339
	6	20.000	10	5	4.480	.520	.448
	7	22.000	10	6	5.235	.765	.523
	8	25.000	10	7	6.343	.657	.634
	9	30.000	10	8	7.931	.069	.793
	10	36.000	11	9	10.089	-1.089	.917

10.2.6. SPSS komandų sintaksė

Surašysime, kaip atrodo SPSS komandos, nagrinėtajam pradiniam logistinės regresijos modeliui. Jas galima naudoti sintaksės lange, pasirenkant opciją **Run**.

```
* Generalized Linear Models.
GENLIN Y (REFERENCE=LAST) BY K35_1 (ORDER=ASCENDING) WITH K37_1 K33_2
  /MODEL K35_1 K37_1 K33_2 INTERCEPT=YES
 DISTRIBUTION=BINOMIAL LINK=PROBIT
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100
MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD)
CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
  /SAVE PREDVAL COOK.
```

SPSS sintaksė dozavimo modeliui atrodo taip:

```
PROBIT kiekis OF grupdidumas WITH laikas
  /LOG NONE
  /MODEL PROBIT
  /PRINT FREQ CI
  /CRITERIA P(.15) ITERATE(20) STEPLIMIT(.1).
```

10.3. Probit regresinė analizė su STATA

10.3.1. Duomenys

Pakartosime probit analizę jau atliktą SPSS programa. Tiriame universitetų absolventų integraciją darbo rinkoje, 2009 m. lapkritis – 2010 m. sausis. Failas **LAMS**, kintamieji: *K2* – aukštoji mokykla, *K33_2* – pasiekiant dabartinę padėtį darbo rinkoje studijų metu įgytos žinios (1 - visai nereikšmingos, ..., 5 – labai reikšmingos), *K35_1* – esamo darbo atitkimas bakalauro studijų krypčiai (1 - tikrai taip, 2 – greičiau taip, 3 – greičiau ne, 4 – tikrai ne), *K37_1* – pasitenkinimas esamu

darbu (1 – labai nepat. , ..., 5 – labai pat.) ir *Y*, kuris įgyja dvi reikšmes: 0 – jeigu respondentai nedažnai naudojasi studijų žiniomis (*K36_1* = 1, 2 arba 3) ir 1 – jeigu respondentai dažnai naudojasi studijų žiniomis (*K36_1* = 4 arba 5). Klaipėdos universiteto bakalaurams tirsime modelį:
Y = f(K35_1, K33_2, K37_1).

Visų pirma, pašaliname iš tyrimo respondentus, kurie į klausimus neatsakė (keep if K35_1 <99 & K33_2 <99 & K37_1 < 99). Atsirenkame Klaipėdos universiteto absolventus (keep if K2 == 3).

Pasižiūrime, kokias reikšmes įgyja *Y*. Įvykdome:

```
tabulate Y, nolabel
```

Y	Freq.	Percent	Cum.
0	100	31.06	31.06
1	222	68.94	100.00
Total	322	100.00	

10.3.2. Modelio tyrimas

Tirsime modelį *Y = f(K35_1, K33_2, K37_1).* Modelyje yra vienas kategorinis regresorius *K35_1*. Įvykdome komandą:

```
xi: probit Y i.K35_1 K33_2 K37_1
```

```
Probit regression                                 Number of obs   =        322
                                                  LR chi2(5)      =     163.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -117.57071                       Pseudo R2       =     0.4107
```

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_IK35_1_2	-.5592151	.2150986	-2.60	0.009	-.9808007	-.1376296
_IK35_1_3	-1.315548	.2857085	-4.60	0.000	-1.875527	-.7555698
_IK35_1_4	-1.576967	.3271906	-4.82	0.000	-2.218249	-.9356852
K33_2	.779756	.1151439	6.77	0.000	.5540781	1.005434
K37_1	.2727866	.1140529	2.39	0.017	.049247	.4963261
_cons	-3.276152	.6937848	-4.72	0.000	-4.635945	-1.916358

Regresorius *K35_1* buvo kategorinis, įgyjantis keturias reikšmes. Todėl jis modelyje keičiamas trimis dichotominiais regresoriais, kurie įtraukiami į kintamųjų sąrašą. Kintamasis *_IK35_1_2* atitinka *K35_1* = 2, *_IK35_1_3* atitinka *K35_1* = 3, *_IK35_1_4* atitinka *K35_1* = 4.

Didžiausio tikėtinumo Chi kvadrato kriterijaus statistika ir *p* reikšmė pateikiamos lentelės viršuje dešinėje pusėje. Matome, kad chi kvadrato statistka yra didelė (lygi 163,85) ir statistiškai reikšminga (jos *p* reikšmė 0,000..). Faktiškai tai informacija, kad su pasirinktais regresoriais

modelis gerai tinka duomenims, nors neaišku, ar visi kintamieji modelyje reikalingi. Kuo chi kvadrato reikšmė didesnė, tuo labiau tikėtina, kad logistinės regresijos modelis tinka. Ji ypač informatyvi tada, kai turime labai daug duomenų. Tada net ir statistiškai reikšminga, bet maža chi kvadrato reikšmė, kelia įtarimų dėl modelio gerumo.

Determinacijos pseudokoeficiento reikšmė pakankama (**Pseudo R2 = 0.41**) ir gerokai viršija 0,20 – tradicinę determinacijos koeficientams taikomą tinkamumo ribą. STATA pateikia Makfadeno determinacijos pseudokoeficientą.

Lentelėje yra modelio koeficientų įverčiai, tų įverčių 95% pasikliautinieji intervalai ir Voldo kriterijaus reikšmės.

STATA probit regresija sudaroma tikimybei Y = 1.

Taigi, sudarysime modelį dažnam studijų žinių naudojimui darbe. Norėdami gauti modelį retam žinių naudojimui, turėtume tiesiog perkoduoti *Y* (pakeisti 0 į 1, o 1 – į 0).

Gavome keturis modelius, kurie skiriasi tik konstanta. Juos galim užrašyti taip:

$$\mathrm{P}(\text{Dažnai naudoja studijų žinias darbe}) = \Phi(z),$$

$$z = -3{,}28 + 0{,}27\,K37\_1 + 0{,}78\,K33\_2 + \begin{cases} -0{,}56, \text{kai } K35\_1 = 2, \\ -1{,}32, \text{kai } K35\_1 = 3, \\ -1{,}58, \text{kai } K35\_1 = 4. \end{cases}$$

Įsitikinsime, kad modelio koeficientų ženklai atitinka sveiko proto reikalavimus. Koeficientas prie *K37_1* yra teigiamas. Jeigu respondentas renkasi didesnę *K37_1* reikšmę (yra labiau patenkintas darbu), tai labiau tikėtina, kad jis dažnai naudoja studijų žinias darbe. Analogiškai darome išvadą, kad kuo labiau suvokia studijų žinių reikšmę, tuo didesnė tikimybė dažnai naudoti tas žinias darbe.

Norėdami gauti klasifikacinę lentelę, įvykdome komandą:

```
estat clas
```

```
Probit model for Y

                 -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       205            34  |        239
     -     |        17            66  |         83
-----------+--------------------------+-----------
   Total   |       222           100  |        322

Classified + if predicted Pr(D) >= .5
True D defined as Y != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   92.34%
Specificity                     Pr( -|~D)   66.00%
Positive predictive value       Pr( D| +)   85.77%
Negative predictive value       Pr(~D| -)   79.52%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   34.00%
False - rate for true D         Pr( -| D)    7.66%
False + rate for classified +   Pr(~D| +)   14.23%
False - rate for classified -   Pr( D| -)   20.48%
--------------------------------------------------
Correctly classified                        84.16%
--------------------------------------------------
```

D kategorija žymi dažną žinių naudojimą ($Y = 1$). Matome, kad iš 222 tokių respondentų teisingai atpažinti 205. Tai sudaro 92,34 % teisingų klasifikavimų. Analogiškai nustatome, kad iš 100 retai naudojančių studijų žinias teisingai klasifikuoti 66 (66%). Bendrasis teisingų klasifikavimų procentas yra 84,16%. Tai rodo gerą modelio tikimą duomenims.

Galima papildomai suskaičiuoti *Pirsono chi kvadrato statistika.* Jeigu p reikšmė $\geq 0{,}05$, tai rodo gerą modelio tikimą. Tuo įsitikiname, įvykdę :

```
estat gof
```

```
Probit model for Naudbak, goodness-of-fit test

       number of observations =       322
 number of covariate patterns =        49
             Pearson chi2(43) =        48.22
                  Prob > chi2 =         0.2700
```

Norėdami gauti informacinių kriterijų AIC ir BIC reikšmes (jie taikomi, kai lyginami du modeliai. Geresnis tas modelis, kurio AIC ir BIC mažesni), įvykdome

Norėdami gauti klasifikacinę lentelę, įvykdome komandą:

```
fitstat
```

```
Measures of Fit for probit of Naudbak

Log-Lik Intercept Only:       -199.494   Log-Lik Full Model:           -117.571
D(316):                        235.141   LR(5):                         163.847
                                         Prob > LR:                       0.000
McFadden's R2:                   0.411   McFadden's Adj R2:               0.381
ML (Cox-Snell) R2:               0.399   Cragg-Uhler(Nagelkerke) R2:      0.561
McKelvey & Zavoina's R2:         0.600   Efron's R2:                      0.457
Variance of y*:                  2.501   Variance of error:               1.000
Count R2:                        0.842   Adj Count R2:                    0.490
AIC:                             0.768   AIC*n:                         247.141
BIC:                         -1589.617   BIC':                         -134.974
BIC used by Stata:             269.789   AIC used by Stata:             247.141
```

Matome, kad ir visi pseudodeterminacijos koeficientai rodo pakankamai gerą modelio tikimą duomenims.

10.3.3. Reikšmių prognozavimas

Tarkime, kad norime prognozuoti *Y,* žinodami, kad jam(jai) *K33_2* = 4, *K37_1* = 4, *K35_1* = 2. Tą padaryti galima, įvykdžius komandą:

```
adjust K33_2 = 4  K37_1=4   _IK35_1_2=1,pr
```

All	pr
	.5166

Key: pr = Probability

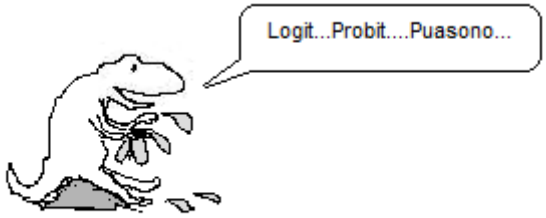
Tikimybė, kad respondentas dažnai naudoja studijų metu įgytas žinias lygi 0,516.

Uždaviniai

1. Naudojant probit regresiją, išspręsti dvinarės logistinės regresijos skyrelio uždavinius. Palyginti gautų modelių klasifikacines lenteles su logistinės regresijos klasifikacinėmis lentelėmis.

2. Sudaryti ir ištirti dozavimo modelį. Buvo tirta dviejų rūšių bičių atsparumas augalų cheminiam užterštumui. Kiekvienos rūšies bitės buvo apipurkštos skirtingos koncentracijos Mechiko smogo koncentratu. Fiksuota, kiek bičių neišgyveno pusvalandžio po apipurškimo.

Kiekis	Iš kiek	Koncentracija	Grupė
0	20	2	1
0	20	0	1
1	20	7	1
2	20	13	1
4	20	16	1
5	20	19	1
6	20	21	1
7	20	24	1
8	20	27	1
10	20	29	1
2	20	2	2
0	20	0	2
3	20	7	2
5	20	13	2
7	20	16	2
7	20	19	2
9	20	21	2
10	20	24	2
12	20	27	2
13	20	30	2



Kaip pasirinkti tinkamą regresijos modelį.

11. LOGTIESINIAI MODELIAI

Kai mus domina dviejų kategorinių požymių priklausomybė, taikome chi kvadrato homogeniškumo kriterijų ir tiriame procentinių sudėčių lentelę. Logtiesiniai modeliai yra sudėtingesnis tokios analizės atvejis, kai kategorinių požymių yra daugiau nei du. Be to, net dviems kategoriniams požymiaims galima įvertinti jų tarpusavio sąveiką. Tipiniai logtiesinių modelių taikymo pavyzdžiai:

- Norime išsiaiškinti, ar kaip susiję antsvoris, lytis ir padidėjęs kraujo spaudimas (viršutinis kraujospūdis > 140 ir $\leq$ 140).
- Tiriame kaip susiję laimingumas, šeiminė padėtis ir amžius.
- Aiškinamės kaip susijusi lytis, profesija ir sėkminga karjera.

Priešingai iki šiol nagrinėtiems modeliams logtiesiniuose modeliuose nėra priklausomo kintamojo ir regresorių. Faktiškai – tai tik kintamųjų tarpusavio priklausomybių analizė. Jeigu tyrimo tikslas nustatyti, kaip vieną kintamąjį veikia visi kiti, tai geriau taikyti logistinę regresiją, o ne logtiesinį modelį.

11.1. Logtiesinio modelio struktūra ir savybės

11.1.1. Modelio lygtis

Įsivaizduokime, kad pagal du požymius A ir B suskirstėme respondentus į grupes. Modeliuojamas respondentų *skaičius* kiekvienoje grupėje (tiksliau skaičiaus logaritmas – iš čia ir pavadinimas kilęs). Sąlyginė modelio schema atrodo taip:

$$\ln(\text{respondentų skaičius grupėje}) = \mu + \lambda_A + \lambda_B + \lambda_{A\times B}\,.$$

Čia μ žymi bendrą vidutinį respondentų skaičių grupėje, $\lambda_A$atspindi požymio A įtaką, λ_B atspindi požymio B įtaką, λ_{AB} nurodo požymių A ir B tarpusavio sąveiką. Norint, kad modelio užrašas būtų neformalus, reikėtų visus λ keisti keliais dichotominiais kintamaisiais (jų būtų tiek, kiek požymiai turi kategorijų). Užrašytasis modelis vadinamas *pilnuoju* (angl. saturated), nes jame nurodytos visos galimos kintamųjų sąveikos. Pilnasis modelis visiškai tiksliai aprašo kintmųjų priklausomybes. Logtiesinio modeliavimo tikslas – surasti paprastesnės struktūros modelį, kuris kintamųjų priklausomybes aprašytų beveik taip gerai, kaip ir pilnasis modelis.

Geras logtiesinis modelis nuo pilnojo modelio statistiškai reikšmingai nesiskiria, bet yra paprastesnės struktūros.

Pavyzdžiui, jeigu tiktų modelis $\ln(\text{respondentų skaičius grupėje}) = \mu + \lambda_{A\times B}$, tai darytume išvadą, kad patys po vieną požymiai nėra itin svarbūs, o svarbi tik jų sąveika.

Kai kintamųjų daug (dalis jų gali būti ir intervaliniai), pilnieji modeliai tampa labai gremėzdiški. Pavyzdžiui, pilnasis logtiesinis modelis, kai yra trys kategoriniai požymiai, atrodo taip:

$$\boxed{\ln(\text{skaičius grupėje}) = \mu + \lambda_A + \lambda_B + \lambda_C + \lambda_{A\times B} + \lambda_{A\times C} + \lambda_{B\times C} + \lambda_{A\times B\times C}\,.}$$

Todėl sudarytas modelis turėtų būti paprastesnis. Matematiškai respondentų skaičius grupėje modeliuojamas, laikant jį Puasono atsitiktiniu dydžiu. Faktiškai gauname Puasono regresijos analogą, kai respondentų skaičius grupėje užima priklausomo kintamojo vietą.

11.1.2. Rezultatų interpretavimas

Formaliai sudaryti duomenims tinkamą modelį nėra taip sudėtinga. Juolab, kad galima naudoti žingsninį sudarymo metodą – jis šiuo atveju vadinamas hierarchiniu modeliavimu. Problemos kyla, kai tenka gautuosius rezultatus interpretuoti. Faktiškai žiūrima kokie ryšiai liko modelyje. Tarkime, kad tiriame požymius A, B, C. Pilnasis modelis šiai situacijai užrašytas ankstesniame skyrelyje.

Jeigu gerai duomenis aprašo modelis

$$\ln(\text{skaičius grupėje}) = \mu + \lambda_A + \lambda_B + \lambda_C\,,$$

tai darome išvadą, kad visi požymiai (A, B, C) yra nepriklausomi.

Jeigu tinka modelis

$$\ln(\text{skaičius grupėje}) = \mu + \lambda_C + \lambda_{A\times B}\,,$$

tai darome išvadą, kad C nepriklauso nuo A ir B, o A ir B susiję.

Jeigu tinka modelis

$$\ln(\text{skaičius grupėje}) = \mu + \lambda_{A\times C} + \lambda_{B\times C},$$

tai darome išvadą, kad A susijęs su B, B susijęs su C, bet A ir B nepriklausomi, *jeigu atsižvelgsime* į C. Tai vadinamsis *sąlyginis* požymių nepriklausomumas. Beje, taikydami chi kvadratą A ir B požymiams galime gauti, kad jie priklausomi. Analogiškai interpretuojami visi kiti modeliai.

Po to, kai nustatėme kintamųjų sąsajas norisi išsiaiškinti, kokios yra tos priklausomybės. Pats lengviausias būdas tą atlikti – surašyti atitinkamas dažnių lenteles. Pavyzdžiui, jeigu tirdami rūkymo (A), lyties (B) ir antsvorio (C) priklausomybę, gavome modelį, kuriame yra $\lambda_{A\times B\times C}$, tai galima pateikti ir aprašyti procentines porinių dažnių AxC lenteles vyrams ir moterims. Arba nurodyti kiek procentų rūkančių vyrų turi antsvorio, kiek procentų rūkančių moterų turi antsvorio ir t.t. Taigi pats aprašymas priklauso nuo tyrėjo.

11.1.3. Galimybių santykis

Kaip ir logistinėje regresijoje, logtiesiniuose modeliuose galima apibrėžti galimybių santykį. Kadangi nagrinėjami tik kategoriniai kintamieji (požymiai), tai pradžioje nusakysime, kas yra galimybių santykis porinių dažnių lentelei. Tarkime, kad visus respondentus suskirstėme pagal rūkymą ir patirtą infarktą.

	Infarktas	
	Taip	Ne
Rūko	a	b
Nerūko	c	d

Tikimybė, kad rūkantis respondentas patirs infarktą yra a/ (a+b). kad nepatirs – b/(a+b). Galimybė (tikimybių santykis) rūkantiems patirti infarktą yra a/b. Analogiškai gauname, kad galimybė patirti infarktą nerūkantiems yra c/d. Galimybių santykis (angl. odds ratio) yra

$$\frac{a}{b} : \frac{c}{d} = \frac{ad}{bc}.$$

Darysime išvadą, kad rūkymas padidina infarkto galimybę ad/bc karto.

Jeigu kintamieji yra nepriklausomi, tai galimybių santykis yra lygus vienetui.

Kuo galimybių santykis labiau skiriasi nuo vieneto, tuo priklausomybė stipresnė. Galimybių santykis 0,2 rodo stipresnę kintamųjų priklausomybę, nei galimybių santykis lygus 0,8.

Panašiai interpretuojami galimybių santykiai ir logtiesiniuose modeliuose. Nors tam galima naudoti modelio parametrų įverčių eksponentes, tačiau praktikoje tiesiog atsižvelgiama į modelio kintamuosius ir tikrinamos atitinkamos lentelės. Tarkime, kad turime tokį modelį:

$$\ln(\text{skaičius grupėje}) = \mu + \lambda_{A\times C} + \lambda_{B\times C} + \lambda_{A\times B\times C}.$$

Tada galimybių santykį požymiams AxB, reikia skaičiuoti visoms C kategorijoms atskirai. Jeigu modelyje yra tik

$$\ln(\text{skaičius grupėje}) = \mu + \lambda_{A\times C} + \lambda_{B\times C},$$

Tai sąlyginis galimybių santykis AxB yra lygus 1 (kai fiksuojame C kategoriją). Jegu modelyje yra tik antros eilės sąveikos (AxB, AxC, BxC), ir nėra trečios eilės sąveikos AxBxC, tai galimybių santykiai AxB yra tie patys visoms C kategorijoms. Tas pats galioja kitiems galimybių santykiams.

11.1.4. Reikalavimai duomenims

Duomenys turi tenkinti tokius reikalavimus:

1) *Kintamieji yra kategoriniai arba intervaliniai.* Kaip taisyklė visi kintamieji yra kategoriniai. Būtent jų priklausomybes siekiame nustatyti. Intervaliniai kintamieji turi prasmę, jeigu labai domimės kaip suskaičiuoti, kiek respondentų pateks į konkrečią grupę.
2) *Stebėjimai turi būti nepriklausomi.* Tiksliau kalbant, respondentų skaičius vienoje grupėje nepriklauso nuo to, kiek respondentų pakliuvo į kitą grupę. Jeigu tyrime buvo 17 rūkančių vyrų, tai neturi įtakos nerūkančių moterų skaičiui.
3) *Duomenų turi būti pakankamai daug.* Jeigu jau suskirstėme respondentus pagal kelis požymius, tai kiekvienoje grupėje turi būti atstovų. Žmonėms mėgstantiems tikslias instrukcijas (pvz., kiek gramų druskos dėti į litrą kopūstų sriubos) nurodome, kad stebėjimų turi būti bent penkis kartus daugiau, nei yra grupių. Jei taip nėra, tai eikite gaudyti naujų respondentų, arba stambinkite kintamųjų kategorijas. Pavyzdžiui, jeigu respondentus skirstome pagal tai, ar jie rūko, ar turi antsvorio ir lytį, tai skirtingų grupių bus 8, o repondentų reikia bent 40[18]. Dažniausiai reikalaujama, kad tarp stebimų *Y* reikšmių vienetų (nulių) būtų ne mažiau penktadalio.
4) Jeigu modelyje yra daug kategorinių kintamųjų, tai *kiekvienam tų kategorijų deriniui turi būti pakankamai daug respondentų* (bent po 5).
5) Regresoriai neturi labai stipriai tarpusavyje koreliuoti (neturi būti regresorių multi-kolinearumo).

11.1.5. Modelio tinkamumas

Bendruoju atveju modelio tikimą duomenims parodo tokie rodikliai:

1) *Didžiausio tikėtinumo chi kvadrato statistika*. Parodo, ar modelis statistiškai reikšmingai skiriasi nuo pilnojo modelio. Gerai, jeigu nesiskiria, t.y., jeigu statitiskos *p* reikšmė didesnė už 0,05.
2) *Pirsono chi kvadrato statistika*. Parodo, ar modelis statistiškai reikšmingai skiriasi nuo pilnojo modelio. Gerai, jeigu nesiskiria, t.y., jeigu statitiskos *p* reikšmė didesnė už 0,05.
3) *Voldo testai kintamiesiems*. Padeda nuspręsti ar kintamasis šalintinas iš modelio. Jeigu konkretaus testo *p* reikšmė < 0,05, tai dažniausiai jį modelyje paliekame. Kategoriniams kintamiesiems statistinis reikšmingumas tikrinamas visiems kategorijų reikšmių deriniams

[18] Ne tiksliai 40, o bent 40. Visai puikiai tiks ir 92 tiriamieji. Arba 103.

(pvz., A= 1, B = 2). Dažnai keli deriniai gali būti nereikšmingi, o keli reikšmingi. Ar tada šalinti kintamąjį iš modelio sprendžia pats tyrėjas.

4) *Koreguotų liekanų Q-Q grafikas.* Jis turi priminti tiesę.
5) *Koreguotų liekanų Q-Q grafikas, pašalinus trendą.* Turi nerodyti jokio reguliaraus liekamųjų paklaidų elgesio – taškai turi būti išsibarstę atsitiktinai.

Paskutinieji du grafikai siejami su noru (nelabai pagrįstu), kad modelio liekanos elgtųsi analogiškai normaliajam skirstiniui. Kadangi visi kintamieji kategoriniai, tai nereikėtų labai šių rodiklių sureikšminti. Modelis tobulinamas, šalinant statistiškai nereikšmingus kintamuosius.

Labai gerai duomenims tinkančiame modelyje:

- Didžiausio tikėtinumo chi kvadrato $p \geq 0,05$.
- Pirsono chi kvadrato $p \geq 0,05$.
- Visi kintamieji statistiškai reikšmingi. Jiems Voldo kriteirjuas $p < 0,05$.
- Koreguotų liekanų Q-Q grafikas primena įstrižainę.
- Koreguotų liekanų Q-Q grafikas be trendo nerodo jokio reguliarumo.

11.2. Logtiesinių modelių tyrimas su SPSS

11.2.1. Duomenys

Kaip konstruoti logtiesinius modelius, parodysime, tirdami moterų ir vyrų požiūrius į lyčių santykius. Failas **ZTLT1,** kintamieji:

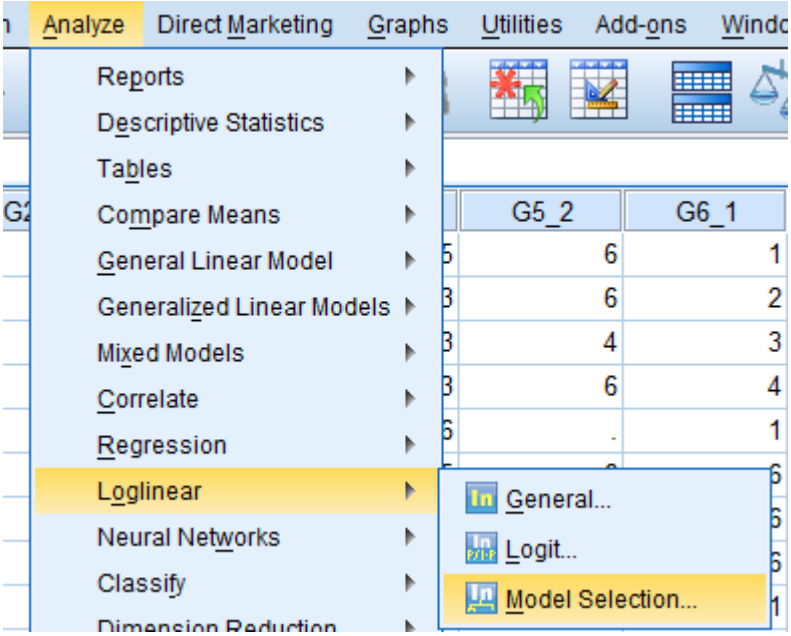
- *S1* – lytis (1 – vyr., 2 – mot.)
- *G7_7* – šeimai geriau, kai moteris prižiūri namus, vaikus (1 – sutinku, 2 – nei taip, nei ne, 3 – nesutinku),
- *G16_1* – sunku suderinti šeimą ir verslą (1 - trukdo verslui, 2 – netrukdo, 3 – nežinau),
- *G13* – noras dirbti pagal lankstų darbo grafiką (1 – taip, 2 – ne , 3 – sunku pasakyti, 4 – nedirbu).

Tirsime, kaip dirbantiems respondentams(ėms) susiję *S1, G7_7* ir *G16_1*. Tirsime tik tuos, kurie turi nuomonę apie tai, ar šeima trukdo verslui. Su **Select Cases** atrenkame respondentus, kuriems *G13* < 4 & *G16_1* < 3.

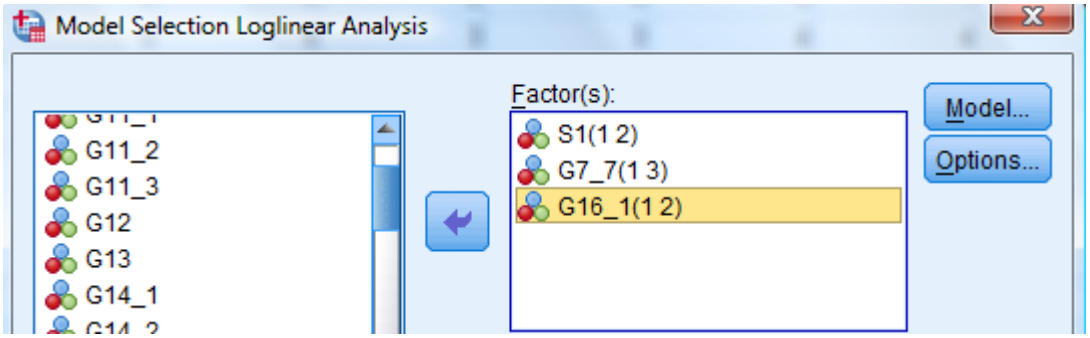
Kadangi išankstinės nuomonės apie modelio savybes neturime, tai atliksime hierarchinį logtiesinį modeliavimą, t.y. modelį sudarysime automatiškai pagal formalius statistinio reikšmingumo požymius.

11.2.2. SPSS parinktys hierarchiniam logtiesiniam modeliui

Atsidarome langelį **Analyze** ir renkamės **Loglinear → Model Selection.**



Atsidariusiame meniu renkamės sukeliame kintamuosius *S1, G16_1* ir *G7_7* į langą **Factors.** Nurodome kiekvieno kintamojo reikšmes. Spaudžiame **OK.**



Rezultatų išklotinė pradedama informacija, kad pradžioje tiriamas pilnasis modelis. Generuojantis reiškinys (*S1*G7_7*G16_1*) nurodo didžiausią tarpusavio sąveiką modelyje.

Convergence Information

Generating Class	S1*G7_7*G16_1
Number of Iterations	1
Max. Difference between Observed and Fitted Marginals	.000
Convergence Criterion	.250

Pilnojo modelio tikimas duomenims yra idealus, todėl lentelėje **Goodness-of-Fit Tests** chi kvadrato statistika lygi nuliui, o *p* reikšmė neskaičiuojama. Lentelėje **K-Way and Higher-Order Effects** pateikiama informacija apie tai, kokios eilės kintamųjų priklausomybės yra statistiškai reikšmingos (kelių kintamųjų sandaugos įtrauktinos į modelį). Matome, kad statistiškai nereikšminga yra trečios eilės priklausomybė. Jai *p* reikšmė (p = 0,451) viršija 0,05. Taigi modelyje nereikės visų trijų kintamųjų sandaugos.

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio Chi-Square	Likelihood Ratio Sig.	Pearson Chi-Square	Pearson Sig.	Number of Iterations
K-way and Higher Order Effects[a]	1	11	189.155	.000	238.505	.000	0
	2	7	31.677	.000	31.797	.000	2
	3	2	1.590	.451	1.592	.451	4
K-way Effects[b]	1	4	157.478	.000	206.708	.000	0
	2	5	30.086	.000	30.205	.000	0
	3	2	1.590	.451	1.592	.451	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Lentelėje **Step Summary** nurodyti tirtieji modeliai. Paskutinis modelis statistiškai geriausias. Tai dar nereiškia, kad jis geras tyrimo prasme.

Step Summary

Step[a]		Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]	S1*G7_7*G16_1	.000	0	.	
	Deleted Effect 1	S1*G7_7*G16_1	1.590	2	.451	4
1	Generating Class[b]	S1*G7_7, S1*G16_1, G7_7*G16_1	1.590	2	.451	
	Deleted Effect 1	S1*G7_7	15.261	2	.000	2
	2	S1*G16_1	.109	1	.741	2
	3	G7_7*G16_1	13.346	2	.001	2
2	Generating Class[b]	S1*G7_7, G7_7*G16_1	1.700	3	.637	
	Deleted Effect 1	S1*G7_7	15.946	2	.000	2
	2	G7_7*G16_1	14.031	2	.001	2
3	Generating Class[b]	S1*G7_7, G7_7*G16_1	1.700	3	.637	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than .050.

b. Statistics are displayed for the best model at each step after step 0.

c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Taigi hierarchinė logtiesinė regresija siūlo tirti modelį:

$$\ln(\text{skaičius grupėje}) = \mu + S1*G7\_7 + G16\_1*G7\_7.$$

Lentelėje **Cell Counts and Residuals** nurodyta, kaip šis modelis aprašo kintamųjų elgesį. Visi standartizuotieji liekamieji nariai turi būti maži.

Cell Counts and Residuals

S1 Lytis	G7_7 Šeimai geriau, kai vyras uždirba pinigus, o moteris prižiūri namus,	G16_1 Sunku suderinti šeimą ir verslą	Observed Count	Observed %	Expected Count	Expected %	Residuals	Std. Residuals
1 Vyrai	1 Sutinku	1 Trukdo verslui	106.000	25.7%	102.309	24.8%	3.691	.365
		2 Netrukdo verslui	37.000	9.0%	40.691	9.9%	-3.691	-.579
	2 Nei taip, nei ne	1 Trukdo verslui	29.000	7.0%	29.651	7.2%	-.651	-.120
		2 Netrukdo verslui	22.000	5.3%	21.349	5.2%	.651	.141
	3 Nesutinku	1 Trukdo verslui	12.000	2.9%	13.500	3.3%	-1.500	-.408
		2 Netrukdo verslui	15.000	3.6%	13.500	3.3%	1.500	.408
2 Moterys	1 Sutinku	1 Trukdo verslui	70.000	17.0%	73.691	17.9%	-3.691	-.430
		2 Netrukdo verslui	33.000	8.0%	29.309	7.1%	3.691	.682
	2 Nei taip, nei ne	1 Trukdo verslui	21.000	5.1%	20.349	4.9%	.651	.144
		2 Netrukdo verslui	14.000	3.4%	14.651	3.6%	-.651	-.170
	3 Nesutinku	1 Trukdo verslui	28.000	6.8%	26.500	6.4%	1.500	.291
		2 Netrukdo verslui	25.000	6.1%	26.500	6.4%	-1.500	-.291

Rezultatų išklotinės pabaigoje yra didžiausio tikėtinumo ir Pirsono chi kvadratų statistikų reikšmės ir jų *p* reikšmės. Abidvi *p* reikšmės gerokai viršija 0,05 , todėl galima kalbėti apie tinkamą duomenims logtiesinį modelį. Dabar jo išsamiau neaptarinėsime. Padarysime tai, atlikę tiesioginį modeliavimą kitame skyrelyje.

Goodness-of-Fit Tests

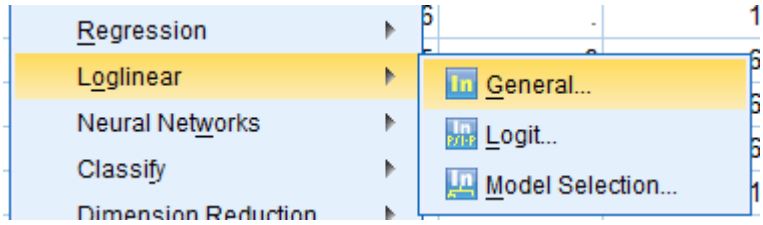
	Chi-Square	df	Sig.
Likelihood Ratio	1.700	3	.637
Pearson	1.705	3	.636

11.2.3. SPSS parinktys logtiesiniam modeliui

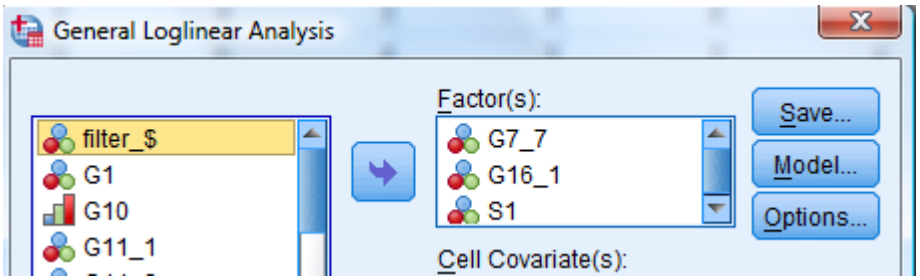
Hierarchiniu modeliavimu nustatėme, koks gali būti tinkamas modelis. Parodysime, kaip reikėtų atlikti analizę, kai spėjama modelio struktūra pasirenkama iš anksto. Pasirinksime tą patį, jau gautą modelį, kurį sąlyginai galime užrašyti taip:

$$\ln(\text{skaičius grupėje}) = \mu + S1*G7\_7 + G16\_1*G7\_7.$$

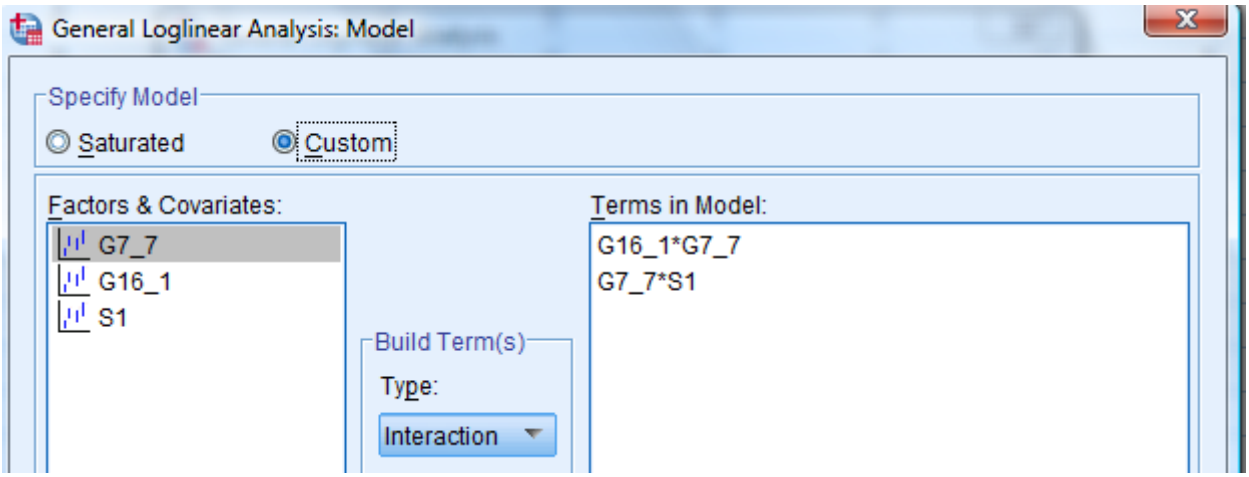
Atsidarome langelį **Analyze** ir renkamės **Loglinear → General.**



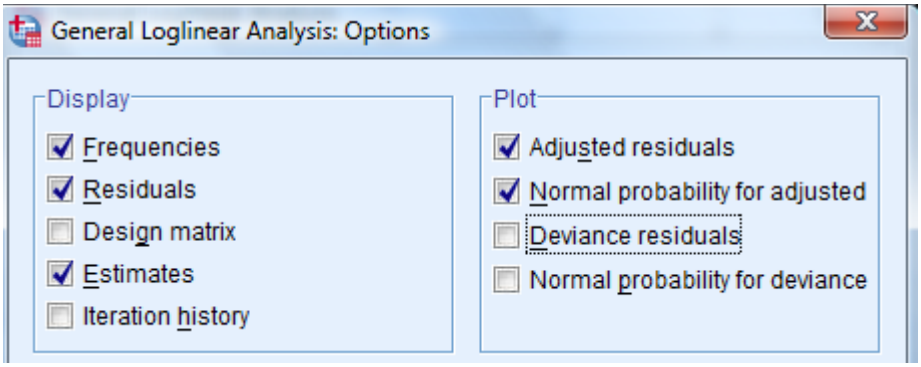
Visus tris kintamuosius įkeliame į langelį **Factor(s)** ir spaudžiame **Model**.



Atsidariusiame lange pažymime **Custom**, pasirenkame **Type: Interaction** ir perkeliame kintamųjų sandaugas. Spaudžiame **Continue**.



Pasirenkame **Options** ir papildomai pažymime **Estimates**. Renkamės **Continue** ir **OK**.



11.2.4. Rezultatai

Rezultatų išklotinė pradedama lentele **Goodness-of-Fit Tests**, kurioje yra didžiausio tikėtinumo ir Pirsono chi kvadratų statistikų reikšmės ir jų *p* reikšmės. Abidvi *p* reikšmės gerokai viršija 0,05 , todėl galima kalbėti apie tinkamą duomenims logtiesinį modelį. Ši lentelė nesiskiria nuo gautosios hierarchiniame logtiesiniame modelyje. Visai analogiška ir lentelė **Cell Counts and Residuals,** todėl jos nebeaptarsime.

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	1.700	3	.637
Pearson Chi-Square	1.705	3	.636

a. Model: Poisson

b. Design: Constant + G7_7 * G16_1 + G7_7 * S1

Lentelėje **Parameter Estimates** yra modelio parametrai ir jų statistinis reikšmingumas. Kadangi visi kintamieji kategoriniai, tai ir informacija pateikiama visiems jų reikšmių deriniams. Negerai, kai daug derinių yra statistiškai nereikšmingi. Šiuo atžvilgiu matome, kad *G7_7*G16_1* yra tik pusiau su bėda priimtinas modelio narys. Galbūt reikėtų pagalvoti apie *G7_7* kategorijų stambinimą.

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	3.277	.177	18.503	.000	2.930	3.624
[G7_7 = 1] * [G16_1 = 1]	1.023	.207	4.950	.000	.618	1.428
[G7_7 = 1] * [G16_1 = 2]	.101	.226	.445	.656	-.343	.545
[G7_7 = 2] * [G16_1 = 1]	-.264	.261	-1.011	.312	-.776	.248
[G7_7 = 2] * [G16_1 = 2]	-.593	.276	-2.148	.032	-1.133	-.052
[G7_7 = 3] * [G16_1 = 1]	5.952E-17	.224	.000	1.000	-.438	.438
[G7_7 = 3] * [G16_1 = 2]	0[a]	.	.	.	.	.
[G7_7 = 1] * [S1 = 1]	.328	.129	2.539	.011	.075	.581
[G7_7 = 1] * [S1 = 2]	0[a]	.	.	.	.	.
[G7_7 = 2] * [S1 = 1]	.376	.219	1.715	.086	-.054	.807
[G7_7 = 2] * [S1 = 2]	0[a]	.	.	.	.	.
[G7_7 = 3] * [S1 = 1]	-.674	.236	-2.853	.004	-1.138	-.211
[G7_7 = 3] * [S1 = 2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

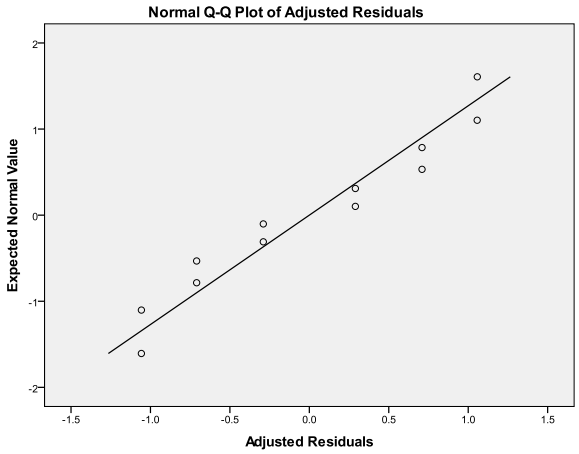
c. Design: Constant + G7_7 * G16_1 + G7_7 * S1

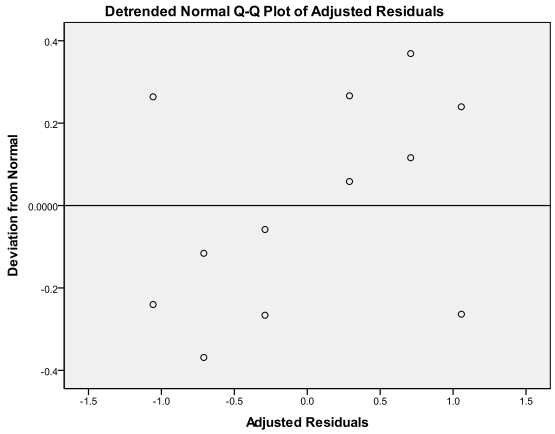
Parametrų įverčius galime panaudoti, jeigu norime sumodeliuoti, koks respondentų skaičius turėtų patekti į kiekvieną grupę. Tarkime respondentų vyrų (*S1* = 1), kurie mano, kad šeimai geriau, kai moteris prižiūri namus (*G7_7* = 1), o šeimą ir verslą suderinti sunku (*G16_1* = 1) skaičius pagal modelį turėtų būti:

skaičius = exp{ 3,27 + 1,02 + 0,328} = 101,29.

Žinoma tokios prognozės vargiai yra prasmingos, nes dar reikia atsižvelgti į tai, kiek tyrime respondentų buvo (412 respondentų).

Koreguotųjų liekamųjų paklaidų grafikas rodo, kad modelis yra pakankamai geras (grafikas primena tiesę). Pašalinus trendą, vaizdas kiek pablogėja (taškai nėra visiškai atsitiktinai išsibarstę), bet jų per mažai, kad įtartume kažkokį rimtą reguliarumą.





Darome išvadą, kad sudarytasis logtiesinis modelis adekvačiai aprašo duomenis.

11.2.5. Išvados

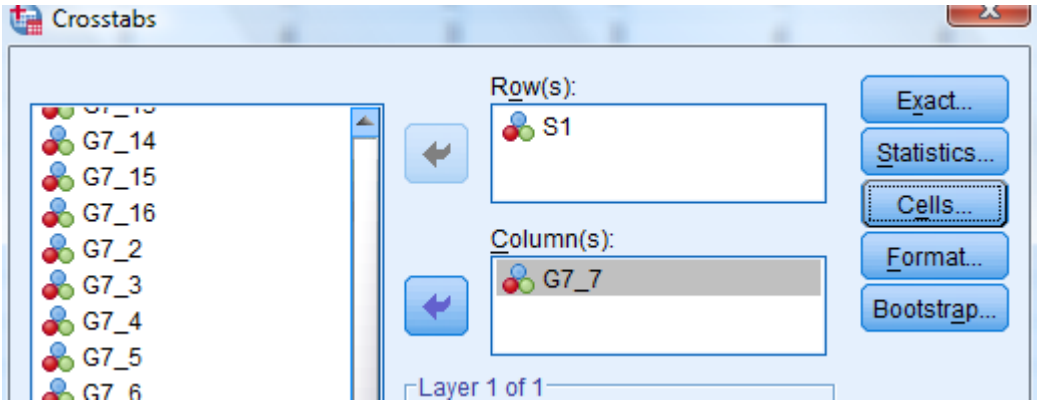
Turėme modelį, kurį sąlyginai galime užrašyti taip:

$$\ln(\text{skaičius grupėje}) = \mu + S1*G7\_7 + G16\_1*G7\_7.$$

Ką rodo šis modelis? Visų pirma, matome, kad kintamieji *S1* (lytis) ir *G16_1* (ar šeima trukdo verslui) yra sąlyginai nepriklausomi, kai atsižvelgiama į *G7_7* (ar moteriai geriau prižiūrėti namus). Kintamieji *S1* ir *G7_7* yra priklausomi. Taip pat priklausomi yra kintamieji *G16_1* ir *G7_7*. Modelyje nėra pirmos eilės priklausomybių, t.y. nėra atskirai įtrauktų kintamųjų *S1, G7_7* ir *G16_1*. Tai rodo, kad po vieną šie kintamieji kažkokios naujos informacijos, lyginant ją su jau esančiomis antros eilės priklausomybėmis neduoda.

Faktiškai rezultatų aprašymas dar turėtų būti papildytas atitinkamomis procentinėmis porinių dažnių lentelėmis. Sakėme, kad *S1* ir *G7_7* yra priklausomi. Atitinkama procentinė lentelė parodo,

kokia ta priklausomybė. Lentelę gausime pasirinkę **Analyze** → **Descriptive Statistics** → **Crosstabs** ir įkėlę *S1* į **Row(s)**, o *G7_7* į **Column(s).** Cells pasirenkame **Row.** Spaudžiame **OK**.



Gautoji procentinė lentelė rodo su požiūriu į tai, kad šeimai geriau, kai moteris prižiūri namus nesutinka 12,2% vyrų ir 27,7% moterų.

S1 * G7_7 Crosstabulation

			G7_7 Šeimai geriau, kai vyras uždirba pinigus, o moteris prižiūri namus, vaikus			Total
			1 Sutinku	2 Nei taip, nei ne	3 Nesutinku	
S1 |Lytis	1 Vyrai	Count	143	51	27	221
		% within S1 |Lytis	64.7%	23.1%	12.2%	100.0%
	2 Moterys	Count	103	35	53	191
		% within S1 |Lytis	53.9%	18.3%	27.7%	100.0%
Total		Count	246	86	80	412
		% within S1 |Lytis	59.7%	20.9%	19.4%	100.0%

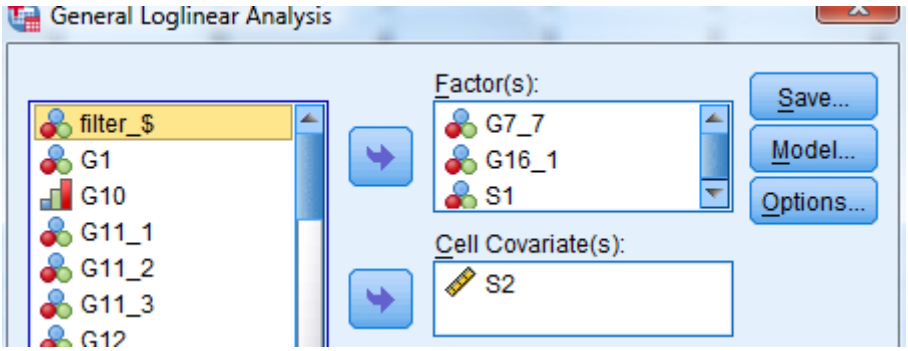
Analogiškai reikėtų ištirti ir aprašyti priklausomybę *G7_7*G16_1*. Galimybių santykių tyrimas šiam pavyzdžiui nėra labai prasmingas, nes kintamasis *G7_7* įgyja daugiau nei dvi reikšmes.

11.2.6. Intervalinis kintamasis

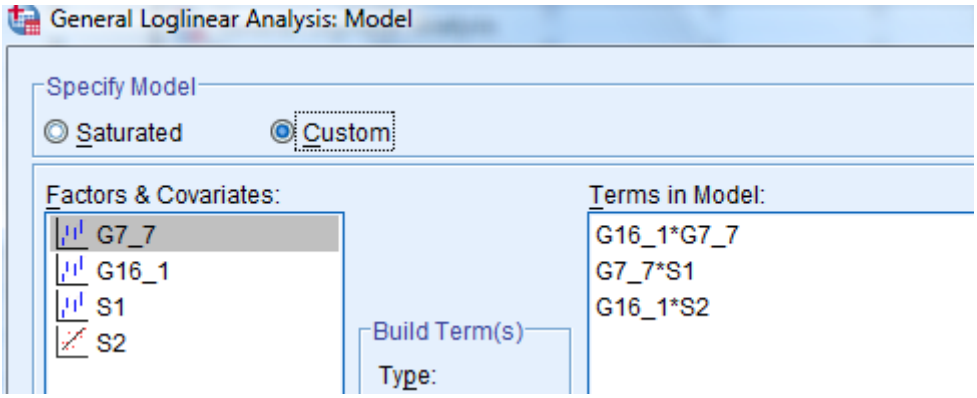
Į modelį galima įtraukti ir intervalinį kintamąjį. Įtraukime į ankstesnį modelį respondento amžių *S2*. Tirsime modelį, kurį sąlyginai galime užrašyti taip:

$$\ln(\text{skaičius grupėje}) = \mu + S1*G7\_7 + G16\_1*G7\_7 + S2*G16\_1.$$

Taigi, tariame, kad požiūris į darbą ir šeimą yra susijęs su respondento amžiumi. Visos parinktys lieka, kaip anksčiau, tik *S2* įkeliame į langelį **Cell Covariate(s)**.



Pasirinkę **Model,** į langelį **Terms in Model** perkeliame regresorių sandaugas (sąveikas):



Rezultatai rodo, kad chi kvadrato statistikos rodo statistiškai priimtiną modelio tikimą duomenims (abi *p* reikšmės didesnės už 0,05):

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	1.396	1	.237
Pearson Chi-Square	1.397	1	.237

a. Model: Poisson

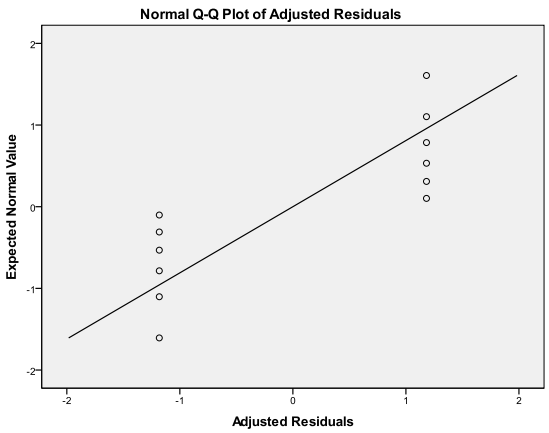
b. Design: Constant + G7_7 * G16_1 + G7_7 * S1 + G16_1 * S2

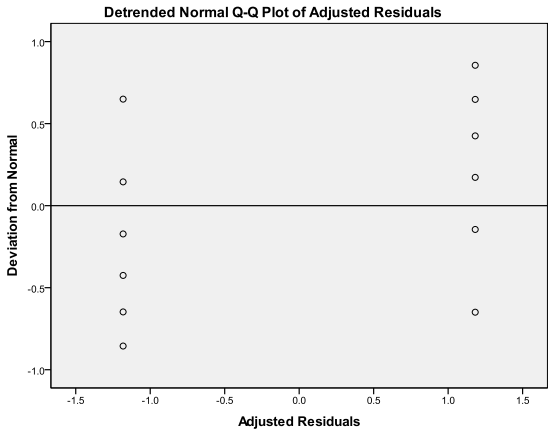
Vis dėlto, negalime sakyti, kad papildytas modelis yra geresnis. Kaip matyti iš lentelės **Parameter Estimates** (pateikiame tik jos fragmetą), kintamasis *S2* yra statistiškai nereikšmingas abiems *G16_1* reikšmėms:

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
[G16_1 = 1] * S2	.105	.191	.551	.582	-.269	.480
[G16_1 = 2] * S2	-.011	.037	-.310	.757	-.084	.061

Be to labai netikę tampa liekamųjų paklaidų grafikai. Darome išvadą, kad naujasis modelis yra netikęs.





11.3. Logtiesiniai modeliai su STATA

11.3.1. Duomenys

Nagrinėsime fiktyvius duomenis. Tarkime, kad apie respondentus žinome ar jie vartoja alkoholį (kintamasis *a*, 1 – taip , 2 – ne), valgo mėsą (kintamasis *m*, 1 – taip, 2 – ne), rūko (kintamasis *r*, 1 – taip, 2 – ne). Turima informacija pateikta lentele:

a	m	r	Kiek tokių
1	1	1	600
1	1	2	300
1	2	1	50
1	2	2	200
2	1	1	20
2	1	2	40
2	2	1	10
2	2	2	150

Sukuriame STATA failą, įvykdydami

```
. input a m r w

             a           m           r           w
  1. 1 1 1 600
  2. 1 1 2 300
  3. 1 2 1 50
  4. 1 2 2 200
  5. 2 1 1 20
  6. 2 1 2 40
  7. 2 2 1 10
  8. 2 2 2 150
  9. end
```

Visada galima ptikrinti, ar duomenys gerai suvesti, įvykdant list a m r w.

11.3.2. Analizė

Logtiesiniams modeliams analizuoti rekomenduojama įdiegti STATA modulį **ipf**. Jį galima atsisiųsti surinkus:

```
findit ipf
```

Patikrinsime, kaip duomenims tinka modelis:

$$\ln(\text{skaičius grupėje}) = \mu + a*m + a*r + m*r.$$

Duomenys jau yra grupuoti. Todėl juos teks pasverti. Įvykdome:

```
ipf [fw=w], fit(a*m+a*r+m*r) exp nolog
```

Svorio (stebėjimų skaičiaus kintamasis) nurodytas laužtiniuose skliaustuose. Jeigu duomenys nėra grupuoti, tai šių skliaustų nėra. Modelis įrašomas į fit(...). Opcija *exp* leis vizualiai palyginti tikrąjį stebėjimų skaičių grupėje su modeliuojamuoju. Gavome:

```
. ipf [fw=w], fit(a*m+a*r+m*r) exp nolog
Deleting all matrices......

Expansion of the various marginal models
------------------------------------------
marginal model 1 varlist :  a m
marginal model 2 varlist :  a r
marginal model 3 varlist :  m r
unique varlist  a m r

N.B.  structural/sampling zeroes may lead to an incorrect df
Residual degrees of freedom = 1
Number of parameters        = 7
Number of cells             = 8


Goodness of Fit Tests
---------------------
df = 1
Likelihood Ratio Statistic G² =   0.0196 p-value = 0.889
Pearson Statistic          X² =   0.0197 p-value = 0.888
```

a	m	r	Efreq	Ofreq	prob
1	1	1	599.69678	600	.43773488
1	1	2	300.30321	300	.21919942
1	2	1	50.303174	50	.03671765
1	2	2	199.69691	200	.14576417
2	1	1	20.303216	20	.01481987
2	1	2	39.696794	40	.02897576
2	2	1	9.6968257	10	.00707797
2	2	2	150.30309	150	.10971028

Matome, kad ir didžiausio tikėtinumo chi kvadrato statistika ir Pirsono chi kvadrato statistika yra statistiškai nereikšmingos (p = 0,889 > 0,05). Tai rodo, kad parinktas modelis gerai aprašo duomenis. Lentelėje pateiktos stebėtieji respondentų skaičiai grupėse (**Ofreg**) ir modeliuojamieji (**Efreg**). Matome, kad modeliuojamosios reikšmės mažai skiriasi nuo stebimųjų. Tai dar vienas argumentas, kad modelis duomenims tinka. Vis dėlto, nėra aišku, ar visi modelio kintamieji yra reikalingi. Todėl įvertinsime jų statistinį reikšmingumą.

Sukuriame kintamuosius, kurie yra *a,m* ir *r* sandaugos:

```
gen am = a*m
gen ar = a*r
gen mr = m*r
```

Įvykdome:

```
glm w am ar mr , fam(pois) link(log)
```

```
                                                  AIC              =  100.9917
Log likelihood   = -399.9667182                   BIC              =  741.5685

------------------------------------------------------------------------------
             |                 OIM
           w |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          am |  -.8693316   .0466876   -18.62   0.000    -.9608376   -.7778255
          ar |  -.2092347   .0365288    -5.73   0.000    -.2808299   -.1376395
          mr |   .2438784   .0301915     8.08   0.000      .184704    .3030527
       _cons |   6.655864   .0783955    84.90   0.000     6.502212    6.809517
------------------------------------------------------------------------------
```

Voldo kriterijus rodo, kad visi kintamieji yra statistiškai reikšmingi (visos p reikšmės yra mažesnės už 0,05). Taigi, juos šalinti iš modelio neturime pagrindo.

Aprašydami gautuosius rezultatus, nurodome, kad visi kintamieji yra poromis priklausomi. Be to, gavome, kad mūsų modelis statistiškai reikšmingai nesiskiria nuo pilnojo modelio, t.y. modelio su trečios eilės kintamųjų sąveika. Darome išvadą, kad trečios eilės sąveika modelyje nereikalinga. Norėdami išaiškinti, kaip susiję kintamieji, turėtume išnagrinėti porinių dažnių lenteles. Čia to nebedarysime, bet aptarsime, kaip suskaičiuoti galimybių santykius.

Surasime galimybių santykį kintamiesiems *ar*. Į mūsų modelį nėra įtraukta trečios eilės kintamųjų priklausomybė, t.y. nėra visų trijų kintamųjų sandaugos. Todėl užtenka nagrinėti galimybių santykį, bet kuriai *m* reikšmei. Galimybių santykiui apskaičiuoti naudojami ne pradiniai duomenus, o sumodeliuotieji. Tą lentelę jau gavome aukščiau. Užrašysime ją patogesniu formatu:

	r	
a	1	2
1	599,7	300,3
2	20,3	39,7

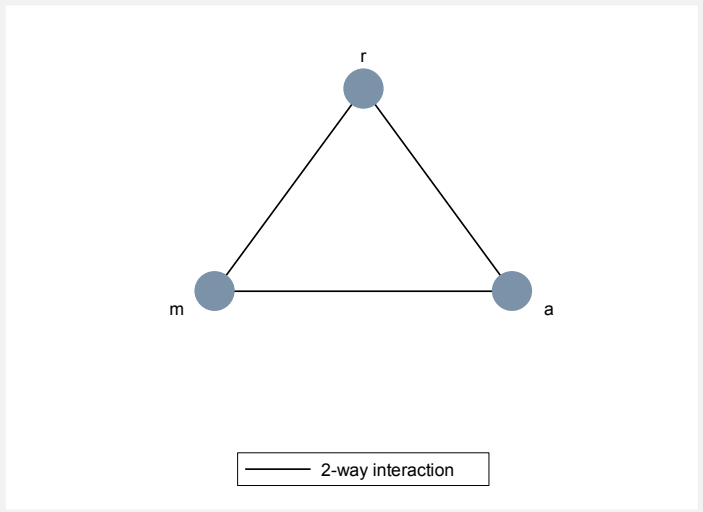
Galimybė, kad alkoholį vartojantis respondentas rūkys yra 599,7/300,3 = 1,997. Galimybė, kad negeriantis respondentas rūkys yra 20,3 / 39,7 = 0,511. Galimybių santykis:

$$\frac{1{,}997}{0{,}511} = 3{,}908.$$

Darome išvadą, kad alkoholio vartojimas padidina rūkymo galimybę 3,9 karto.

Skyrelio pabaigai paminėsime, kad įdiegus papildomą paketą **gipf**, galima nusibraižyti modelio schemą. Įvykdome:

```
gipf, model(a*m+a*r+m*r)
```



Uždaviniai

1. Failas **VS**. Žalingų įpročių paplitimas Lietuvoje, 2004 m. gruodis. Kintamieji

- *Any_drug* – vartojo narkotikus nors kartą gyvenime (1 – vartojo, 2 – nevartojo),
- *B1* – rūko tabaką (1 – taip, 2 – ne),
- *C1* – vartojo alų ar kitus alkoholinius gėrimus per paskutinius 12 mėn. (1 – taip, 2 – ne),
- *Z1* – lytis (1 – vyr., 2 – mot.),
- *Z2* – amžius.

Tirsime žmones iki 35 m. Faile nenurodytos praleistos reikšmės. Todėl reikia atsirinkti tik tuos respondentus, kuriems *Any_drug* ≤ 2 & *B1* ≤ 2 & *C1* ≤ 2 & *Z1* ≤ 2 & *Z2* ≤ 35. Kintamiesiems *Any_drug, B1, C1, Z1* sudaryti tinkamą logtiesinį modelį ir jį ištirti. Kiek kartų padidina galimybę

rūkyti narkotikų vartojimas? Ar modelyje reikalingos trečios eilės priklausomybės? Kurie kintamieji sąlyginai nepriklausomi?

2. Lentelėje pateikti duomenys apie respondentų lytį (kintamasis *L*, 1 – vyr., 2 – mot.), gyvenamą vietą (kintamasis *V* , 1 – Vilnius, 2 – kita), Eurovizijos žiūrėjimą (kintamasis *E*, 1 – žiūrėjo, 2 – nežiūrėjo), ir balsavimą telefonu (kintamasis *B*, 1 – nebalsavo, 2 – balsavo). Sudaryti tinkamą logtiesinį modelį ir aptarti kintamųjų priklausomybes.

L	*V*	*E*	*B*	Kiekis
1	1	1	1	70
1	1	1	2	50
1	1	2	1	100
1	1	2	2	20
1	2	1	1	30
1	2	1	2	19
1	2	2	1	60
1	2	2	2	10
2	1	1	1	110
2	1	1	2	15
2	1	2	1	120
2	1	2	2	5
2	2	1	1	60
2	2	1	2	30
2	2	2	1	150
2	2	2	2	50

Apie statistikos terminus - politkorektiškai.

Duomenų šaltiniai

Duomenys buvo imti iš ES ir LiDA atvirai prieinamų internetinių šaltinių.

- Failas **ZTLT** (Žmogaus teisės Lietuvoje). Tyrimas: Lietuvos tolerancijos profiliai, 2003 m. http://www.lidata.eu/data/quant/LiDA_ZTLT_0012_STUDY_01.
- Failas **ESS4LT1.sav.** Socialinis Lietuvos respondentų tyrimas. http://www.lidata.eu/data/quant/LIDA_ESS_0150_STUDY_01.
- Failas **LAMS** . Universitetų absolventų integracija darbo rinkoje, 2009 m. lapkritis – 2010 m. sausis. http://www.lidata.eu/data/quant/LiDA_LAMS_0146_STUDY_01.
- Failas **ZTLT1.** Vyrai ir moterys, 2006 m. spalis. http://www.lidata.eu/data/quant/LiDA_ZTLT_0160_STUDY_01.
- Failas **VS**. Žalingų įpročių paplitimas Lietuvoje, 2004 m. gruodis. http://www.lidata.eu/data/quant/LiDA_VS_0161_STUDY_01
- Failai **ESS4EE, ESS4PT, ESS4CZ, ESS4SE, ESS4IL, ESS4F**R – tai 2008 metų Europos Sąjungos socialinio tyrimo įvairių šalių duomenys (European Social Survey). http://ess.nsd.uib.no/ess/round4/
- Darbo valandų skaičius ir atlyginimas, 2008 m birželis. http://www.lidata.eu/data/quant/LiDA_SEB_0111_STUDY_01

Literatūra

1. V. Čekanavičius, G. Murauskas. Statistika ir jos taikymai. I., Vilius, TEV, 2000.
2. V. Čekanavičius, G. Murauskas. Statistika ir jos taikymai. II., Vilius, TEV, 2002.
3. A. Agresti. Categorical data analysis. (sec. edition), 2002, Wiley.
4. S.Chatterjee, A.S.Hadi. Regression analysis by Example (fourth edition). 2006, Wiley.
5. A.Field. Discovering statistics using SPSS (and sex, drugs and rock‘n‘roll),(sec. edition). 2005, Sage.
6. J D. Hosmer, S. Lemeshow. Applied logistic regression, (sec. ed.) 2000, Wiley.
7. IBM SPSS Regression 19.
8. D. G. Kleinbaum, M. Klein. Logistic Regression. A Self-Learning Text. 2002, Springer- Verlag.
9. M. Kutner, Ch.J.Nachtsheim, J.Neter, W.Li. Applied Linear Statistical Models (fifth edition), 2005, McGraw- Hill.
10. Scott Long, J. Freese. Regression Models for Categorical Dependent Variables Using Stata, (sec. edition), 2005.
11. B. G. Tabachnick, L.S. Fidell. Using Multivariate Statistics (fifth edition), 2007, Pearson.
12. R. Tarling. Statistical Modelling for Social Researchers. Principles and practice. 2009, Routledge.
13. T.P.Ryan, Modern regression methods. 1997, Wiley.

Naudingi tinklalapiai:

- http://www.ats.ucla.edu/stat/
- http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm