Vydas Čekanavičius
Gediminas Murauskas

STATISTIKA
IR JOS TAIKYMAI

TEV
VILNIUS 2011

UDK 311(075.8)
Ce82

2009-10-01 Nr. 09-06 Aukštųjų mokyklų vadovėlių kokybės vertinimo komisijos rekomenduota

Redaktorė *Zita Manstavičienė*

Programinė įranga: *Tadeuš Šeibak*

Kompiuterinė grafika: *Edita Tatarinavičiūtė*

Maketavo *Aldona Žalienė*

Konsultantas *Elmundas Žalys*

Vadovėlio interneto svetainė http://stat.vadoveliai.lt

© Leidykla TEV, Vilnius, 2009
© Vydas Čekanavičius, 2009
© Gediminas Murauskas, 2009
© Edita Tatarinavičiūtė, 2009

ISBN 978-9955-879-91-6

TURINYS

Pratarmė 5
Žymenys 6

I dalis. TIESINIAI MODELIAI 7

1. Tiesiniai Gauso modeliai 10
   1.1. Struktūrinis LNM užrašas 11
   1.2. Modelių formuluočių pavyzdžiai 12
   1.3. LNM parametrų įverčių skaičiavimas 16
   1.4. Modelio tinkamumas duomenims 20
   1.5. Hipotezių tikrinimas 21
   1.6. ANCOVA 25
   1.7. Tiesinių Gauso modelių analizė naudojantis SPSS (PASW Statistics) programa 33
   1.8. Tiesinių Gauso modelių analizė naudojantis R ir SAS programomis 40
   Uždaviniai 43

2. Apibendrintieji tiesiniai modeliai 45
   2.1. GLM struktūra 45
   2.2. Eksponentinių skirstinių šeima 47
   2.3. GLM formuluočių pavyzdžiai 49
   2.4. GLM parametrų įverčių skaičiavimas 52
   2.5. Hipotezių tikrinimas nagrinėjant GLM 55
   2.6. Modelių tinkamumas 57
   2.7. Apibendrintųjų tiesinių modelių analizė naudojantis SPSS, SAS ir R programomis 60
   Uždaviniai 63

II dalis. HIERARCHINIAI MODELIAI 65

1. Tiesiniai hierarchiniai modeliai 67
   1.1. Hierarchiniai duomenys 68
   1.2. Kodėl taikomas HLM 69
   1.3. Besąlyginis HLM modelis 71
   1.4. Bendrasis HLM modelis 76
   1.5. Dviejų lygmenų modelių pavyzdžiai 80
   1.6. Matricinė modelio formuluotė 83
   1.7. Parametrų įverčių skaičiavimas ir modelio charakteristikos 87
   1.8. Hipotezių tikrinimas 90
   1.9. HLM modelių tyrimas naudojantis SPSS programa 96
   1.10. HLM modelių analizė naudojantis SAS programa 109
   1.11. HLM pavyzdžių analizė naudojantis R programa 113
   1.12. Modelių diagnostika 116
   1.13. Daugialygiai modeliai 118
   Uždaviniai 120

2. HLM modelių taikymas panelinių duomenų analizei 122
   2.1. Paneliniai duomenys 123
   2.2. Augimo modeliai 125
   2.3. Koreliuojančiųjų paklaidų modeliai 126
   2.4. Augimo modelių analizė naudojantis SPSS programa 128
   2.5. Augimo modelių analizė naudojantis R programa 131
   Uždaviniai 134

3. Apibendrintieji hierarchiniai modeliai ... 135
3.1. GLMM struktūra ... 136
3.2. Hierarchinė logistinė regresija ... 138
3.3. GLMM parametrų įverčiai ... 139
3.4. GLMM hipotezių tikrinimas ir tinkamumas duomenims ... 142
3.5. Modelių analizė naudojantis statistinėmis programomis ... 144
Uždaviniai ... 147

III dalis. STATISTINIAI PRIEŽASTINĖS ANALIZĖS MODELIAI ... 149

1. Kelių analizė ... 152
1.1. Kelių analizės modelio aprašymas ... 152
1.2. Modelio diagrama ... 156
1.3. Kelių analizės modelio identifikavimas ... 158
1.4. Modelio parametrų įvertinimas ... 160
1.5. Kelių koeficientų išskaidymas ... 164
1.6. Modelio tobulinimas ... 166
Uždaviniai ... 167

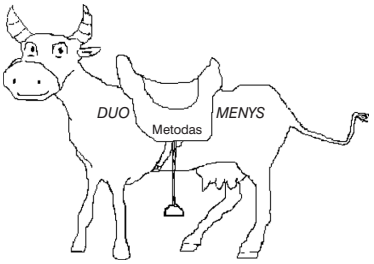
2. Patvirtinančioji faktorinė analizė ... 168
2.1. Dviejų rūšių faktorinės analizės ... 168
2.2. CFA modelio prielaidos ir tyrimo etapai ... 170
2.3. CFA modelio aprašymas ir identifikavimas ... 172
2.4. CFA modelio diagrama ... 175
2.5. Matricinė CFA modelio išraiška ... 176
2.6. CFA modelio parametrų įverčiai ir modelio tinkamumas duomenims ... 178
2.7. CFA modelio parametrų įverčių analizė ... 184
2.8. CFA modelio interpretavimas ir tobulinimas ... 187
Uždaviniai ... 190

3. Bendrieji struktūrinių lygčių modeliai ... 192
3.1. SEM prielaidos ... 193
3.2. SEM modelio aprašymas ... 195
3.3. Modelio identifikavimas ... 199
3.4. SEM modelių pavyzdžiai ... 202
3.5. Matricinis SEM modelio užrašas ... 206
3.6. SEM modelio parametrų vertinimas ir tinkamumas duomenims ... 210
3.7. SEM modelio interpretavimas ir tobulinimas ... 214
3.8. Modeliai su parametrų apribojimais ... 216
3.9. Nenormalieji duomenys ... 220
3.10. SEM analizei skirtos programos ... 224
Uždaviniai ... 230

Priedai ... 232
Vartojamų terminų anglų–lietuvių kalbų žodynėlis ... 237
Dalykinė rodyklė ... 238
Literatūra ... 240

PRATARMĖ

Statistiniuose tyrimuose, ko gero, pati populiariausia yra Prokrusto[1] lovos metodika. Tyrėjas visus (ir save) bando įtikinti, kad jo žinių (kad ir kokios jos menkos būtų) visiškai pakanka turimai informacijai paaiškinti. Metodika paplitusi, be to, gerai atspindėta literatūroje. Pavyzdžiui, ją taikė Pelenės seserys, kai modifikavo savo kojas (kirviu) bandydamos įbrukti jas į krištolinę kurpaitę. Rezultatas žinomas. Panašiai atsitinka, kai naudojami tik keli modeliai ir duomenys į juos „brukami“, nors visiškai netinka. Maždaug taip: nors karvei balnas ir netinka, bet, kai turime tik karvę ir balną... Kai modelis duomenims netinka, statistinės išvados būna paviršutiniškos (dažniausiai) ir statistiškai nepatikimos.



Mūsų tikslas – papildyti skaitytojų statistikos žinių bagažą šiuo metu pasaulyje itin populiariais tiesiniais apibendrintaisiais, hierarchiniais ir struktūrinių lygčių modeliais. ANCOVA, Puasono regresija, kelių analizė ir patvirtinančioji faktorinė analizė – atskiri šių modelių atvejai. Nors pastarąjį dešimtmetį pasaulyje tokie modeliai yra itin populiarūs, Lietuvoje jie taikomi dar nedažnai. Viena iš priežasčių – tinkamos lietuviškos literatūros stoka. Be to, angliškoji specialioji literatūra dažniausiai išsiskiria išsamiais sudėtingo matematinio aparato aprašymais, kurie metodų taikymo nepalengvina.

Ši knyga yra ankstesniųjų[2] tęsinys. Todėl knygos struktūra ir medžiagos pateikimo principai liko tokie patys. Kaip ir ankstesnėse knygose, siekiame skaitytojus išmokyti *taikyti* aprašomus modelius, pateikdami pavyzdžių iš įvairių sričių. Kiekvieno skyriaus pabaigoje yra uždavinių. Duomenų masyvai, reikalingi uždaviniams spręsti, ir patarimai, kaip naudotis SPSS (PASW Statistics), SAS ir R programomis, pateikti interneto puslapyje ***http://stat.vadoveliai.lt***.

Knyga tinkama visiems, kas įstengė perskaityti pirmuosius tomus. Manome, kad ji galėtų tapti įvairių sričių doktorantų, magistrantų ir bakalaurų parankine knyga.

Rašydami šį vadovėlį naudingų pastabų sulaukėme iš V. Kazakevičiaus ir V. Zemlio. Z. Manstavičienė įdėjo labai daug darbo, tvarkydama rankraščio kalbą, kuri tapo gramatiškai taisyklingesnė ir sklandesnė. E. Tatarinavičiūtė vykusiai įkūnijo autorių idėjas piešiniuose. Rengdami šią knygą spaudai jautėme nuolatinį generalinio direktoriaus E. Žalio palaikymą. Visiems jiems nuoširdžiai dėkojame.

Autoriai

---

[1] Legendinis senovės Graikijos guldymo į lovą specialistas, priderindavęs guldomąjį prie lovos įvairiomis *jo* fizinėmis transformacijomis.

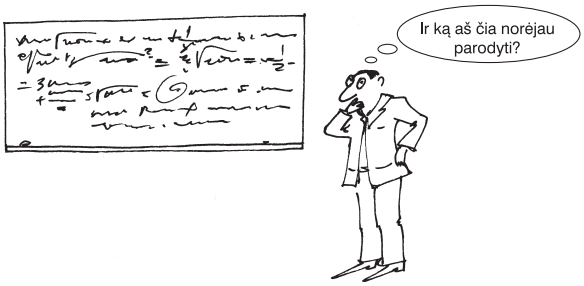
[2] V. Čekanavičius, G. Murauskas, *Statistika ir jos taikymai*. I, II, TEV, Vilnius.

Žymenys

n	imties didumas
$\overline{X}$, $\overline{x}$	imties vidurkis
S^2, s^2	imties dispersija
$\beta, \gamma, \lambda, \ldots$	parametrai (koeficientai)
$\hat{\beta}, \hat{\gamma}, \hat{\lambda}, \ldots$	parametrų (koeficientų) įvertiniai (įverčiai)
$SE(\hat{\beta})$	standartinės $\hat{\beta}$ paklaidos įvertinys
$X, Y, Z, \ldots$	kintamieji
$\mathbf{X}, \mathbf{Y}, \ldots$	kintamųjų reikšmių vektoriai arba matricos
$\mathbf{X}', \mathbf{Y}', \ldots$	transponuotieji vektoriai arba matricos
$\mathbf{E}X$	kintamojo X vidurkis
$\mathbf{D}X$	kintamojo X dispersija
$\mathbb{C}\mathrm{ov}\,\mathbf{Y}$	atsitiktinio vektoriaus $\mathbf{Y}$ kovariacijų matrica
$\mathbf{\Sigma}$	indikatorių kovariacijų matrica
$\mathbf{S}$	empirinė indikatorių kovariacijų matrica
$\mathbf{L}$, $\mathbf{C}$	kontrastų matricos
A^{-1}	atvirkštinė matrica
$f(x)$	tankio funkcija
$X \sim \mathcal{P}(\lambda)$	kintamojo X skirstinys yra Puasono
$X \sim \mathcal{B}i(n, p)$	kintamojo X skirstinys yra binominis
$X \sim \mathcal{N}(\mu, \sigma^2)$	kintamojo X skirstinys yra normalusis
df	laisvės laipsniai
α	reikšmingumo lygmuo
$\chi^2_\alpha(n)$	χ^2 skirstinio su n laisvės laipsnių α lygmens kritinė reikšmė
$t_\alpha(n)$	Stjudento skirstinio su n laisvės laipsnių α lygmens kritinė reikšmė
$F_\alpha(n, m)$	Fišerio skirstinio n ir m laisvės laipsnių α lygmens kritinė reikšmė

1

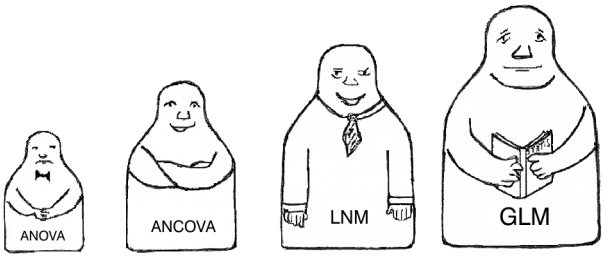
TIESINIAI MODELIAI



Matematika gali būti nusakyta kaip disciplina, kurioje niekada nežinome, nei apie ką kalbame, nei tai, ar tiesa, ką sakome.

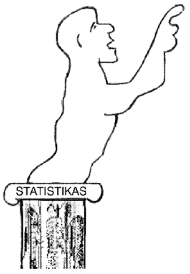
B. Raselas

Yra toks tradicinis rusiškas suvenyras – matrioškos. Didelėje bobulėje (matrioškoje) slepiasi mažesnė, pastarojoje – dar mažesnė, toje – dar mažesnė. Panašiai ir statistikoje – vienas modelis yra kito modelio dalis, šis savo ruožtu yra dar bendresnio modelio dalis, o tas – dar bendresnio ir t. t.



Šioje knygos dalyje aprašysime tiesinius Gauso (normaliuosius, LNM[1]) ir apibendrintuosius tiesinius modelius (GLM[2]). Prie jų priskiriama daug skaitytojui žinomų modelių, pavyzdžiui, tiesinės regresijos ir dispersinės analizės modeliai. Šioje dalyje yra gana daug matematinių užrašų. Todėl kyla natūralus klausimas – kam varginti skaitytoją. Viena vertus, akiračio praplėtimas dar niekam nepakenkė.[3] Kita vertus, didesnioji šios knygos dalis skirta modeliams, išsirutuliojusiems iš tiesinių modelių. Nežinant pagrindinių LNM sąvokų, vargu ar įmanoma suprasti ir sudėtingesnių modelių analizės metodus. Be to, išnagrinėjus tiesinius modelius, bus daug lengviau orientuotis statistikai skirtos literatūros jūroje, ieškant detalesnio vienos ar kitos analizės aprašymo, o statistinių programų procedūrų gausybėje – tinkamos nagrinėjamam modeliui procedūros. Pavyzdžiui, žinant, kad logistinės regresijos modeliai priskiriami prie GLM, analizei galima naudoti SAS programos PROC GENMOD procedūrą ir kt.

Supratus tiesinio modelio struktūrą, galima atlikti išsamesnę statistinę analizę. Įsivaizduokime žmogų, patekusį į staliaus dirbtuvę ir kuriantį paminklą statistikai. Jeigu jis moka tik kirviu mojuoti – pavyks padaryti tik grubiai nutašytą kryžių. Jeigu mokės naudotis kaltu ir obliumi – pavyks tą kryžių statistikos simboliais papuošti. O tikras meistras panaudos daugybę visokių gremžtukų, kaltelių, pjūklelių ir, žiūrėk, išskaptuos skulptūrą. Kuo geriau valdomi įrankiai, tuo rezultatas geresnis.



Panašiai ir statistikoje. Statistinio tyrimo rezultatai labai priklauso nuo naudojamų „įrankių“. Tiesiniams modeliams nagrinėti „įrankių“ yra pakankamai, tik reikia išmokti jais naudotis. Tam skirsime nemažai dėmesio.

---

[1] LNM – angl. *Linear Normal Model.*

[2] GLM – angl. *Generalized Linear Model.*

[3] Įdomu, kad kai kurie Honkongo aktoriai šį posakį suprato tiesiogiai ir ryžosi akių operacijai.

1. TIESINIAI GAUSO MODELIAI

Pabandę aprašyti labai skirtingus gyvenimo reiškinius viena formule, gautume ne itin prasmingą modelį.

Bendrasis gyvenimo modelis

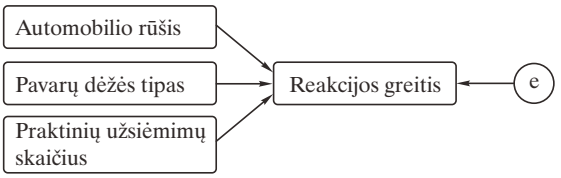
Kaip iš viso statistikoje aprašomas modelis? Paprastai kalbant, stebime kelis mus dominančius kintamuosius ir manome, kad vienas iš jų (tarkime, Y) priklauso nuo likusiųjų, vadinamųjų *aiškinamųjų kintamųjų* (tarkime, $X_1, X_2, \ldots, X_k$). Apytiksliai tokį modelį galima užrašyti taip:

$$Y = f(X_1, X_2, \ldots, X_k);$$

čia $f(\cdot)$ kokia nors funkcija. Pavyzdžiui, tirdami pradedančiųjų vairuotojų reakcijos greitį, galime užrašyti tokį modelį:

Reakcijos greitis $= f($*automobilio rūšis, pavarų dėžės tipas, praktinių užsiėmimų skaičius*$)$.

Žinoma, naivu manyti, kad Y priklauso tik nuo aiškinamųjų kintamųjų X_j. Todėl į tikslų modelio užrašą dar įtraukiama *modelio paklaida* (ją žymėsime simboliu e), t. y. visa, nuo ko dar gali priklausyti Y. Modelis tuo geresnis, kuo ši paklaida mažesnė. Grafiškai tokį modelį galima pavaizduoti taip:



Modelių klasės skiriamos pagal tai, kokius reikalavimus tenkina kintamieji ir kokia funkcija f naudojama. Tiesiniai Gauso modeliai – tai tokie modeliai, kurie užrašomi tiesine priklausomybe, t. y. juose aiškinamieji kintamieji padauginami iš kokių nors koeficientų ir sudedami. Pavyzdžiui, jiems priklauso:

- tiesinės regresijos modeliai;
- dispersinės analizės modeliai;
- kovariančių analizės modeliai.

Šiame skyriuje išsiaiškinsime, kokia yra tiesinių modelių struktūra, kaip vertinami parametrai ir tikrinamos statistinės hipotezės. Primename, kad tiesiniams Gauso modeliams žymėti naudojame santrumpą LNM.

1.1. Struktūrinis LNM užrašas

Matematikai primena prancūzus: kad ir ką jiems sakytum, išverčia į savo kalbą ir tai tampa jau kažkuo visiškai kitu.

J. V. A. Gėtė

Užrašysime LNM tiksliau. Tai patogiausia padaryti naudojantis matricomis, nes užrašas tampa daug trumpesnis ir lengviau suvokiamas. Tarkime, kad turime n normaliojo kintamojo stebėjimų $Y_1, Y_2, \ldots, Y_n$. Tada, naudojantis matricomis, LNM modelį galima užrašyti taip:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{1.1.1}$$

Čia vektorių $\mathbf{Y}$ sudaro priklausomo kintamojo stebėjimai, $\mathbf{e}$ yra paklaidų vektorius, o $\boldsymbol{\beta}$ – nežinomų modelio parametrų vektorius:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_K \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}. \tag{1.1.2}$$

Matrica $\mathbf{X}$ vadinama modelio *struktūros* (plano) matrica[4]. Standartinis $\mathbf{X}$ pavidalas yra toks:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1K} \\ 1 & X_{21} & X_{22} & \ldots & X_{2K} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{nK} \end{pmatrix}; \tag{1.1.3}$$

čia $X_{11}, X_{21}, \ldots, X_{n1}$ yra kintamojo X_1 stebėjimai ir t. t. Galima tarti, kad kintamieji $X_1, X_2, \ldots, X_K$ buvo matuoti n kartų ir gautos reikšmės surašytos į matricą $\mathbf{X}$. Prisiminus, ką reiškia matricų sandauga[5], matyti, kad modelis yra tiesinis – matricos $\mathbf{X}$ elementai dauginami iš koeficientų β ir sudedami. Be to, koeficientas β_0 yra laisvasis modelio narys[6], t. y. jis tiesiog pridedamas prie kitų sandaugų. Norime atkreipti dėmesį, kad yra modelių ir be laisvųjų narių. Tuo atveju pasikeičia ir $\mathbf{X}$ – dingsta pirmasis stulpelis.

Bendrieji LNM struktūros principai:

1. Taikant LNM modelį, žinomos vektoriaus $\mathbf{Y}$ ir į matricą $\mathbf{X}$ įtrauktų aiškinamųjų kintamųjų (regresorių) reikšmės.
2. Vektorius $\mathbf{Y}$ turi daugiamatį normalųjį skirstinį.
3. Tariama, kad į matricą $\mathbf{X}$ įtrauktų aiškinamųjų kintamųjų reikšmės gautos kintamuosius matuojant be paklaidų.
4. Paklaidų vektorius $\mathbf{e}$ turi daugiamatį normalųjį skirstinį su nuliniu vidurkių vektoriumi ir kovariacijų matrica $\boldsymbol{\Sigma}$.

[4] Angl. *Design Matrix.*

[5] Galima ir į knygos priedus žvilgtelėti.

[6] Angl. *Intercept.*

LNM užrašą galima traktuoti kaip vektoriaus $\mathbf{Y}$ išskaidymą į neatsitiktinę ($\mathbf{X}\boldsymbol{\beta}$) ir atsitiktinę ($\mathbf{e}$) dalis. Todėl nesunku suprasti, kad $\mathbf{Y}$ kovariacijų matrica sutampa su $\mathbf{e}$ kovariacijų matrica ir yra lygi $\boldsymbol{\Sigma}$. Pastebėsime, kad apskritai nereikalaujama, jog paklaidos $e_1, e_2, \ldots, e_n$ būtų nepriklausomos. Vadinasi, Y_j stebėjimai gali būti susiję. Tačiau klasikiniuose LNM modeliuose tariama, kad visos modelio paklaidos nekoreliuoja. Be to, jos turi tą pačią dispersiją σ^2, t. y. $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$; čia $\mathbf{I}$ yra vienetinė matrica. Taigi klasikiniuose LNM modeliuose gali būti $K + 2$ nežinomų parametrų: $\beta_0, \beta_1, \ldots, \beta_K$ ir σ^2.

Modelio tyrimo tikslas – išsiaiškinti, ar modelis gerai aprašo turimus duomenis. Kalbant statistikų terminais, reikia įvertinti nežinomus parametrus $\boldsymbol{\beta}$ ir $\boldsymbol{\Sigma}$, patikrinti hipotezes, ar jie statistiškai reikšmingi, ir apskaičiuoti modelio tinkamumo duomenims matą. O po to jau iškyla kiti tikslai, pavyzdžiui, kaip modelį panaudoti prognozavimui.

Kartais nagrinėjami modeliai su atsitiktiniais kintamaisiais $\mathbf{X}$. *Tuo atveju prielaidos formuluojamos sąlyginiams vidurkiams. Pavyzdžiui*: $\mathbf{E}(\mathbf{e}|\mathbf{X}) = 0$, $\mathbf{E}(\mathbf{e}\mathbf{e}'|\mathbf{X}) = \boldsymbol{\Sigma}$. *Modelio analizės metodai nuo to nepasikeičia.*

Toliau knygoje sąvokos *kintamasis* ir *atsitiktinis dydis* laikomos sinonimais. Ankstesnėse [2, 3] knygose didžiosiomis raidėmis žymėjome atsitiktinius dydžius, o mažosiomis – jų realizacijas. Šioje knygoje, vengdami simbolių gausos, paliksime tik didžiąsias raides. Skaitytojams tai neturėtų sukelti didelių problemų.

1.2. Modelių formuluočių pavyzdžiai

Dėstytojas: *„O dabar įrodysime teoremą. Tiesą sakant, aš ją įrodysiu vienas pats.“*

Pateiksime pavyzdžių, kaip naudojantis LNM matricine forma galima užrašyti gerai žinomus modelius. Kituose skyreliuose šiuos pavyzdžius naudosime aprašydami įverčių skaičiavimo ir hipotezių tikrinimo ypatumus. Pradėsime nuo tiesinės regresijos. Primename tipinius regresijos pavyzdžius:

- Tiriama pelno priklausomybė nuo reklamos išlaidų. Modelis (indeksus praleidžiame):
$$Pelnas = \beta_0 + \beta_1 Reklama + e.$$
- Aiškinamasi egzamino rezultatų priklausomybė nuo lankytų paskaitų skaičiaus ir studento IQ. Modelis:
$$Rezultatas = \beta_0 + \beta_1 Paskaitos + \beta_2 IQ + e.$$
- Vertinama namų ūkio išlaidų kultūriniams poreikams priklausomybė nuo bendrųjų pajamų ir vaikų skaičiaus. Modelis:
$$Išlaidos = \beta_0 + \beta_1 Pajamos + \beta_2 Vaikai + e.$$
- Nustatinėjama polinkio į nusikalstamumą priklausomybė nuo socialinio ir ekonominio statuso, draugų įtakos ir respondento amžiaus. Modelis:
$$Polinkis = \beta_0 + \beta_1 Statusas + \beta_2 Draugai + \beta_3 Amžius + e.$$

Visais šiais atvejais tiriame, kaip intervalinis kintamasis Y priklauso nuo intervalinių kintamųjų X_1, X_2, ..., X_K. Turime n kiekvieno kintamojo stebėjimų. Taigi tiesinės regresijos modelis užrašomas taip:

$$Y_j = \beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \cdots + \beta_K X_{jK} + e_j;$$

čia Y_j yra kintamojo Y j-oji reikšmė, X_{j1} yra kintamojo X_1 j-oji reikšmė ir t. t. Matricinis modelio užrašas yra visiškai toks pat, kaip ir bendrojo LNM modelio:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e};$$

čia matrica $\mathbf{X}$ yra (1.1.3) pavidalo, o vektoriai $\mathbf{Y}$, $\boldsymbol{\beta}$ ir $\mathbf{e}$ apibrėžiami (1.1.2) išraiškomis.

Įžangoje minėjome, kad dispersinės analizės modeliai taip pat priskiriami prie tiesinių Gauso modelių. Kaip juos užrašyti LNM matricine forma? Paprastumo dėlei apsiribosime vienfaktore dispersine analize (ANOVA). Iš pradžių priminsime keletą faktų.

ANOVA uždaviniuose populiacijos viena nuo kitos skiriamos pagal vieną požymį (kategorinį kintamąjį, faktorių). Kiekvienoje populiacijoje matuojamas tas pats intervalinis kintamasis. Tikslas – ištirti, ar šio kintamojo populiacijų vidurkiai skiriasi.

Pavyzdžiui:

- ar tiek pat laiko žiūri TV kaimų, miestų ir didmiesčių pensininkai;
- ar skiriasi vidutiniai Lietuvos ir Latvijos programuotojų uždarbiai;
- ar visų rasių žmonės vienodai triukšmingai reiškia emocijas po pralaimėtų rungtynių;
- ar taikant kelias skirtingas terapijas kraujo parametrai normalizuojasi vienodai greitai ir pan.

Turint kintamojo Y i-osios imties j-ąjį stebėjimą, struktūrinis vienfaktorės dispersinės analizės modelis užrašomas taip:

$$Y_{ji} = \mu_i + e_{ji}; \tag{1.1.4}$$

čia e_{ji} yra atsitiktinė paklaida, turinti normalųjį skirstinį, o μ_i – i-osios populiacijos vidurkis. Šį modelį galima užrašyti ir kitaip:

$$Y_{ji} = \mu + (\mu_i - \mu) + e_{ji}; \tag{1.1.5}$$

čia μ yra bendrasis visų populiacijų vidurkis.

Pateiksime tris ANOVA užrašymo LNM matricine forma būdus. Vėliau matysime, kokią įtaką modelio užrašymo forma turi statistinės analizės rezultatams. Bendruoju atveju matricų išraiškos užimtų daug vietos ir būtų sunkiai suprantamos. Perfrazuojant žinomą posakį apie mišką ir medžius, galima pasakyti, kad taikant tokius užrašus bendriesiems dispersinės analizės modeliams per indeksus nebus galima matyti modelio. Todėl šiame skyriuje nagrinėsime konkretų pavyzdį, kai stebėjimų yra nedaug.

Tarkime, kad turime draudimo firmos trijų atsitiktinai parinktų iš *viso* darbo laikotarpio mėnesių išmokas (tūkst. litų)[7].

[7] Tokį paprastą pavyzdį imame tik norėdami paaiškinti teoriją! Turint tiek stebėjimų, kaip nurodyta pavyzdyje, statistinis tyrimas yra beprasmis.

Mėnuo	1	1	2	3	3
Išmoka	5	7	7	4	6

Tariama, kad firma dirba stabilios ekonomikos šalyje. Tyrimo tikslas – įvertinti, ar vidutinės mėnesinės išmokos skiriasi.

Šio uždavinio modelį užrašysime trimis skirtingais būdais.

Pirmasis modelis. Tarkime, kad μ yra bendrasis visų išmokų vidurkis, μ_i – i-ojo mėnesio išmokų vidurkis, e_{ji} – skirtumas tarp mėnesio išmokos ir mėnesio vidurkio. ANOVA tikslas yra nustatyti, ar mėnesių išmokų vidurkiai skiriasi. Naudodamiesi (1.1.5) išraiška, turimą informaciją užrašome tokia lygčių sistema:

$$\begin{aligned}
5 &= \mu + (\mu_1 - \mu) + e_{11},\\
7 &= \mu + (\mu_1 - \mu) + e_{21},\\
7 &= \mu + (\mu_2 - \mu) + e_{12},\\
4 &= \mu + (\mu_3 - \mu) + e_{13},\\
6 &= \mu + (\mu_3 - \mu) + e_{23}.
\end{aligned}$$

Modelio parametras μ laikomas visų vidurkių vidurkiu, t. y.

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3}. \tag{1.1.6}$$

Lygčių sistemą visiškai paprastai galima perrašyti matricomis. Gauname tokį modelį:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}; \tag{1.1.7}$$

čia

$$\mathbf{Y} = \begin{pmatrix} 5\\ 7\\ 7\\ 4\\ 6 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 1 & 0 & 1 & 0\\ 1 & 0 & 0 & 1\\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu\\ \mu_1 - \mu\\ \mu_2 - \mu\\ \mu_3 - \mu \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_{11}\\ e_{21}\\ e_{12}\\ e_{13}\\ e_{23} \end{pmatrix}. \tag{1.1.8}$$

Matricos $\mathbf{X}$ pirmas stulpelis gaunamas sudėjus likusiuosius, t. y. šios matricos rangas yra mažesnis nei stulpelių skaičius. Tokia, atrodytų, nesvarbi detalė sukelia sunkumų skaičiuojant parametrų įverčius. Todėl užrašysime modelį, kuriame ši problema išspręsta.

Antrasis modelis. Pasinaudosime (1.1.6) lygybe. Nesunku patikrinti, kad tada

$$\mu_3 - \mu = -(\mu_1 - \mu) - (\mu_2 - \mu).$$

Todėl paskutines dvi pirmojo modelio lygybes galima užrašyti taip:

$$4 = \mu - (\mu_1 - \mu) - (\mu_2 - \mu) + e_{13}, \qquad 6 = \mu - (\mu_1 - \mu) - (\mu_2 - \mu) + e_{23}.$$

Taigi pirmasis modelis virsta tokiu:

$$\begin{aligned}
5 &= \mu + (\mu_1 - \mu) + e_{11},\\
7 &= \mu + (\mu_1 - \mu) + e_{21},\\
7 &= \mu + (\mu_2 - \mu) + e_{12},\\
4 &= \mu - (\mu_1 - \mu) - (\mu_2 - \mu) + e_{13},\\
6 &= \mu - (\mu_1 - \mu) - (\mu_2 - \mu) + e_{23}.
\end{aligned}$$

Šią lygčių sistemą užrašę matricomis, gauname tokį modelį:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}; \tag{1.1.9}$$

čia

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \mu_1 - \mu \\ \mu_2 - \mu \end{pmatrix}. \tag{1.1.10}$$

Vektoriai $\mathbf{Y}$ ir $\mathbf{e}$ liko tie patys, kaip ir pirmajame modelyje.

Nagrinėjamame pavyzdyje parametrų vektorių $\boldsymbol{\beta}$ sudaro bendrasis visų išmokų vidurkis ir skirtumai tarp pirmojo bei antrojo mėnesių vidutinių išmokų ir bendrojo išmokų vidurkio. Jeigu LNM modelį užrašytume be laisvojo nario, tai parametrų vektoriaus $\boldsymbol{\beta}$ interpretacija pasikeistų.

Trečiasis modelis. Pasinaudoję (1.1.4) išraiška, gauname tokią lygčių sistemą:

$$\begin{aligned}
5 &= \mu_1 + e_{11},\\
7 &= \mu_1 + e_{21},\\
7 &= \mu_2 + e_{12},\\
4 &= \mu_3 + e_{13},\\
6 &= \mu_3 + e_{23}.
\end{aligned}$$

Šį modelį matricomis galima užrašyti taip:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}; \tag{1.1.11}$$

čia

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}. \tag{1.1.12}$$

Matome, kad parametrų vektorių sudaro visų tiriamų populiacijų vidurkiai. Nors (1.1.12) išraiškos yra paprastesnės, tačiau, sprendžiant uždavinius statistinėmis programomis, jos rečiau naudojamos.

Kodėl tiek dėmesio skiriama struktūriniams modelių užrašams? Jų prireikia norint suprasti gautus analizės rezultatus. Pavyzdžiui, statistinėse programose taikomos įvairios LNM užrašo formos. Jų nežinant, sunku suprasti, kokios statistinės hipotezės buvo tikrintos ir kaip reikia interpretuoti gautus rezultatus. Iš pateiktų pavyzdžių matyti, kad, atsižvelgiant į modelio užrašo formą, $\boldsymbol{\beta}$ įgyja skirtingą prasmę.

Tiriant tiesinį Gauso modelį:

1) ***nebūtinas*** jo matricinis užrašas;
2) ***nereikia*** smulkiai žinoti modelio struktūros, norint pasinaudoti statistinėmis programomis;
3) modelio struktūros ***prireikia*** tik išsamiau analizuojant gautus atsakymus.

Žinoma, modelio struktūra svarbi tik atliekant gilesnę analizę. Labai dažnai visiškai pakanka tarp gausybės atsakymų rasti tą vienintelę p reikšmę (dėstytojas dažniausiai tuo atveju sako: „Žiūrėkit čia!“ – ir pirštu baksnoja į ekraną) ir ja remiantis padaryti išvadą.

1.3. LNM parametrų įverčių skaičiavimas

Jei man liktų tik viena diena gyventi, aš ją praleisčiau statistikos paskaitoje. Diena būtų tokia ilga...

Nagrinėsime klasikinį LNM modelį, t. y.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Paklaidų vektorius $\mathbf{e}$ turi normalųjį skirstinį su kovariacijų matrica $\boldsymbol{\Sigma}$, o matricų $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ ir $\mathbf{e}$ matmenys yra atitinkamai $n \times 1$, $n \times (K+1)$, $(K+1) \times 1$ ir $n \times 1$. Mūsų tikslas – rasti matricos $\boldsymbol{\Sigma}$ ir parametrų vektoriaus $\boldsymbol{\beta}$ įverčius. Aptarsime teorinius įverčių gavimo aspektus. Praktiškai įverčiai skaičiuojami naudojantis statistinėmis programomis. Kaip tai atliekama, aprašyta kituose skyreliuose.

Yra keletas įverčių skaičiavimo metodų. Populiariausi – mažiausiųjų kvadratų ir didžiausiojo tikėtinumo (ML[8]) metodai. Šie metodai jau išnagrinėti [2] ir [3] knygose,

[8] ML – angl. *Maximum Likelihood*.

todėl tik trumpai priminsime pagrindines idėjas. Toliau visur tariama, kad stebėjimų yra daugiau, nei vertinamų parametrų ($\boldsymbol{\beta}$ koordinačių), t. y. $n > K + 1$.

Nuo [3] knygos išleidimo praėjo nemažai laiko ir lietuviška statistikos terminija pasipildė naujais žodžiais. Šiuo metu skiriamos dvi sąvokos: *įvertis*[9] ir *įvertinys*[10]. *Įvertinys* – statistika (funkcija), kuri naudojama populiacijos parametrui įvertinti. *Įvertis* – įvertinio realizacija. Trumpai kalbant, įvertis yra tai, *ką* apskaičiuojame, o įvertinys parodo, pagal *kokią formulę* skaičiuojame. Vartojant abi sąvokas, galima lengviau atskirti teorinius samprotavimus nuo konkrečių rezultatų interpretavimo.

Klasikinio Gauso modelio mažiausiųjų kvadratų įvertiniai. Taikant šį metodą, ieškoma tokio $\boldsymbol{\beta}$ įvertinio $\hat{\boldsymbol{\beta}}$, kad apskaičiuotasis skirtumas $\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}$ būtų kuo mažesnis. Kalbant tiksliau, $\hat{\boldsymbol{\beta}}$ minimizuoja funkciją $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$.

Tarkime, kad turime klasikinį LNM modelį su nekoreliuojančiomis paklaidomis, t. y. $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Be to, $\mathbf{X}$ turi didžiausią įmanomą rangą (lygų stulpelių skaičiui $K + 1$). Tada mažiausiųjų kvadratų metodu gaunamas toks $\boldsymbol{\beta}$ įvertinys:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.1.13}$$

Šis įvertinys turi geras savybes. Visų pirma, jis yra nepaslinktasis, o jo kovariacijų matricos struktūra labai paprasta:

1. $\mathbf{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.
2. $\mathbb{C}\mathrm{ov}\,\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Be to, $\hat{\boldsymbol{\beta}}$ turi normalųjį skirstinį ir yra BLUE[11], t. y. visų nepaslinktųjų įvertinių aibėje jo dispersija yra mažiausia. Šis faktas dar vadinamas Gauso–Markovo teorema.

Panaudojus $\hat{\boldsymbol{\beta}}$, galima rasti paklaidų įvertinius:

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{1.1.14}$$

Dispersija σ^2 nėra žinoma, todėl skaičiuojamas jos įvertinys:

$$s^2 = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-K-1} = \frac{\hat{e}_1^2 + \cdots + \hat{e}_n^2}{n-K-1}. \tag{1.1.15}$$

[9] Angl. *Estimate*.

[10] Angl. *Estimator*.

[11] BLUE – angl. *Best Linear Unbiased Estimator.*

Visi įvertiniai nurodyti 1.1.1 lentelėje.

1.1.1 lentelė. LNM parametrų mažiausiųjų kvadratų įvertiniai

Modelis	Įvertiniai
$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$	
$\boldsymbol{\beta}' = (\beta_0, \ldots, \beta_K)$	$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
$\mathbf{e}' = (e_1, \ldots, e_n)$	$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
$\mathbf{e} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$	$s^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/(n - K - 1)$

1.1.1 pavyzdys. Rasime nagrinėto vidutinių draudimo išmokų modelio parametrų įverčius. Naudosimės 1.1.1 lentelės formulėmis. Apskaičiuojame matricą

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & -1 & -1 \\ 0 & 0 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 5 & 0 & -1 \\ 0 & 4 & 2 \\ -1 & 2 & 3 \end{pmatrix}.$$

Ši matrica turi tokią atvirkštinę:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{18} \begin{pmatrix} 4 & -1 & 2 \\ -1 & 7 & -5 \\ 2 & -5 & 10 \end{pmatrix}.$$

Pasinaudoję (1.1.13) formule, gauname

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} \\ \hat{\mu}_1 - \hat{\mu} \\ \hat{\mu}_2 - \hat{\mu} \end{pmatrix} = \frac{1}{18} \begin{pmatrix} 4 & -1 & 2 \\ -1 & 7 & -5 \\ 2 & -5 & 10 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & -1 & -1 \\ 0 & 0 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 5 \\ 7 \\ 7 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ 1 \end{pmatrix}.$$

Taigi $\hat{\mu} = 6$, $\hat{\mu}_1 - \hat{\mu} = 0$, $\hat{\mu}_2 - \hat{\mu} = 1$, t. y. vidutinės mėnesinės išmokos yra tokios: $\hat{\mu}_1 = 6$, $\hat{\mu}_2 = 7$ ir $\hat{\mu}_3 = 5$. Naudodamiesi (1.1.14) ir (1.1.15) formulėmis, apskaičiuojame paklaidų ir dispersijos įverčius:

$$\hat{\mathbf{e}} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \qquad s^2 = \frac{1 + 0 + 1 + 1 + 1}{5 - 3} = 2. \tag{1.1.16}$$

Pasižiūrėkime, kaip atrodys parametrų įverčiai, pasirinkus kitą LNM formą.

1.1.2 pavyzdys. Rasime trečiojo draudimo išmokų modelio parametrų įverčius. Be vargo apskaičiuojame:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \\ 5 \end{pmatrix}. \tag{1.1.17}$$

Kaip ir 1.1.1 pavyzdyje, $\hat{\mu}_1 = 6$, $\hat{\mu}_2 = 7$, $\hat{\mu}_3 = 5$. Nepasikeitė ir $\hat{\mathbf{e}}$ bei s^2 įverčiai.

Taigi, nors vektoriai $\boldsymbol{\beta}$ abiejuose pavyzdžiuose skirtingi, vidutinių mėnesinių išmokų įverčiai yra tokie pat. Tai nėra atsitiktinumas.

Nagrinėjant dispersinės analizės modelius, grupių vidurkių įverčiai nepriklauso nuo pasirinkto modelio formos.
Tačiau $\boldsymbol{\beta}$ įverčiai skiriasi ir nevienodai interpretuojami.
Todėl analizuojamojo modelio struktūrą reikia žinoti ir naudojant statistines programas.

Kai LNM yra tik intervaliniai kintamieji (t. y. modelis tapatus tiesinei regresijai), $\hat{\boldsymbol{\beta}}$ yra tiesiog daugiklių prie atitinkamų aiškinamųjų kintamųjų įverčiai. Primename, kad regresijoje dar skaičiuojami ir standartizuotieji daugikliai, į kuriuos atsižvelgiama nustatant aiškinamųjų kintamųjų įtakos priklausomam kintamajam stiprumą.

Bendresnių Gauso modelių mažiausiųjų kvadratų įvertiniai. Jeigu pabandytume pritaikyti LNM mažiausiųjų kvadratų formules pirmajam draudimo išmokų modeliui, susidurtume su rimta problema. Paaiškėtų, kad matrica $\mathbf{X}'\mathbf{X}$ neturi atvirkštinės. Taip yra todėl, kad matricos $\mathbf{X}$ rangas lygus 3, o stulpelių joje – 4. Pasirodo, kad tuo atveju (1.1.13) formulė vis tiek tinka, tik ieškoma *apibendrintosios atvirkštinės matricos*. Taip vadinama bet kuri matrica $(\mathbf{X}'\mathbf{X})^-$, tenkinanti lygybę

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}.$$

Taigi bendruoju atveju

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}.$$

Jei atvirkštinė matrica egzistuoja, tai ji sutampa su apibendrintąja atvirkštine matrica.

Pavyzdžiui, galima įsitikinti, kad pirmojo modelio matrica $\mathbf{X}'\mathbf{X}$ turi tokią apibendrintąją atvirkštinę matricą:

$$(\mathbf{X}'\mathbf{X})^- = \frac{1}{8}\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 3 & -2 & -1 \\ 1 & -2 & 5 & -2 \\ 0 & -1 & -2 & 3 \end{pmatrix}. \tag{1.1.18}$$

Paklaidų įverčiams skaičiuoti tinka (1.1.14) formulė, o (1.1.15) išraišką reikia pakeisti tokia:

$$s^2 = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-r}; \tag{1.1.17}$$

čia r yra matricos $\mathbf{X}'\mathbf{X}$ rangas.

Yra dar viena situacija, kai parametrų įverčiai skaičiuojami beveik taip pat, kaip ir klasikiniu atveju. Tarkime, kad modelio paklaidos koreliuoja. Tiksliau, kad $\mathbf{e}$ turi kovariacijų matricą $\boldsymbol{\Sigma} = \sigma^2\mathbf{M}$; čia $\mathbf{M}$ yra žinoma teigiamai apibrėžta matrica, o σ^2 – nežinomas parametras. Tada naudojamas *apibendrintasis mažiausiųjų kvadratų metodas* ir parametrų įvertiniai yra tokie:

$$\hat{\boldsymbol{\beta}}_M = (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{M}^{-1}\mathbf{Y}, \qquad \hat{\sigma}_M^2 = \frac{1}{n-K-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{M}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}).$$

Jeigu $\mathbf{M} = \mathbf{I}$, tai turime klasikinius mažiausiųjų kvadratų įvertinius.

Šioje knygoje dar ne kartą susidursime su mažiausiųjų kvadratų metodu. Todėl pakartosime svarbiausią formulę pačiu bendriausiu atveju.

Jeigu $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, o paklaida $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, tai apibendrintuoju mažiausiųjų kvadratų metodu gautas įvertinys yra toks:

$$\hat{\boldsymbol{\beta}}_{\Sigma} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}.$$

Įvertinio išraiškoje yra kovariacijų matrica $\boldsymbol{\Sigma}$. Jeigu ji nežinoma, tai skaičiuojant $\hat{\boldsymbol{\beta}}$ galima naudoti jos įvertį. Tiesa, be papildomos informacijos kovariacijų matricos įvertį ne visada pavyksta apskaičiuoti. Taip yra todėl, kad turint n kintamojo Y reikšmių negalima įvertinti $n(n+1)/2$ kovariacijų.

Didžiausiojo tikėtinumo įvertiniai. Primename didžiausiojo tikėtinumo metodo pagrindinę idėją. Žinome, kaip nuo parametrų priklauso stebimo kintamojo skirstinys. Turint konkrečią imtį, parametrai parenkami taip, kad tokios imties realizacijos tikimybė būtų didžiausia. Esant tolydžiam kintamajam, maksimizuojama ne tikimybė, bet tankių sandaugos logaritmas (logtikėtinumo funkcija, žr. [2]). Tarkime, kad nagrinėjame klasikinį Gauso modelį, t. y. $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Tada didžiausiojo tikėtinumo funkciją sudaro normaliųjų tankių sandauga, o jos logaritmas lygus

$$\ln L = \ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Maksimumas pasiekiamas, kai

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \qquad \hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}.$$

Matome, kad klasikinių Gauso modelių įvertinys $\hat{\boldsymbol{\beta}}$ yra toks pat, kaip ir gautas skaičiuojant mažiausiųjų kvadratų metodu. Tuo tarpu dispersijos σ^2 įvertinys skiriasi nuo s^2, pateikto (1.1.15) formule.

1.4. Modelio tinkamumas duomenims

Jei trikampiai turėtų dievą, jis būtų trišonis.

Žydų patarlė

Sudarius tiesinį modelį, t. y. apibrėžus, nuo kokių kintamųjų X_j gali priklausyti kintamasis Y, reikia ištirti, ar jis tinka duomenims. Dar kartą priminsime, kad nagrinėjame tokį klasikinį Gauso modelį:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e};$$

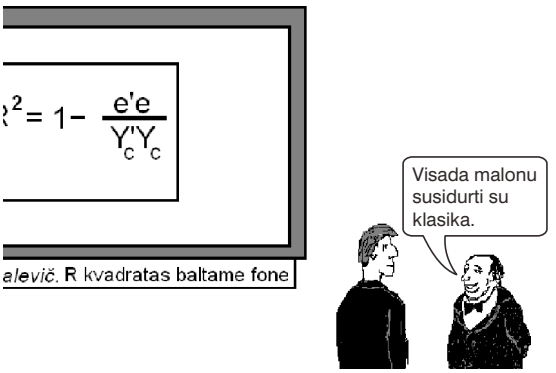
čia $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$. Užrašysime bendrąjį modelio tinkamumo duomenims matą. Nagrinėjant Gauso modelius, jis apibrėžiamas taip:

$$R^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}_c'\mathbf{Y}_c}; \tag{1.1.20}$$

čia $\mathbf{e}$ – paklaidų įvertinys, $\mathbf{Y}_c$ yra vektorius, gautas iš kiekvienos $\mathbf{Y}$ koordinatės atėmus visų Y stebėjimų vidurkį, t. y.

$$\mathbf{Y}_c = \begin{pmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \dots \\ Y_n - \overline{Y} \end{pmatrix}, \quad \overline{Y} = \frac{Y_1 + Y_2 + \dots + Y_n}{n}.$$

Tiesinės regresijos atveju R^2 yra determinacijos koeficientas, o ANOVA atveju – sutampa su dispersinėje analizėje naudojamu koeficientu η^2. Primename, kad η^2 reikšmė gali būti interpretuojama kaip grupės įtakos svarba. Kaip ir regresijoje, kuo R^2 didesnis, tuo geriau LNM modelis aprašo turimus duomenis. Statistikai vengia griežtų nuorodų, kokį R^2 laikyti pakankamai dideliu, nes visada dar reikia atsižvelgti į daugybę kitų rodiklių – imties didumą, kintamųjų normalumą, koreliuotumą ir pan. Vis dėlto galima priminti vieną negriežtą taisyklę, pagal kurią R^2 didumą siūloma vertinti taip: jeigu $R^2 < 0{,}20$, labai abejotina, ar modelis duomenims tinka.



1.1.3 pavyzdys. Apskaičiuosime antrojo vidutinių draudimo išmokų modelio R^2 reikšmę. Bendrasis stebėjimų vidurkis yra $\overline{Y} = (5+7+7+4+6)/5 = 5{,}8$. Todėl $\mathbf{Y}_c' = (-0{,}8;\ 1{,}2;\ 1{,}2;\ -1{,}8;\ -0{,}2)$. Pasinaudoję $\hat{\mathbf{e}}$, apibrėžta (1.1.16) formule, gauname

$$R^2 = 1 - \frac{1+1+1+1}{0{,}8^2 + 1{,}2^2 + 1{,}8^2 + 0{,}2^2} = \frac{2{,}8}{6{,}8} = 0{,}4117\dots.$$

Matome, kad R^2 reikšmė pakankamai didelė. Vadinasi, modelis gerai tinka duomenims. Nagrinėjame dispersinės analizės modelį, todėl R^2 sutampa su ANOVA modeliuose naudojamu rodikliu η^2.

1.5. Hipotezių tikrinimas

Vien tik matematika leidžia mums pajusti savo mąstymo ribotumą. Darant eksperimentą visada galima pasiteisinti, kad jis nesuprantamas, nes turima per mažai duomenų. Matematikoje turime visus duomenis ir vis tiek jos nesuprantame.
S. Veilas

Aprašysime, kaip tikrinamos hipotezės turint klasikinį Gauso modelį:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e};$$

čia $\mathbf{Y} = (Y_1, \dots, Y_n)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)'$.

Nurodysime, kokios statistinės hipotezės tikrinamos ir kokios formulės tam taikomos. Atkreipsime dėmesį, kad neklasikiniuose tiesiniuose modeliuose taikomi kitokie statistiniai kriterijai, tačiau hipotezių formuluotės ir gautųjų išvadų interpretacija lieka tokia pat, kaip ir klasikiniu atveju.

Bendroji hipotezė apie β lygybę nuliui. Hipotezė formuluojama tiems LNM, kurie turi laisvuosius narius β_0. Primename, kad tada pirmąjį struktūros matricos $\mathbf{X}$ stulpelį sudaro vien tik vienetai. Iškeliama hipotezė, kad visi likusieji koeficientai $\beta_1, \beta_2, \ldots, \beta_K$ yra lygūs nuliui. Tai ekvivalentu tyrimui, ar Y priklauso nors nuo vieno X_j. Pažymėkime $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_K)$.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \tilde{\boldsymbol{\beta}} = \mathbf{0}, \\ H_1\colon\ \tilde{\boldsymbol{\beta}} \neq \mathbf{0}. \end{cases}$$

2 *Kriterijaus statistika.* Apskaičiuojame

$$F = \frac{n-K-1}{K} \cdot \frac{R^2}{1-R^2};$$

čia R^2 apibrėžtas (1.1.20) formule, o n yra visų stebėjimų skaičius.

3 *Sprendimo priėmimo taisyklė.* Nulinė hipotezė atmetama, jeigu $F > F_\alpha(K, n-K-1)$; čia $F_\alpha(K, n-K-1)$ yra Fišerio skirstinio su K ir $n-K-1$ laisvės laipsnių α lygmens kritinė reikšmė, α – pasirinktasis reikšmingumo lygmuo.

4 *Sprendimo priėmimo, atsižvelgiant į p reikšmę, taisyklė.* Nulinė hipotezė atmetama, jeigu $p < \alpha$. Tuo atveju daroma išvada, kad yra bent vienas X_j, kurio įtaka Y statistiškai reikšminga.

1.1.4 pavyzdys. Grįžkime prie antrojo draudimo išmokų modelio. Patikrinsime hipotezę, kad $\tilde{\boldsymbol{\beta}} = \mathbf{0}$. Tarkime, reikšmingumo lygmuo $\alpha = 0{,}05$. Pasinaudoję 1.1.3 pavyzdyje gautu R^2 įverčiu, apskaičiuojame

$$F = \frac{5-3}{2} \cdot \frac{2{,}8}{4} = 0{,}70.$$

Iš lentelių (galima pasinaudoti [2, 3] knygomis) randame reikiamą kritinę reikšmę: $F_{0{,}05}(3, 2) = 19{,}16$. Kadangi $0{,}70 < 19{,}16$, tai nulinės hipotezės neatmetame. Darome išvadą, kad koeficientus β_1 ir β_2 galima

laikyti statistiškai reikšmingai nesiskiriančiais nuo nulio. Prisiminkime, kad nagrinėjamame modelyje $\beta_1 = \mu_1 - \mu_3$, o $\beta_2 = \mu_2 - \mu_3$. Taigi iš tikrųjų patikrinome ANOVA hipotezę (ir jos neatmetėme), kad visų mėnesių vidutinės draudimo išmokos vienodos: $\mu_1 = \mu_2 = \mu_3$. Beje, ši išvada prieštarauja 1.1.4 pavyzdžio rezultatams. Todėl dar kartą akcentuojame, kad tai teorinis pavyzdys – duomenų per mažai, kad statistinės išvados būtų prasmingos (kai duomenų mažai, labai retai nulinės hipotezės atmetamos).

Hipotezių apie atskirų parametrų lygybę nuliui tikrinimas. Tikslinant modelį, neretai prireikia patikrinti hipotezę, kad $\beta_j = 0$. Tokią hipotezę galima interpretuoti kaip tyrimą, ar Y priklauso nuo X_j.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \beta_j = 0, \\ H_1\colon\ \beta_j \neq 0. \end{cases}$$

2 *Kriterijaus statistika.* Apskaičiuojame

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)};$$

čia $\hat{\beta}_j$ yra atitinkama $\hat{\boldsymbol{\beta}}$ koordinatė, o $SE(\hat{\beta}_j)$ yra $\hat{\beta}_j$ standartinės paklaidos įvertinys. Jis lygus kvadratinei šakniai iš matricos $s^2(\mathbf{X}'\mathbf{X})^{-1}$ pagrindinės įstrižainės elemento, atitinkančio $\hat{\beta}_j$.

3 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo yra α. Nulinę hipotezę atmetame ir darome išvadą, kad Y statistiškai reikšmingai priklauso nuo X_j, jeigu $|t| > t_\alpha(n-K-1)$; čia $t_\alpha(n-K-1)$ yra Stjudento skirstinio su $n-K-1$ laisvės laipsnių α lygmens kritinė reikšmė.

4 *Sprendimo priėmimo, atsižvelgiant į p reikšmę, taisyklė.* Nulinė hipotezė atmetama, jeigu $p < \alpha$.

1.1.5 pavyzdys. Patikrinsime hipotezę, ar antrojo draudimo išmokų modelio koeficientas $\beta_0 = 0$. Tarkime, reikšmingumo lygmuo $\alpha = 0{,}05$. Pasinaudoję 1.1.4 pavyzdyje gautais rezultatais, apskaičiuojame

$$\left(SE(\hat{\beta}_0)\right)^2 = \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{11} = \frac{4}{9}, \qquad t = \frac{6}{\sqrt{4/9}} = 9.$$

Iš lentelių randame reikiamą kritinę reikšmę: $t_{0{,}025}(2) = 4{,}3$. Kadangi $9 > 4{,}3$, tai nulinę hipotezę atmetame ir darome išvadą, kad koeficientas β_0 statistiškai reikšmingai skiriasi nuo nulio. Ekonomikoje šio fakto interpretacija nėra itin prasminga. Prisiminę, kad nagrinėjamame modelyje $\beta_0 = \mu = (\mu_1 + \mu_2 + \mu_3)/3$, darome išvadą, kad vidutinių išmokų suma $\mu_1 + \mu_2 + \mu_3$ statistiškai reikšmingai skiriasi nuo nulio.

Tiesinės regresijos atveju (t. y. kai yra tik intervaliniai X) šis Stjudento kriterijus taikomas siekiant nustatyti statistiškai nereikšmingus kintamuosius. Jeigu kuris nors $\hat{\beta}_j$ statistiškai reikšmingai nesiskiria nuo nulio, tai reikia pabandyti atitinkamą struktūros matricos $\mathbf{X}$ stulpelį (arba atitinkamą nepriklausomą kintamąjį) iš modelio pašalinti.

Hipotezės apie parametrų tiesinių darinių lygybę skaičiui. Bendresniu atveju hipotezes apie parametrų reikšmes galima suformuluoti apibrėžus apribojimų (kontrastų) matricą $\mathbf{L}$.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{c}, \\ H_1\colon\ \mathbf{L}\boldsymbol{\beta} \neq \mathbf{c}; \end{cases}$$

čia $\mathbf{c}$ yra iš anksto pasirinktas konstantų vektorius. Jis pasirenkamas atsižvelgiant į norimą tikrinti hipotezę. Dažniausiai $\mathbf{c} = \mathbf{0}$.

2 *Kriterijaus statistika.* Apskaičiuojame

$$F = \frac{(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})}{qs^2};$$

čia q – matricos $\mathbf{L}$ rangas (tikrinamosiose hipotezėse jis sutampa su matricos eilučių skaičiumi).

3 *Sprendimo priėmimo taisyklė.* Nulinė hipotezė atmetama, jeigu $F > F_\alpha(q, n-K-1)$; čia $F_\alpha(q, n-K-1)$ yra Fišerio skirstinio su q ir $n-K-1$ laisvės laipsnių α lygmens kritinė reikšmė, α – pasirinktas reikšmingumo lygmuo.

4 *Sprendimo priėmimo, atsižvelgiant į p reikšmę, taisyklė.* Nulinė hipotezė atmetama, jeigu $p < \alpha$.

Pastaba. Jeigu $\mathbf{X}$ rangas yra $r < K+1$, tai vietoj $(\mathbf{X}'\mathbf{X})^{-1}$ naudojama apibendrintoji atvirkštinė matrica $(\mathbf{X}'\mathbf{X})^{-}$, o F lyginama su $F_\alpha(q, n-r)$.

Tiesinės regresijos atveju kontrastai leidžia patikrinti bet kokią spėjamą koeficientų β_j tarpusavio priklausomybę.

1.1.6 pavyzdys. Tiriame Lietuvos 1995–2005 m. ekonominių rodiklių ketvirtinius duomenis. Sudaryta darbo užmokesčio logaritmo (*LDU*) regresijos lygtis. Aiškinamaisiais kintamaisiais pasirinkti sąlyginio darbuotojų skaičiaus logaritmas (*LDSK*) ir prekių bei paslaugų eksporto defliatoriaus logaritmas (*LDEF*). Duomenų aibę sudaro 44 kiekvieno kintamojo stebėjimai. Skaičiavimus atlikus SPSS programa, gauti tokie įverčiai: $\hat{\beta}_0 = 21{,}07$, $\hat{\beta}_1 = -2{,}05$, $\hat{\beta}_2 = 1{,}73$. Taigi regresijos lygtis (LNM dalinis atvejis) yra tokia:

$$LDU = 21{,}07 - 2{,}05LDSK + 1{,}73LDEF.$$

Tarkime, norime patikrinti hipotezę, kad koeficientas prie *LDSK* yra dukart didesnis už koeficientą prie *LDU*. Tiksliau, tikrinsime hipotezę, ar

$$\beta_1 + 2\beta_2 = 0,$$

esant reikšmingumo lygmeniui $\alpha = 0{,}05$. Šią hipotezę galima patikrinti pasirinkus $L = (0, 1, 2)$. Vėliau paaiškinsime, kaip tai galima atlikti naudojantis statistine programa. Dabar tik galime pasakyti, kad SPSS programa patikrinta hipotezė nebuvo atmesta ($p = 0{,}17$).

Kategoriniams aiškinamiesiems kintamiesiems kontrastai sudaromi atsižvelgiant į $\boldsymbol{\beta}$ interpretaciją.

1.1.7 pavyzdys. Patikrinsime hipotezę, kad antrojo draudimo išmokų modelio vidutinė trečiojo mėnesio išmoka lygi pusei sumos, vidutiniškai išmokamos per kitus du mėnesius. Tiksliau, šią hipotezę galima užrašyti taip: $\mu_1 + \mu_2 = 2\mu_3$. Statistinę hipotezę reikia formuluoti vektoriui $\boldsymbol{\beta}$. Prisiminę, kad $\mu_3 - \mu = -(\mu_1 - \mu +$

$\mu_2 - \mu$), matome, kad ši hipotezė ekvivalenti teiginiui $\mu_1 - \mu + \mu_2 - \mu = 0$, t. y. $\mathbf{L}\boldsymbol{\beta} = 0$; čia $\mathbf{L} = (0, 1, 1)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $q = 1$ ir $\mathbf{c} = \mathbf{0}$. Pasinaudoję 1.1.4 pavyzdyje gautais rezultatais, apskaičiuojame:

$$\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}' = \frac{7}{18}, \qquad \mathbf{L}\hat{\boldsymbol{\beta}} = 1, \qquad F = \frac{18/7}{1 \cdot 2} = \frac{9}{7} = 1{,}285\ldots.$$

Iš lentelių randame reikiamą kritinę reikšmę: $F_{0{,}05}(1, 2) = 18{,}5$. Kadangi $1{,}285 < 18{,}5$, tai nulinės hipotezės neatmetame ir darome išvadą, kad duomenys statistiškai reikšmingai neprieštarauja spėjimui, jog $\mu_1 + \mu_2 = 2\mu_3$.

1.1.8 pavyzdys. Prisiminkime trečiąjį draudimo išmokų modelį iri patikrinkime hipotezę, kad $\mu_1 + \mu_2 = 2\mu_3$. Šiuo atveju $\mathbf{L} = (1, 1, -2)$, o $\mathbf{c} = \mathbf{0}$. Naudodamiesi 1.1.1 pavyzdyje gautais rezultatais, apskaičiuojame:

$$\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}' = \frac{1}{2}\begin{pmatrix} 1 & 1 & -2 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} = \frac{7}{2}, \qquad F = \frac{9}{7} = 1{,}285\ldots.$$

Taigi statistikos reikšmė nepasikeitė. Ją reikia lyginti su ta pačia kritine reikšme $F_{0{,}05}(1, 2) = 18{,}5$. Todėl darome tokią pat, kaip ir ankstesniame pavyzdyje, išvadą.

Matome, kad modelio užrašo forma įtakos rezultatams neturi. Tai teisinga ir bendru atveju.

Tikrinamos hipotezės rezultatas nepriklauso nuo pasirinkto modelio užrašo, reikia tik tinkamai pasirinkti matricą $\mathbf{L}$.

Tą pačią hipotezę galima aprašyti skirtingomis kontrastų matricomis. Pavyzdžiui, nagrinėjant trečią modelį, hipotezę $\mu_1 = \mu_2 = \mu_3$ galima patikrinti pasirinkus bet kurią iš kontrastų matricų:

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \quad \text{arba} \quad \mathbf{L} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

Galimybė patikrinti spėjamus parametrų ryšius – vienas pačių svarbiausių gilesnio LNM tyrimo aspektų. Deja, ne bet kokia kontrastų matrica $\mathbf{L}$ tinka hipotezėms tikrinti (netinkamų kontrastų pavyzdžiai pateikiami 1.6 skyrelyje). Vis dėlto norime nuraminti skaitytoją, kad tikrinant sudėtingas hipotezes toliau aprašomomis statistinės analizės programomis nebūtina pačiam „lipdyti“ kontrastų matricą. Ją galima tiesiog pasirinkti iš leistinų kontrastų sąrašo. Šį sąrašą galima išsispausdinti naudojantis statistine programa.

1.6. ANCOVA

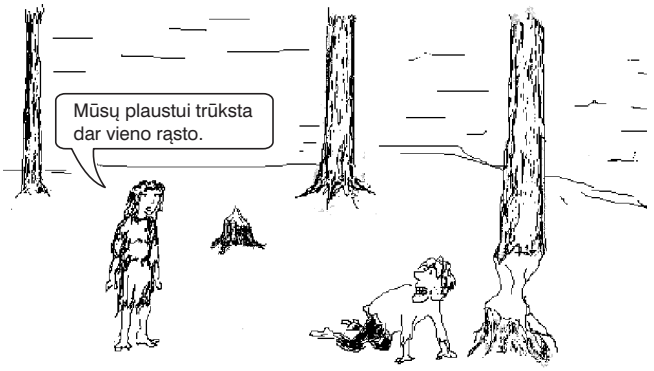
– Tu matematiką myli labiau nei mane!
– Ką tu! Tave aš myliu labiau už viską pasaulyje.
– Įrodyk!
– Gerai. Tarkime, aibę A sudaro visi mylimi objektai...

Vienas iš plačiausiai taikomų LNM dalinių atvejų yra kovariančių analizės (ANCOVA[12]) modelis. ANCOVA galima traktuoti kaip regresijas modelį, kuriame yra pseudokintamųjų. Vis dėlto labiau įprasta ANCOVA laikyti modeliu, skirtu grupių vidurkiams lyginti,

[12] ANCOVA – angl. *Analysis of Covariance.*

kai atsižvelgiama į papildomus, galinčius turėti įtakos intervalinius kintamuosius. Jie vadinami *kovariantėmis*. Taigi ANCOVA laikoma tam tikru patikslintu dispersinės analizės (ANOVA) variantu.

Tarkime, kad pasirinkus keletą maisto rūšių, stebėta, kiek laiko aktyvia fizine veikla užsiima kiekvienas, atitinkamu maistu pavaišintas, respondentas. Pritaikius ANOVA, gauta statistinė išvada, kad labiausiai fizinę veiklą skatina kepti sliekai. Būtent taip negyvenamoje saloje besimaitindami respondentai plušėjo nuo ankstyvo ryto iki gilių sutemų. Vis dirbo, ir dirbo, ir dirbo, ir dirbo plaustą...



Aišku, kad šiuo atveju padaryta išvada nelabai galima tikėti, nes nebuvo atsižvelgta į svarbų faktorių – respondentų motyvaciją. Pateikiame dar kelis pavyzdžius, kuriuose svarbu atsižvelgti į kovariantes.

- Tiriama, ar šeimų didumai besivystančių valstybių kaimuose ir miestuose statistiškai reikšmingai skiriasi (standartinis ANOVA klausimas), kai atsižvelgiama į respondento socialinį ir ekonominį statusą (standartinis ANCOVA patikslinimas).
- Norima sužinoti, ar vienodai geri du gydymo metodai (matuota trukmė iki visiško pasveikimo). Papildomai atsižvelgiama į respondentų amžių ir jautrumą įtaigai.
- Norima nustatyti, kuris iš dviejų mokymo metodų geresnis, kai, vertinant rezultatus, atsižvelgama į mokinių motyvaciją ir IQ.
- Tiriama, ar besimokantys vairuoti, važiuodami dideliais automobiliais, padaro daugiau klaidų nei važiuodami mažais. Papildomai atsižvelgiama į vairuotojo amžių.
- Mobilius telefonus gaminanti bendrovė nori išsiaiškinti, ar naujasis modelis klientui teikia didesnį pasitenkinimą. Apklausa parodė, kad klientai, turintys naująjį modelį, kritiškiau jį vertina, nei senojo modelio turėtojai. Be to, naująjį modelį įsigijo daug jaunesni respondentai. Lieka neaišku, ar modelis yra blogas, ar tiesiog jaunimas labiau linkęs viską kritikuoti. Šiuo atveju reiktų lyginti nuomones apie modelius (tai turėtų būti intervalinis kintamasis – pavyzdžiui, pažymys), papildomai atsižvelgus į respondento amžių.

Vienfaktorė ANCOVA. Ji labai primena vienfaktorę dispersinę analizę. Populiacijos viena nuo kitos skiriamos pagal vieną požymį (kategorinį kintamąjį, faktorių). Kiekvienoje populiacijoje matuojamas tas pats intervalinis kintamasis, o mus domina, ar šio kintamojo

populiacijų vidurkiai (toliau – grupių vidurkiai) skiriasi. Be to, atsižvelgiama į papildomų intervalinių kintamųjų reikšmes. Kad būtų paprasčiau, toliau atsižvelgsime tik į vieną papildomą intervalinį kintamąjį (kovariantę).

Tiriama hipotezė, kad kintamojo Y populiacijų vidurkiai skiriasi.
Atsižvelgiama į X reikšmes.

Taigi ANCOVA modelis užrašomas taip[13]:

$$Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \overline{X}) + e_{ij}; \tag{1.1.21}$$

čia Y_{ij} yra priklausomo kintamojo j-osios imties i-asis stebėjimas, X_{ij} – kovariantės j-osios imties i-asis stebėjimas, μ – bendrasis visų populiacijų kintamojo Y vidurkis, $\tau_j = \mu_j - \mu$, $\mu_j = \mathbf{E}Y_{ij}$, $\overline{X}$ – bendrasis visų imčių kintamojo X vidurkis, $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ – atsitiktinė paklaida, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. Modelio interpretacija tokia: rezultatai priklauso nuo grupės (τ_j), nuo kovariantės X_{ij} ir nuo individualios paklaidos e_{ij}. Taigi ANCOVA modelį užrašome tokia lygtimi:

$$\begin{aligned} \textit{Rezultatas} = {} & \textit{bendrasis vidurkis} + \textit{grupės įtaka} \\ & + \textit{kovariantės įtaka} + \textit{individualumas}. \end{aligned}$$

Tikslas – nustatyti, ar yra grupės įtaka rezultatams.

Perkėlę kovariantę į kairiąją lygties pusę, gauname

$$Y_{ij} - \beta(X_{ij} - \overline{X}) = \mu + \tau_j + e_{ij}.$$

Dešiniojoje lygybės pusėje yra ANOVA modelis, kairiojoje – kintamojo Y reikšmė, „pataisyta“ atsižvelgus į kovariantę. Todėl galime sakyti, kad lyginsime „pataisytas“ imtis.

Jei $\tau_j = 0$, tai (1.1.21) modelis virsta paprasčiausiu tiesinės regresijos modeliu, o kai $\beta = 0$ – ANOVA modeliu. Taigi ANCOVA yra regresijos ir dispersinės analizės modelių mišinys. Todėl ir reikalavimai kintamiesiems yra reikalavimų, keliamų nagrinėjant regresijos ir dispersinės analizės modelius, mišinys:

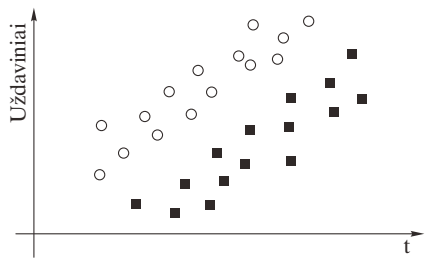
1 Kovariantė yra intervalinė ir matuojama be paklaidos (e_{ij} yra Y matavimo paklaida, o ne X).

2 Kovariantė nėra susijusi su grupavimo faktoriumi (nėra sąveikos tarp grupės ir kovariantės). Pavyzdžiui, negali būti, kad gydant vienu metodu amžius daug svarbesnis nei gydant kitu metodu. Alternatyviai šį reikalavimą galima suformuluoti taip: skirtingose grupėse priklausomybę tarp X ir Y nusakančios regresinės tiesės yra lygiagrečios. Taigi turint (1.1.21) modelį reikalaujama, kad nagrinėjant visas populiacijas koeficientas β yra vienas ir tas pats.

3 Jeigu yra kelios kovariantės, tai jos stipriai nekoreliuoja (nėra multikolinearumo problemos).

[13] Tai tik vienas iš keleto galimų užrašų.

4 Visos liekamosios paklaidos e_{ij} yra nepriklausomos, normaliai pasiskirsčiusios, jų vidurkis lygus nuliui, o dispersija σ^2 yra ta pati, t. y. $e_{ij} \sim \mathcal{N}(0, \sigma^2)$.



1.1.1 pav. Grafinė ANCOVA iliustracija

Pabandysime vienfaktorę ANCOVA paaiškinti grafiškai. Dviejų grupių studentų išspręstų uždavinių skaičiaus priklausomybė nuo sprendimo laiko pavaizduota 1.1.1 paveiksle. Iš jo matyti, kad kuo daugiau laiko studentas sprendžia, tuo daugiau uždavinių išsprendžia. Be to, akivaizdžiai geriau juos sprendžia vienos grupės studentai. Jeigu ignoruotume rezultatų priklausomybę nuo laiko, tai lygintume trumpai sprendusių vienos grupės studentų rezultatus su kitos grupės ilgai sprendusiųjų rezultatais ir didelio skirtumo tarp grupių nepamatytume. Taikydami ANCOVA lyginame vienodą laiką uždavinius sprendusių respondentų rezultatus.

Taigi, neatsižvelgdami į testavimo trukmę, padarytume klaidingą išvadą. Rezultatų priklausomybė nuo laiko abiejose grupėse vienoda – „debesėliai beveik lygiagretūs“. Todėl sąveikos tarp grupės ir laiko nėra. Negalima teigti, kad viena grupė efektyviau išnaudoja laiką uždaviniams spręsti nei kita.

1.1.9 pavyzdys. Išsiaiškinsime, ar trijų grupių testo rezultatai (*REZ*) statistiškai reikšmingai skiriasi. Atsižvelgsime ir į respondentų motyvaciją (*MOT*). Pasirenkame reikšmingumo lygmenį $\alpha = 0,05$. Duomenys yra tokie:

Grupė	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
REZ	4	5	6	8	6	9	10	11	12	13	6	7	8	9	10
MOT	1	4	3	4	5	6	8	8	9	10	2	4	3	4	5

Grupė	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
REZ	12	12	13	14	15	7	8	9	9	13	12	13	14	15	16
MOT	6	7	8	9	9	3	2	3	4	6	6	7	8	9	10

Grupių rezultatų vidurkiai yra atitinkamai 8,4, 10,6 ir 11,6 taško. Palyginę šiuos vidurkius su vienfaktore dispersine analize gautais rezultatais, matome, kad jie statistiškai reikšmingai nesiskiria: $p = 0{,}079 > 0{,}05$ (žr. 1.1.2 pav.).

Jeigu atsižvelgiama į motyvaciją, tai grupių rezultatų vidurkių skirtumai tampa statistiškai reikšmingi, $p < 0{,}05$ (žr. 1.1.3 pav.). Darome išvadą, kad ne visi mokymo metodai vienodai geri.

ANOVA

REZ

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	53,600	2	26,800	2,792	,079
Within Groups	259,200	27	9,600		
Total	312,800	29			

1.1.2 pav. ANOVA rezultatai

Tests of Between-Subjects Effects

Dependent Variable: REZ

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	293,808[a]	3	97,936	134,074	,000
Intercept	74,779	1	74,779	102,372	,000
MOT	240,208	1	240,208	328,843	,000
GR	54,558	2	27,279	37,344	,000
Error	18,992	26	,730		
Total	3434,000	30			
Corrected Total	312,800	29			

a. R Squared = ,939 (Adjusted R Squared = ,932)

1.1.3 pav. ANCOVA rezultatai

Ką iš tikrųjų lygina ANCOVA? Pagrindinę hipotezę galima interpretuoti taip: skirtingose grupėse nustatytos regresinės, *REZ* ir *MOT* siejančios tiesės, nors ir yra lygiagrečios, tačiau nesutampa. Toks pasakymas visiškai teisingas, tačiau nespecialistui skamba absoliučiai nesuprantamai, todėl jo reikėtų vengti. Šį tvirtinimą reikėtų suformuluoti taip: turinčių tą pačią motyvaciją studentų rezultatai statistiškai reikšmingai skiriasi.

Post hoc testai, taikomi vienfaktorei ANCOVA. Nagrinėjome tris grupes. Reikėtų patikslinti, kurios iš jų statistiškai reikšmingai skiriasi. Dispersinėje analizėje tam skirti *post hoc* testai. Vieningos nuomonės, ką daryti ANCOVA atveju, nėra. Siūlomi du variantai: 1) pasirinkti dvi grupes ir jas palyginti naudojantis specialiai sudarytu kontrastu; 2) surasti koreguotuosius grupių vidurkius ir juos palyginti taikant dispersinės analizės *post hoc* testus.

Pirmasis variantas nepatogus tuo, kad norimus kontrastus reikia mokėti užrašyti. Be to, gautosios p reikšmės nebus koreguojamos atsižvelgiant į porinių lyginimų skaičių. Todėl, jeigu tokių lyginimų bus daug, labai išaugs tikimybė bent vieno lyginimo metu padaryti klaidingą sprendimą. Šį metodą rekomenduojame taikyti tik tada, kai grupių yra nedaug (ne daugiau kaip 5).

1.1.10 pavyzdys. Patikrinsime, ar anksčiau nagrinėtame pavyzdyje antroji ir trečioji grupės skiriasi. Norint parašyti tinkamą kontrastą, reikia žinoti statistinėse programose naudojamą ANCOVA modelio užrašą. Jis iš esmės nesiskiria nuo aprašyto nagrinėjant pirmąjį draudimo išmokų modelį, tik $\mathbf{X}$ matricoje prisideda stulpelis *MOT*, o vektorius $\boldsymbol{\beta}$ pasipildo koeficientu β (iš jo bus dauginamas kintamasis *MOT*), t. y. $\boldsymbol{\beta}' = (\mu, \mu_1 - \mu, \mu_2 - \mu, \mu_3 - \mu, \beta)$.

Bendroji kontrasto $\mathbf{L}$ kiekvienos eilutės struktūra yra $(L_0, L_1, L_2, L_3, L_4)$, o principinė tikrinamos hipotezė išraiška yra tokia:

$$L_0\mu + L_1(\mu_1 - \mu) + L_2(\mu_2 - \mu) + L_3(\mu_3 - \mu) + L_4\beta = 0.$$

Test Results

Dependent Variable: REZ

Source	Sum of Squares	df	Mean Square	F	Sig.
Contrast	3,961	1	3,961	5,423	,028
Error	18,992	26	,730		

1.1.4 pav. Dviejų grupių palyginimas

Todėl, norint patikrinti, ar μ_2 ir μ_3 skiriasi, galima naudotis kontrastu $\mathbf{L} = (0, 0, 1, -1, 0)$. Pasirenkame reikšmingumo lygmenį $\alpha = 0{,}05$. SPSS programa atlikę skaičiavimus, gauname, kad abi grupės pagal vidutinį išspręstų uždavinių skaičių statistiškai reikšmingai skiriasi (žr. 1.1.4 pav.).

Norėdami *post hoc* testams naudoti koreguotus grupių vidurkius, turime susitarti, kas tai yra. Tradiciškai, pasinaudojus ANCOVA modelio užrašu, visi imčių vidurkiai pakeičiami į tokius, kokie jie būtų, kovariantėms įgijus vidutines reikšmes. Taigi j-osios grupės kintamojo Y koreguotasis vidurkis $\widetilde{Y}_i$ skaičiuojamas pagal formulę

$$\widetilde{Y}_j = \overline{Y}_j - \hat{\beta}(\overline{X}_j - \overline{X}); \tag{1.1.22}$$

čia $\overline{Y}_j$ ir $\overline{X}_j$ yra j-osios grupės kintamųjų Y ir X reikšmių empiriniai vidurkiai, $\overline{X}$ – bendrasis visų imčių kintamojo X vidurkis, $\hat{\beta}$ – (1.1.21) lygtyje esančio koeficiento β įvertis.

1.1.11 pavyzdys. Pasinaudosime 1.1.9 pavyzdžio duomenimis ir apskaičiuosime koreguotuosius kintamojo *REZ* vidurkius. Žinome, kad $\overline{REZ}_1 = 8{,}4$, $\overline{REZ}_2 = 10{,}6$ ir $\overline{REZ}_3 = 11{,}6$. Be to, nesunkiai apskaičiuojame $\overline{MOT}_1 = 5{,}8$, $\overline{MOT}_2 = 5{,}7$, $\overline{MOT}_3 = 5{,}8$ ir $\overline{MOT} = 5{,}77$. SPSS programa įvertinę ANCOVA modelį, gauname $\hat{\beta} = 1{,}098$. Todėl koreguotieji rezultatų vidurkiai yra tokie (pateikiame 2 ženklus po kablelio):

$$\widetilde{REZ}_1 = 8{,}4 - 1{,}098(5{,}8 - 5{,}77) = 8{,}36, \qquad \widetilde{REZ}_2 = 10{,}67, \qquad \widetilde{REZ}_3 = 11{,}56.$$

Pairwise Comparisons

Dependent Variable: REZ

(I) GR	(J) GR	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a] Lower Bound	Upper Bound
1,00	2,00	-2,310*	,382	,000	-3,288	-1,332
	3,00	-3,200*	,382	,000	-4,178	-2,222
2,00	1,00	2,310*	,382	,000	1,332	3,288
	3,00	-,890	,382	,084	-1,868	,088
3,00	1,00	3,200*	,382	,000	2,222	4,178
	2,00	,890	,382	,084	-,088	1,868

Based on estimated marginal means

*. The mean difference is significant at the ,05 level.

a. Adjustment for multiple comparisons: Bonferroni.

1.1.5 pav. Koreguotųjų 1.1.9 pavyzdžio vidurkių palyginimas

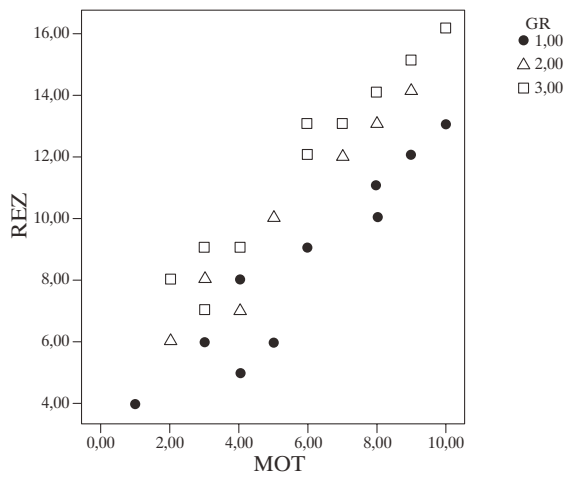
Ar šie vidurkiai statistiškai reikšmingai skiriasi, tikrinsime naudodamiesi Bonferonio kriterijumi. Kaip alternatyva dar naudojamas Sidako kriterijus.

1.1.12 pavyzdys. Naudodamiesi SPSS programa, palyginkime 1.1.9 pavyzdžio koreguotuosius kintamojo *REZ* vidurkius. Gauti rezultatai pateikiami 1.1.5 paveiksle. Tarkime, kad pasirinktasis reikšmingumo lygmuo $\alpha = 0{,}05$. Matome, kad pirmosios grupės koreguotasis vidurkis statistiškai reikšmingai skiriasi nuo likusių dviejų koreguotųjų vidurkių ($p < 0{,}05$). Tuo tarpu antrosios ir trečiosios grupių vidurkiai statistiškai reikšmingai nesiskiria.

Ar *post hoc* testams taikyti kontrastus, ar lyginti koreguotuosius grupių vidurkius, sprendžia pats tyrėjas. Lyginti koreguotuosius vidurkius daug lengviau, nes tai standartinė procedūra, realizuota daugelyje statistinių programų. Ją taikant nereikia galvoti apie kontrastų išraiškas.

Prielaidų tikrinimas. Daugelis žmonių ANCOVA laiko tikra panacėja, nesivargindami patikrinti, ar duomenys jai tinkami. Visų pirma, kintamieji turi būti bent apytiksliai normalūs. Be to, regresijos modeliuose yra prielaida, kad visos paklaidos turi tą pačią dispersiją σ^2 (duomenys homoskedastiški). Jeigu ši prielaida negalioja, sakome, kad duomenys heteroskedastiški. Tuo atveju tiesinio modelio mažiausiųjų kvadratų įvertiniai vis tiek yra nepaslinktieji, suderintieji ir asimptotiškai normalūs, tačiau nėra efektyvūs (visų nepaslinktųjų įvertinių aibėje jų dispersija nėra mažiausia). Kokios galimos heteroskedastiškumo priežastys? Paminėsime tik keletą jų: a) į modelį neįtrauktas svarbus nepriklausomas kintamasis, b) priklausomo ir nepriklausomų kintamųjų sąryšiai nėra tiesiniai.

Dažnai heteroskedastiškumą galima pastebėti, nusibraižius priklausomojo kintamojo ir kovariančių priklausomybės grafikus. Žinoma, grafiką prasminga braižyti tik tada, kai duomenų nėra labai daug. Jeigu pabandysime atidėti tūkstantį taškų, grafikas primins ne debesėlius, o vientisą blyną. Duomenys nėra heteroskedastiški, jeigu gauname „beveik lygiagrečius vienodo storio debesėlius“.



1.1.6 pav. Kintamųjų *REZ* ir *MOT* sklaidos diagrama

Nubraižę 1.1.9 pavyzdžio kintamųjų *REZ* ir *MOT* (žr. 1.1.6 pav.) sklaidos diagramą, matome, kad taškai išsibarstę tolygiai ir duomenys nėra heteroskedastiški.

ANCOVA prielaidos nenumato populiacijos tipo (grupavimo kintamojo) ir kovariančių sąveikos (interakcijos). Todėl visada rekomenduojama sudaryti LNM, į kurį būtų įtraukta grupavimo kintamojo ir kovariantės sandauga, ir patikrinti, ar koeficientas prie šios sandaugos statistiškai reikšmingas. Jeigu taip, tai yra kintamųjų sąveika ir klasikinė ANCOVA netinka. Modelį reikia keisti.

1.1.13 pavyzdys. Patikrinsime, ar nėra grupės ir motyvacijos sąveikos. Šios sąveikos interpretacija yra paprasta: skirtingų grupių motyvacijos įtaka rezultatams yra skirtinga.

Įtraukus kintamųjų sąveiką, modelis užrašomas taip:

$$Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \overline{X}) + \gamma(X_{ij} - \overline{X})\tau_j + e_{ij}.$$

SPSS programa patikrintos hipotezės apie naujojo modelio koeficientų statistinį reikšmingumą pateiktos 1.1.7 paveiksle. Matome, kad grupės ir motyvacijos sąveika nėra statistiškai reikšminga. Taigi gavome patvirtinimą, kad vienas iš svarbiausių ANCOVA reikalavimų tenkinamas.

Tests of Between-Subjects Effects

Dependent Variable: REZ

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	294,814[a]	5	58,963	78,680	,000
Intercept	72,200	1	72,200	96,344	,000
GR	6,102	2	3,051	4,071	,030
MOT	240,283	1	240,283	320,632	,000
GR * MOT	1,006	2	,503	,671	,520
Error	17,986	24	,749		
Total	3434,000	30			
Corrected Total	312,800	29			

a. R Squared = ,943 (Adjusted R Squared = ,931)

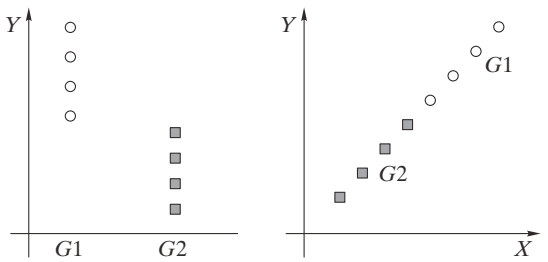
1.1.7 pav. Modelio su parametrų sąveika ANCOVA lentelė

Keletas pastabų. Jeigu kovariantė yra ne viena, tai modelio analizė praktiškai nesikeičia. Tiesiog reikia atsižvelgti į daugiau galimų sąveikų.

Galima nagrinėti ir dvifaktorį ANCOVA modelį. Jį galima įsivaizduoti kaip dvifaktorį ANOVA modelį, papildytą kovariantėmis. Tačiau tokio modelio rezultatus aprašyti pakankamai sudėtinga.

Jeigu yra grupavimo kintamojo ir kovariantės sąveika, tai klasikinis ANCOVA modelis netinka. Tuo atveju galima nagrinėti bendresnį LNM, kuriame yra kintamųjų sąveikos. Problemų gali kilti tik interpretuojant rezultatus.

Nereikia tikėtis, kad ANCOVA visada padės rasti vidurkių skirtumus, kurių nenustatoma su ANOVA. Visiškai įmanomas ir priešingas rezultatas – kol neįtraukiama kovariantė, vidurkiai statistiškai reikšmingai skiriasi. Tuo tarpu atsižvelgus į kovariantę tie skirtumai dingsta (žr. 1.1.8 pav.). Iš kairiojo paveikslėlio aiškiai matyti, kad kintamojo Y reikšmės grupėse skiriasi, iš dešiniojo – kad tie skirtumai atsirado tik dėl skirtingų kovariantės X reikšmių.



1.1.8 pav. Kai vidurkių skirtumai dingsta

1.7. Tiesinių Gauso modelių analizė naudojantis SPSS (PASW Statistics) programa

Tikroji grėsmė – ne tai, kad kompiuteriai pradės mąstyti kaip žmonės, bet tai, kad žmonės pradės mąstyti kaip kompiuteriai.

S. Haris

Praktiškai LNM modeliams tirti naudojamos statistinės programos. Šiame skyrelyje parodysime, kaip tai atlikti naudojantis SPSS programa. Šios knygos pavyzdžiai spręsti naudojantis versija SPSS 17. Nors naujausia 2009 m. išleista programos versija pavadinta PASW Statistics 18, tačiau visos šioje knygoje pateiktos programos jai taip pat tinka. Kituose skyreliuose panagrinėsime R ir SAS programas. Tačiau norėtume atkreipti dėmesį, kad:

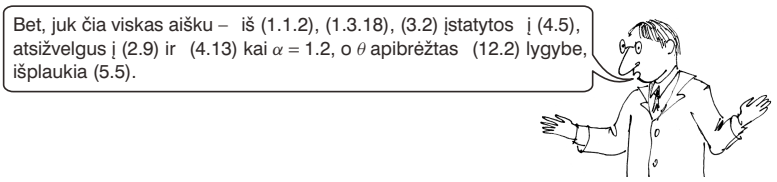
1 Šis vadovėlis nėra programų SPSS, SAS ar R pradžiamokslis. Taigi skaitytojas, norintis pasinaudoti pateikiama medžiaga, turi būti susipažinęs su bent viena iš jų.

2 Nesistengėme atskleisti *visų* aprašomų procedūrų galimybių (tam neužtektų ir kur kas storesnės knygos). Šio skyrelio tikslas – parodyti, kokios yra nagrinėjamų modelių analizės galimybės. Vieniems to pakaks, kitiems galbūt prireiks papildomai žvilgtelėti į statistinės programos dokumentaciją (ar elektroninį pagalbininką).

3 SPSS ir SAS sakiniai gali būti rašomi tiek didžiosiomis, tiek mažosiomis raidėmis. R programos sakiniuose didžiosios ir mažosios raidės skiriamos.

Aptarsime SPSS programos procedūrą UNIANOVA. Svarbiausi UNIANOVA sintaksės paaiškinimai pateikiami 1.1.2 lentelėje.

Nagrinėsime draudimo išmokų modelį, aprašytą 1.1 skyrelyje. Kad būtų patogiau, sąlygą pakartosime.



1.1.2 lentelė. UNIANOVA procedūros sintaksės paaiškinimai

1	UNIANOVA Y BY V Z WITH X W	Išvardijami modelio kintamieji. Po procedūros pavadinimo UNIANOVA nurodomas priklausomas kintamasis – Y, po BY išvardijami kategoriniai kintamieji, vadinamieji faktoriai (V, Z), po WITH išvardijamos kovariantės (X, W)
2	/METHOD = SSTYPE(3)	Įverčių skaičiavimo metodas (METHOD). Parinktis SSTYPE(3) yra numatytoji. Galima nurodyti vieną iš keturių metodų
3	/INTERCEPT = INCLUDE	Nurodoma, ar modelyje yra laisvasis narys. INCLUDE yra numatytoji reikšmė ir reiškia, kad laisvasis narys yra
4	/CRITERIA = ALPHA(.05)	Nurodoma, kad visų pasikliautinųjų intervalų pasikliovimo lygmuo lygus 0,95. Be to, F kriterijaus galia skaičiuojama esant reikšmingumo lygmeniui 0,05
4a	/EMMEANS = TABLE(V * Z) compare(V)	Nurodoma, kokius priklausomo kintamojo vidurkius, esant fiksuotoms regresorių reikšmėms, skaičiuoti
5	DESIGN = V Z V*Z	Nurodoma modelio struktūra. Pavyzdžiui, V*Z reiškia, kad į modelį įtraukta kintamųjų V ir Z sandauga
6	/LMATRIX Intercept a X b c ...	Tikrinama hipotezė apie parametrų tiesinių darinių lygybę nuliui. Aprašomas kontrastas **L**, t. y. išvardijami koeficientai prie laisvojo nario (intercept) ir prie aiškinamųjų kintamųjų X (a, b ir c)

Žinome draudimo firmos trijų atsitiktinai parinktų iš *viso* darbo laikotarpio mėnesių penkias draudimo firmos išmokas (tūkst. litų).

Mėnuo	1	1	2	3	3
Išmoka	5	7	7	4	6

Tarkime, kad μ_i yra i-ojo mėnesio išmokų vidurkis, e_{ji} – išmokos ir mėnesio vidurkio skirtumas. Tada visas modelis užrašomas tokia matricine lygtimi:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e};$$

čia

$$\mathbf{Y} = \begin{pmatrix} 5 \\ 7 \\ 7 \\ 4 \\ 6 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \mu_1 - \mu \\ \mu_2 - \mu \\ \mu_3 - \mu \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_{11} \\ e_{21} \\ e_{12} \\ e_{13} \\ e_{23} \end{pmatrix}. \tag{1.1.23}$$

Modelio parametras μ laikomas visų vidurkių vidurkiu, t. y.

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3}. \tag{1.1.24}$$

Reikia ištirti tokias problemas:

1. Įvertinti vidutines kiekvieno mėnesio išmokas.
2. Patikrinti hipotezę, kad visų mėnesių vidutinių išmokų dydis yra toks pat.
3. *Post hoc* testais nustatyti, kurių mėnesių vidutinės išmokos statistiškai reikšmingai skiriasi.
4. Patikrinti hipotezę, kad išmokų vidurkius sieja lygybė $\mu_1 + \mu_2 = 2\mu_3$.

Duomenyse mėnesio numeris (grupavimo kintamasis) yra kintamasis G. SPSS programa yra tokia:

```
UNIANOVA Y BY G
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/POSTHOC = G (BONFERRONI)
/PLOT = PROFILE(G)
/EMMEANS = TABLES(G)
/PRINT = DESCRIPTIVE TEST(LMATRIX)
/LMATRIX Intercept 0 G 1 1 -2
/CRITERIA = ALPHA(.05)
/DESIGN = G .
```

Trumpai aptarsime programos parametrus. Sakinys UNIANOVA Y BY G nurodo, kad modelyje bus tik kategorinis kintamasis (mėnesio numeris G). Statistikose įverčiams reikalingos kvadratų sumos bus skaičiuojamos trečiuoju metodu (METHOD = SSTYPE(3)). Smulkiai aprašinėti, kuo skiriasi kvadratų sumų skaičiavimai, yra beprasmiška. Tektų įsivelti į daugybę formulių ir matematizuotų komentarų. Todėl tik atkreipiame dėmesį, kad SSTYPE(3) metodas yra tinkamiausias beveik visoms standartinėms situacijoms, pavyzdžiui, kai grupėse yra nevienodas stebėjimų skaičius.

Parametro /INTERCEPT reikšmė INCLUDE nurodo, kad modelyje bus laisvasis narys.

Modeliai, kuriuose yra kategoriniai kintamieji, turi tokias savybes:

- Jeigu į modelį įtraukiamas laisvasis narys, UNIANOVA procedūra lygina grupių vidurkius.
- Jeigu į modelį laisvasis narys neįtraukiamas, tai tikrinama hipotezė, ar visi vidurkiai lygūs nuliui.

Parametro /POSTHOC reikšmė G (BONFERRONI) nurodo, kad vidutinės mėnesių išmokos lyginamos poromis pagal Bonferonio kriterijų. Ši parinktis nenaudojama, jeigu modelyje yra intervalinių kintamųjų. Parametro /PLOT reikšmė PROFILE(G) nurodo, kad reikia nubraižyti grupių vidurkių grafiką.

Įtraukus į modelį parametrą /EMMEANS = TABLES(G), randami kiekvienos grupės (kiekvieno mėnesio) vidurkiai (vidutinės išmokos) ir jų pasikliautinieji intervalai.

Parametras /PRINT turi dvi reikšmes. Pirmoji reikšmė DESCRIPTIVE rašoma norint išspausdinti kiekvienos grupės vidurkį. Modelyje yra tik vienas kategorinis kintamasis,

todėl naudojant DESCRIPTIVE gaunami tie patys įverčiai, kaip ir naudojant /EMMEANS. Antroji reikšmė (TEST(LMATRIX)) nurodo, kad bus pateikti visi kontrastai, kurie buvo naudoti tikrinant, ar koeficientų tiesiniai dariniai lygūs nuliui (t. y. ar yra parametrų sąryšiai).

Parametras /LMATRIX aprašo, kokiam kontrastui bus tikrinama hipotezė. Reikšmės Intercept 0 G 1 1 -2 nurodo, kad bus tikrinama mūsų sugalvota hipotezė $\mu_1+\mu_2 = 2\mu_3$. Iš tikrųjų kontrastų matrica lygi $\mathbf{L} = (0, 1, 1, -2)$. Padauginkime ją iš $\boldsymbol{\beta}$, apibrėžto (1.1.23)

Warnings

Post hoc tests are not performed for G because at least one group has fewer than two cases.

Descriptive Statistics

Dependent Variable: Y

G	Mean	Std. Deviation	N
1,00	6,0000	1,41421	2
2,00	7,0000	.	1
3,00	5,0000	1,41421	2
Total	5,8000	1,30384	5

Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2,800[a]	2	1,400	,700	,588
Intercept	162,000	1	162,000	81,000	,012
G	2,800	2	1,400	,700	,588
Error	4,000	2	2,000		
Total	175,000	5			
Corrected Total	6,800	4			

a. R Squared = ,412 (Adjusted R Squared = -,176)

Contrast Coefficients (L' Matrix)

Intercept

Parameter	Contrast L1
Intercept	1,000
[G=1,00]	,333
[G=2,00]	,333
[G=3,00]	,333

The default display of this matrix is the transpose of the corresponding L matrix. Based on Type III Sums of Squares.

G

Parameter	Contrast L2	Contrast L3
Intercept	0	0
[G=1,00]	1	0
[G=2,00]	0	1
[G=3,00]	-1	-1

The default display of this matrix is the transpose of the corresponding L matrix. Based on Type III Sums of Squares.

1.1.9 pav. UNIANOVA rezultatai (pradžia)

formule:

$$\mathbf{L}\boldsymbol{\beta} = 0 \cdot \mu + 1 \cdot (\mu_1 - \mu) + 1 \cdot (\mu_2 - \mu) - 2 \cdot (\mu_3 - \mu) = \mu_1 + \mu_2 - 2\mu_3.$$

Todėl lygybė $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ ekvivalenti lygybei $\mu_1 + \mu_2 = 2\mu_3$.

Parametro /CRITERIA reikšmė ALPHA(.05) nurodo, kad visų pasikliautinųjų intervalų pasikliovimo lygmuo lygus 0,95, o parinktis DESIGN = G – modelio pavidalą. Modelyje yra tik vienas kategorinis kintamasis.

Gauti rezultatai pavaizduoti 1.1.9, 1.1.10 ir 1.1.11 paveiksluose. Aptarsime juos išsamiau.

Custom Hypothesis Tests

Contrast Coefficients (L' Matrix)

Parameter	Contrast L1
Intercept	,000
[G=1,00]	1,000
[G=2,00]	1,000
[G=3,00]	-2,000

The default display of this matrix is the transpose of the corresponding L matrix.

Contrast Results (K Matrix)[a]

Contrast			Dependent Variable Y
L1	Contrast Estimate		3,000
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		3,000
	Std. Error		2,646
	Sig.		,374
	95% Confidence Interval for Difference	Lower Bound	-8,384
		Upper Bound	14,384

a. Based on the user-specified contrast coefficients (L') matrix number 1

Test Results

Dependent Variable: Y

Source	Sum of Squares	df	Mean Square	F	Sig.
Contrast	2,571	1	2,571	1,286	,374
Error	4,000	2	2,000		

1.1.10 pav. UNIANOVA rezultatai (tęsinys)

Estimated Marginal Means

Contrast Coefficients (L' Matrix)

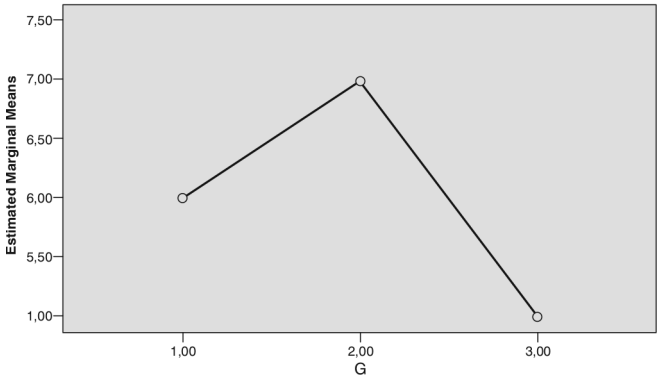
Parameter	G		
	1,00	2,00	3,00
Intercept	1	1	1
[G=1,00]	1	0	0
[G=2,00]	0	1	0
[G=3,00]	0	0	1

Estimates

Dependent Variable: Y

G	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1,00	6,000	1,000	1,697	10,303
2,00	7,000	1,414	,915	13,085
3,00	5,000	1,000	,697	9,303

Estimated Marginal Means of Y



1.1.11 pav. UNIANOVA rezultatai (pabaiga)

Vidutinės kiekvieno mėnesio išmokos pateiktos lentelėse *Descriptive Statistics* (žr. 1.1.9 pav.) ir *Estimates* (žr. 1.1.11 pav.). Lentelėje *Estimates* dar nurodyti ir įverčių pasikliautinieji intervalai bei vidurkių trendas. Abiejų lentelių įverčiai sutapo, bet taip įvyko tik todėl, kad modelyje buvo vienintelis kategorinis kintamasis G. Jeigu modelyje būtų intervalinių kintamųjų, tai lentelėje *Descriptive Statistics* būtų grupių vidurkiai, kai į intervalinius kintamuosius neatsižvelgiama, o lentelėje *Estimates* – kai atsižvelgiama (koreguotieji vidurkiai). Plačiau apie koreguotuosius vidurkius jau rašėme ankstesniame skyrelyje. Vidurkių trendo grafikas pateiktas 1.1.11 paveikslo apačioje.

Hipotezės, kad visų mėnesių vidutinių išmokų dydis yra toks pat, tikrinimo rezultatą matome lentelėje *Tests of Between-Subjects Effects* (žr. 1.1.9 pav.). Kadangi $p = 0{,}588 > 0{,}05$, tai hipotezės apie vidurkių lygybę neatmetame. Darome išvadą, kad statistiškai reikšmingų skirtumų tarp vidutinių mėnesinių išmokų nėra.

Bonferonio kriterijaus pritaikyti nepavyko, nes antro mėnesio buvo tik vienas stebėjimas (per mažai duomenų, žr. komentarą 1.1.9 pav. viršuje). Jeigu duomenų yra pakankamai, tai gaunama lentelė, kuri niekuo nesiskiria nuo analogiškos dispersinės analizės lentelės.

Hipotezės $\mu_1 + \mu_2 = 2\mu_3$ tikrinimo rezultatus matome 1.1.10 paveiksle. Iš pradžių užrašoma taikyta kontrastų matrica. Lentelėje *Test results* galima rasti tikrinamos hipotezės p reikšmę. Kadangi $p = 0{,}374$, tai duomenys statistiškai reikšmingai neprieštarauja spėjimui, kad $\mu_1 + \mu_2 = 2\mu_3$ (reikšmingumo lygmuo $\alpha = 0{,}05$).

Lentelėje *Contrast Results* pateikiama informacija apie kiekvienos kontrastų matricos eilutės ir $\boldsymbol{\beta}$ sandaugą. Taip pat patikrinama hipotezė, ar ši sandauga statistiškai reikšmingai skiriasi nuo nulio. Nagrinėjamuoju atveju kontrastų matrica turi tik vieną eilutę, todėl p reikšmė sutampa su gautąja lentelėje *Test Results*.

Taigi jau galime atsakyti į iškeltus klausimus. Reikšmingumo lygmuo $\alpha = 0{,}05$.

Atsakymai į suformuluotas problemas:

1. Vidutinės mėnesinės išmokos yra 6, 7 ir 5.
2. Hipotezė, kad visų mėnesių vidutinių išmokų dydis yra toks pat, nebuvo atmesta.
3. *Post hoc* testų nepavyko pritaikyti, nes antrojo mėnesio stebėjimų buvo per mažai.
4. Hipotezė, kad išmokų vidurkius sieja lygybė $\mu_1 + \mu_2 = 2\mu_3$, nebuvo atmesta.

Dar šiek tiek pakomentuosime, kaip sudaromi kontrastai ir kaip jie taikomi. Programoje parašėme TEST(LMATRIX), todėl tarp gautųjų rezultatų nurodytos kontrastų matricos, kurias naudojome pagrindinėms hipotezėms tikrinti (žr. 1.1.9 pav.). Matome, kad hipotezei apie vidurkių lygybę tikrinti buvo naudojama kontrastų matrica

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Lygybė $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ iš tikrųjų ekvivalenti dviem lygybėms: $\mu_1 - \mu - \mu_3 + \mu = 0$ ir $\mu_2 - \mu - \mu_3 + \mu = 0$, arba $\mu_1 = \mu_2 = \mu_3$.

Kontrastai taikomi ir tikrinant, ar grupių vidurkiai sudaro tiesinį arba kvadratinį trendą. Norėdami patikrinti, ar vidurkiai μ_1, μ_2 ir μ_3 sudaro tiesinį trendą, turime naudoti kontrastą $\mathbf{L} = (0, -1, 0, 1)$. Vis dėlto ne visi kontrastai dispersinėje analizėje prasmingi. Taip yra todėl, kad modelio parametrai tarpusavyje susiję. Nagrinėto pavyzdžio koeficientus sieja lygybė $\mu = (\mu_1 + \mu_2 + \mu_3)/3$. Todėl $\mu_3 - \mu = -(\mu_1 - \mu + \mu_2 - \mu)$. Rašant kontrastą, į tai reikia atsižvelgti. Todėl reikalaujama, kad $\mathbf{L} = (L_0, L_1, L_2, L_3)$ koordinatės tenkintų sąryšį $\mathbf{L} = (L_0, L_1, L_2, L_0 - L_1 - L_2)$. Pavyzdžiui, kontrastas $\mathbf{L} = (0, 1, 1, -3)$ yra netinkamas.

Kai kintamieji intervaliniai, tinka visi kontrastai.

1.1.14 pavyzdys. Užrašysime 1.1.6 pavyzdžio apie Lietuvos ekonominius rodiklius SPSS programos sintaksę. Darbo užmokesčio logaritmo (*LDU*) regresijos lygtis yra tokia:

$$LDU = \beta_0 + \beta_1 LDSK + \beta_2 LDEF + e;$$

čia *LDSK* žymi sąlyginio darbuotojų skaičiaus logaritmą, o *LDEF* – prekių ir paslaugų eksporto defliatoriaus logaritmą. Norint patikrinti hipotezę $\beta_1 + 2\beta_2 = 0$, reikia naudoti kontrastą $\mathbf{L} = (0, 1, 2)$, o SPSS procedūra užrašoma taip:

```
UNIANOVA LDU WITH LDSK LDEF
/METHOD = SSTYPE(3) /INTERCEPT = INCLUDE
/PRINT = DESCRIPTIVE /LMATRIX
Intercept 0 LDSK 1 LDEF 2
/CRITERIA = ALPHA(.05) /DESIGN = LDSK LDEF .
```

Užrašysime, kaip atrodo UNIANOVA procedūra, kai nagrinėjami ANCOVA modeliai. Nagrinėjant 1.1.9 pavyzdį, ši procedūra užrašoma taip:

```
UNIANOVA REZ BY GR WITH MOT
/METHOD = SSTYPE(3) /INTERCEPT = INCLUDE
/PRINT = DESCRIPTIVE /CRITERIA = ALPHA(.05)
/DESIGN = MOT GR .
```

Norint palyginti koreguotuosius vidurkius, reikia įtraukti

```
/EMMEANS = TABLES(GR) WITH(MOT=MEAN)
COMPARE ADJ(BONFERRONI)
```

Tikrinant, ar yra kintamųjų *GR* ir *MOT* sąveika, paskutinį programos sakinį reikia pakeisti tokiu: /DESIGN = GR MOT GR*MOT.

1.8. Tiesinių Gauso modelių analizė naudojantis R ir SAS programomis

Kompiuteris padeda daryti klaidas greičiau nei bet koks kitas žmonijos išradimas, išskyrus galbūt ginklus ir tekilą.

M. Ratlifas

Trumpai aprašysime programų R ir SAS procedūras, skirtas LNM analizei. Sintaksės pavyzdžių pateiksime, nagrinėdami išmokų uždavinį. Manome, kad tai padės skaitytojams lengviau suprasti R ir SAS sintaksių ypatumus.

LNM analizė taikant R programą. Šioje programoje LNM analizei yra naudojama funkcija *lm*. Jos sintaksė yra tokia:

```
lm(formula, data, subset, weights, na.action,
method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Parametras *formula* yra būtinas. Juo nurodomas modelio pavidalas. Pavyzdžiui, užrašas $Y \sim X$ reiškia, kad turi būti analizuojamas paprasčiausias tiesinės regresijos modelis $Y = \beta_0 + \beta_1 x + e$. Nagrinėjant dvifaktorės dispersinės analizės modelį, parametro *formula* reikšmė užrašoma taip: $Y \sim A + B + A : B$ (arba $Y \sim A * B$); čia Y yra priklausomas kintamasis, A, B – faktoriai. Beje, modeliuose laisvasis narys yra numatytasis.

Pateiksime uždavinio apie penkias draudimo išmokas sprendimą su R programa ir paaiškinsime naudotus sakinius bei parametrus. Tarkime, kad norime ištirti tas pačias ankstesniame skyrelyje suformuluotas problemas: 1) įvertinti vidutines kiekvieno mėnesio išmokas; 2) patikrinti hipotezę, kad visų mėnesių vidutinių išmokų dydis yra toks

pat; 3) *post hoc* testais nustatyti, kurių mėnesių vidutinės išmokos skiriasi statistiškai reikšmingai; 4) patikrinti hipotezę, kad išmokų vidurkius sieja lygybė $\mu_1 + \mu_2 = 2\mu_3$.

R programa yra tokia:

```
y=c(5,7,7,4,6)
G=factor(c(1,1,2,3,3))
anv=lm(Y ~ G)
summary(anv)
library(gmodels)
fit.contrast(anv,G,c(1,1,-2))
tapply(y, G, mean)
model.matrix(anv)
```

Pirmieji du sakiniai skirti duomenims įvesti. Sakinys *anv=lm(Y ~ G)* nurodo, kad bus atliekama vienfaktorė dispersinė analizė ir remiantis gautais rezultatais sukuriamas objektas *anv*. Nurodžius *summary(anv)*, spausdinamas modelio suderinamumo matas (R kvadratas), parametrų įverčiai ir hipotezių tikrinimo apie parametrų lygybę nuliui rezultatai. Hipotezė $\mu_1 + \mu_2 = 2\mu_3$ tikrinama naudojantis funkcija *fit.contrast(anv,G,c(1,1,-2))*. Sakiniu *library(gmodels)* nurodoma funkcijų biblioteka, kurioje yra *fit.contrast* funkcija. Nurodžius *tapply(y, G, mean)*, spausdinami priklausomojo kintamojo vidurkiai, esant fiksuotoms faktoriaus G reikšmėms. Sakinys *model.matrix(anv)* nurodo, kad bus spausdinama modelio struktūros matrica.

Skaičiavimo rezultatų nepateikiame, nes jie sutampa su anksčiau pateiktais SPSS rezultatais. R programos sintaksės fragmentus, skirtus kitų šio skyriaus uždavinių sprendimui, galima rasti internete (***http://stat.vadoveliai.lt***).

LNM analizė naudojantis SAS programa. Šioje programoje LNM analizei skirta PROC GLM procedūra. Žinome, kad tiek tiesinės regresijos modeliai, tiek dispersinės analizės modeliai yra atskiri LNM atvejai. Taigi yra labai daug uždavinių, kuriems šią procedūrą galima naudoti. SAS programoje dispersinės analizės ir regresijos modeliams dar yra procedūros PROC ANOVA ir PROC REG. Taigi kokią procedūrą reikėtų rinktis konkrečiam uždaviniui spręsti? Bendresnę PROC GLM procedūrą verta rinktis tada, kai norima duomenims taikyti kelis tiesinius modelius (tik truputį modifikavus PROC GLM, galima atlikti skirtingas analizes) arba kai klasikinės prielaidos (pvz., paklaidų nepriklausomumo sąlyga) negalioja.

Čia nevardysime visų PROC GLM galimybių. Mūsų tikslas – pateikti skaitytojui bazinius PROC GLM elementus. Juos žinodamas, tyrėjas (tikimės) galės atlikti tiesinio modelio analizę. Detalesnį šios procedūros, kaip, beje, ir kitų SAS procedūrų, aprašymą galima rasti SAS programos techninėje dokumentacijoje.

Pirmiausia pastebėsime, kad tik sakinys MODEL procedūroje GLM yra privalomas. Plačiau aprašysime sakinius CLASS, MODEL, CONTRAST, ESTIMATE, LSMEANS. Jie yra naudojami daugelyje SAS procedūrų, skirtų tiesinių (ir, kaip matysime vėliau, bendresnių) modelių analizei.

1.1.3 lentelė. SAS sintaksė LNM tirti

Sakinys	Aprašymas	Parametrai
PROC GLM	Iškviečiama (paleidžiama) procedūra	DATA=duomenų rinkinys; METHOD=analizės metodas;
BY	Vienu iškvietimu PROC GLM vykdomas daug kartų	Nėra
CLASS	Išvardijami faktoriai (kategoriniai, grupavimo kintamamieji). Modelyje jie transformujami į lygmenų indikatorius	Nėra
ID	Nurodomas sąrašas papildomų kintamųjų, įtraukiamų į prognozuojamų reikšmių lenteles	Nėra
MODEL	Užrašoma modelio struktūra (išvardijami priklausomi ir nepriklausomi kintamieji)	SOLUTION – išvedami parametrų įverčių skaičiavimo rezultatai, E – spausdinama įvertinamų funkcijų bendroji forma (naudinga rašant CONTRAST ir ESTIMATE sakinius)
CONTRAST	Aprašomos vartotojo suformuluotos hipotezės apie tiesinius parametrų darinius	E – parodoma kontrastų matrica **L**
ESTIMATE	Įvertinamas $\mathbf{L}\boldsymbol{\beta}$ (pasirinktas vektorinis kontrastas)	CL – išvedami pasikliovimo rėžiai
LSMEANS	Skaičiuojami priklausomo kintamojo mažiausiųjų kvadratų įverčiai, esant fiksuotoms nurodytų faktorių reikšmėms	DIFF – skaičiuojami mažiausiųjų kvadratų vidurkių skirtumai, esant fiksuotoms faktorių reikšmėms

Pateikiame keletą MODEL taikymo pavyzdžių, liudijančių, kad su PROC GLM analizuojamų modelių įvairovė įspūdinga (žr. 1.1.4 lentelę). Grupavimo kintamieji (faktoriai) žymimi raidėmis A, B, C, intervaliniai kintamieji – simboliais X1, X2, X3.

1.1.4 lentelė. MODEL taikymo pavyzdžiai

MODEL Y = X1 X2 X3	Daugialypė tiesinė regresija
MODEL Y = X1 X1*X1 X1*X1*X1	Polinominė regresija
MODEL Y = A	Vienfaktorė dispersinė analizė
MODEL Y = A B	Dvifaktorė dispersinė analizė nesant faktorių sąveikos
MODEL Y = A B A*B	Dvifaktorė dispersinė analizė esant faktorių sąveikai
MODEL Y = A B C A*B A*C B*C A*B*C	Trifaktorė dispersinė analizė

Sakinys CONTRAST naudojamas hipotezėms $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ tikrinti. Daugiausia neaiškumų, kaip turi būti užrašomas šis sakinys, kyla norint patikrinti hipotezes, keliamas dispersinės analizės uždaviniams. Tačiau atsiminkime, kad ne visi kontrastai tam tinkami. Pavyzdžiui, nagrinėtojo pirmojo vienfaktorės dispersinės analizės modelio kontrastas $\mathbf{L} = (L_1, L_2, L_3, L_4)$ turi tenkinti sąryšį $L_4 = L_1 - L_2 - L_3$. SAS programoje yra galimybė išsispausdinti galimas matricos **L** formas. Grįžkime prie mūsų nagrinėjamo pavyzdžio. Žinome, kad $\boldsymbol{\beta}' = (\mu, \mu_1 - \mu, \mu_2 - \mu, \mu_3 - \mu)$. Norime patikrinti hipotezę, kad $\mu_1 + \mu_2 - 2\mu_3 = 0$. Parenkame $\mathbf{L} = (0, 1, 1, -2)$. Sakinys užrašomas taip:

```
CONTRAST ' mu1+mu2=2mu3 ' intercept 0 G 1 1 -2;
```

Kokia sakinio ESTIMATE paskirtis? Jis papildo sakinį CONTRAST, kuris tikrina hipotezes. ESTIMATE padeda rasti $\mathbf{L}\boldsymbol{\beta}$ *įverčius*. Naudojami tie patys kontrastai, kaip ir sakinyje CONTRAST. Pavyzdžiui,

```
ESTIMATE 'mu1+mu2-2mu3' intercept 0 G 1 1 -2;
```

Jeigu norime įvertinti $\mu_1 - \mu_2$, tai rašome

```
ESTIMATE 'lygmenų G1, G2 skirtumo įvertis' G 1 -1;
```

Nenurodytas daugiklis laikomas lygiu nuliui. Tą patį rezultatą gautume užrašę taip:

```
ESTIMATE 'lygmenų G1, G2 skirtumo įvertis' intercept 0 G 1 -1 0;
```

Sakinys LSMEANS naudingas, jei modelyje yra faktorių (kategorinių kintamųjų). Nagrinėjant vienfaktorės dispersinės analizės uždavinius, svarbūs vidurkių įverčiai. Jiems apskaičiuoti procedūros GLM sakinio LSMEANS sintaksė yra tokia: LSMEANS G.

Surinkę visus minėtus sakinius, gauname tokią SAS programėlę:

```
PROC GLM
DATA =work.ismokos;
CLASS G;
MODEL y=G /E;
CONTRAST ' mu1+mu2=2mu3 ' intercept 0 G 1 1 -2;
ESTIMATE 'mu1+mu2-2mu3' intercept 0 G 1 1 -2
ESTIMATE 'lygmenų G1, G2 skirtumo įvertis' G 1 -1;
LSMEANS G;
run;
```

ANCOVA	Gauso modelis	mažiausiųjų kvadratų metodas
didžiausiojo tikėtinumo metodas	kontrastų matrica	struktūros matrica

UŽDAVINIAI

1. Naudodamiesi apibendrintąja atvirkštine matrica (žr. 1.3 skyrelį), raskite pirmojo draudimo išmokų modelio parametrų įverčius.

2. Užrašykite pirmąjį draudimo išmokų modelį matricine forma, naudodami tokią struktūros matricą:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Kaip galima interpretuoti $\boldsymbol{\beta}$ koordinates?

3. Pasinaudokite 2 uždavinio duomenimis ir patikrinkite hipotezę $\mu_1+\mu_2 = 2\mu_3$. Reikšmingumo lygmuo $\alpha = 0{,}05$.

4. Turime klasikinį LNM su nekoreliuojančiomis paklaidomis: $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. Įrodykite, kad mažiausiųjų kvadratų LNM įvertis $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ turi tokias savybes:

$$\mathbf{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta},$$
$$\mathbb{C}\mathrm{ov}\,\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

5. Pasinaudokite failo **stabdymas.sav** duomenimis ir ištirkite, ar stabdymo kelias (kintamasis *ssustojo*) priklauso nuo dangos (kintamasis *danga*). Be to, atsižvelkite į automobilio greitį (kintamasis *greitis*).

6. Norėdamas ištirti dviejų rūšių trąšų efektyvumą, sodininkas mėgėjas patręšė savo balkone vazonėliuose auginamus svogūnus. Svogūnai buvo auginami 5 vazonėliuose. Žemės paimtos iš dviejų skirtingų dirvožemių. Skaičiuota, kiek svogūnų laiškų pavyko nuskinti. Gauta:

Trąša	Dirvožemis	
I	33 34	25
II	35	32

Tai tipinis dvifaktorės analizės modelis. Jo matricinis pavidalas yra toks:

$$\begin{pmatrix} 33 \\ 34 \\ 25 \\ 35 \\ 32 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_{1\cdot} \\ \tau_{2\cdot} \\ \tau_{\cdot 1} \\ \tau_{\cdot 2} \\ \tau_{11} \\ \tau_{12} \\ \tau_{21} \\ \tau_{22} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix};$$

čia

$$\mu = \frac{\mu_{11}+\mu_{12}+\mu_{21}+\mu_{22}}{4}, \qquad \mu_{\cdot j} = \frac{\mu_{1j}+\mu_{2j}}{2},$$
$$\mu_{i\cdot} = \frac{\mu_{i1}+\mu_{i2}}{2}, \qquad \tau_{i\cdot} = \mu_{i\cdot} - \mu, \qquad \tau_{\cdot j} = \mu_{\cdot j} - \mu,$$
$$\tau_{ij} = (\mu_{ij} - \mu) - (\mu_{i\cdot} - \mu) - (\mu_{\cdot j} - \mu), \quad i = 1, 2, \quad j = 1, 2.$$

Užrašykite redukuotą modelį ir kontrastą dirvožemio ir trąšų sąveikai patikrinti.

2. APIBENDRINTIEJI TIESINIAI MODELIAI

Idėja – tai apibendrinimas, o gebėjimas apibendrinti yra mąstymo savybė. Apibendrinti – reiškia mąstyti.

G. Hėgelis

Troškimas kurti kuo bendresnius, visaapimančius modelius įaugęs žmonėms į kraują. Prisiminkime mobiliojo telefono raidą. Iš pradžių sąvoka *mobilusis telefonas* reiškė nešiojamąjį telefoną. Dabar tai gali reikšti „penkis viename“: telefoną + fotoaparatą + muzikos grotuvą + interneto naršyklę + žaidimų automatą. Netrukus galima tikėtis aparato su skutimosi mašinėle (vyrams), pudrine (moterims), lenktiniu peiliuku (turistams), pinigine, žirklėmis, veidrodėliu, inhaliatoriumi nuo slogos, žibintuvėliu, žiebtuvėliu, aparatu kraujospūdžiui matuoti, elektrošoku (gynybai), masažuokliu (pavargusiems), didinamuoju stiklu (besivaržantiems tolregystės), lūpine armonikėle (muzikaliems) ir *jo-jo* (sportiškiems).

Panašiai ir statistikoje. Modeliai tampa vis universalesni, bendresni, sudėtingesni ir mažiau suprantami. Ankstesniame skyriuje įsitikinome, kad daug klasikinių statistikos modelių galima aprašyti kaip LNM. Tačiau – ne visus. Pavyzdžiui, logistinė regresija nėra Gauso modelis. Šiame skyriuje susipažinsime su apibendrintaisiais tiesiniais modeliais (GLM). Kodėl prireikia GLM? Gauso modelių prielaidos yra labai griežtos. Natūralu, kad sprendžiant daugelį praktinių uždavinių jos negalioja ir tada prireikia kitokių statistinės analizės metodų.

Nesieksime smulkiai išnagrinėti visų GLM atvejų. Tam prireiktų atskiros knygos. Mūsų tikslas daug kuklesnis – susipažinti su pagrindinėmis sąvokomis ir modelio analizės principais.

2.1. GLM struktūra

Apibendrinti – reiškia būti idiotu. Bendrosios žinios yra tai, ką žino idiotas.

V. Bleikas

Klasikinį LNM modelį galima užrašyti dviem ekvivalenčiomis formomis. Pirmoji aprašo kintamojo Y tiesinę priklausomybę nuo aiškinamųjų kintamųjų X_j, ją nagrinėjome ankstesniame skyriuje. Antroji aprašo kintamojo Y vidurkį μ (užrašysime tik principinę priklausomybės schemą, praleisdami dalį indeksų):

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K. \tag{1.2.1}$$

GLM – tai tiesinis modelis, apibūdinantis priklausomo kintamojo vidurkio funkciją:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K. \tag{1.2.2}$$

Akivaizdu, kad GLM yra LNM apibendrinimas. Funkcija $g(\cdot)$ vadinama *jungties funkcija*[14], arba jungtimi. Ji parenkama atsižvelgiant į tiriamą problemą.

Natūralu, kad modelis turi tenkinti tam tikrus reikalavimus:

1 Priklausomi kintamieji Y turi turėti pakankamai geras savybes. Tiksliau, Y skirstinys turi priklausyti eksponentinių skirstinių šeimai. Plačiau šią šeimą aptarsime kitame skyrelyje.

2 Jungties funkcija g turi būti tolydi ir turėti atvirkštinę funkciją g^{-1} (kad (1.2.2) lygtį būtų galima išspręsti μ atžvilgiu).

Naudojant atvirkštinę jungties funkciją, GLM galima užrašyti kaip vidurkio modelį:

$$\mu = g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K).$$

Šis modelis yra tiesinis tik tada, kai $g(x) = x$.

Yra nusistovėję GLM sudėtinių dalių pavadinimai. Priklausomas kintamasis Y dar vadinamas *atsitiktine modelio komponente*. Į modelio lygtį įtrauktas jo vidurkis μ. Tiesinė modelio dalis su nežinomais parametrais β_k ir matuojamais kintamaisiais X_k vadinama modelio *prediktoriumi*[15]. Dažnai prediktoriui žymėti vartojama santrumpa $\eta(\boldsymbol{\beta})$:

$$\eta(\boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K. \tag{1.2.3}$$

Kol kas užrašėme principinį vieno priklausomo kintamojo Y modelį. Bendruoju atveju užrašomas kintamojo Y_j, kuris yra imties $(Y_1, Y_2, \ldots, Y_n)$ narys (atsitiktinis dydis, o ne realizacija), modelis.

Bendrasis GLM modelis užrašomas taip:

$$g(\mu_j) = \beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \cdots + \beta_K X_{jK}, \quad j = 1, 2, \ldots, n.$$

Čia:

- $\mu_j = \mathbf{E}Y_j$, o Y_j skirstinys priklauso eksponentinių skirstinių šeimai.
- Dešiniojoje lygybės pusėje esantis reiškinys vadinamas modelio prediktoriumi ir žymimas $\eta_j(\boldsymbol{\beta})$:

$$\eta_j(\boldsymbol{\beta}) = \beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \ldots + \beta_K X_{jK}.$$

- Jungties funkcija $g(\cdot)$ turi atvirkštinę, t. y. $\mu_j = g^{-1}(\eta_j(\boldsymbol{\beta}))$.
- Parametrai $\boldsymbol{\beta}$ nežinomi, juos reikia įvertinti.
- Gali reikėti įvertinti ir kai kuriuos Y_j skirstinio parametrus.

[14] Angl. *Link Function*.

[15] Angl. *Predictor*.

Jeigu norėtume GLM užrašyti matricine forma, tai reikėtų tiesiog susitarti, kad $g(\boldsymbol{\mu}) := (g(\mu_1), g(\mu_2), \ldots, g(\mu_n))'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$. Tada gautume, kad

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta};$$

čia, kaip ir LNM atveju, $\mathbf{X}$ yra struktūros matrica, o $\boldsymbol{\beta}$ – parametrų vektorius. Dažnai matricos $\mathbf{X}$ j-oji eilutė dar žymima simboliu $\mathbf{X}'_j$. Tada $\eta_j(\boldsymbol{\beta}) = \mathbf{X}'_j\boldsymbol{\beta}$, o GLM modelis užrašomas

$$g(\mu_j) = \mathbf{X}'_j\boldsymbol{\beta}.$$

Toliau šiame skyriuje indeksus rašysime, kai norėsime pabrėžti, kad tai stebėjimas. Aprašydami kintamųjų savybes, indeksus praleisime.

***GLM sudarymo ir tyrimo etapai*:**

1. Formuluojama prielaida apie priklausomo kintamojo Y skirstinį.
2. Atsižvelgiant į Y skirstinio struktūrą, parenkama jungties funkcija $g(\cdot)$.
3. Parenkami regresoriai $X_1, X_2, \ldots, X_K$. Sudaroma lygtis.
4. Įvertinami koeficientai $\boldsymbol{\beta}$ ir nežinomi priklausomo kintamojo Y skirstinio parametrai.
5. Tikrinama, ar modelis tinka imties duomenims.
6. Interpretuojami gauti rezultatai.

Kituose skyreliuose susipažinsime su prielaidomis, kurias turi tenkinti Y, ir parametrų vertinimo metodais.

2.2. Eksponentinių skirstinių šeima

Matematika primena meilę – idėja paprasta, bet galimos komplikacijos.

R. Drabekas

Ne bet koks priklausomo kintamojo Y skirstinys tinkamas GLM. Reikalaujama, kad jis būtų iš eksponentinių skirstinių šeimos. Kokia tai šeima? Skirstinys priklauso eksponentinių skirstinių šeimai, jeigu jo tankis (diskrečiuoju atveju – tikimybė) yra tokios struktūros:

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}; \tag{1.2.4}$$

čia θ, ϕ – parametrai, o $a(\phi)$, $b(\theta)$, $c(y, \phi)$ – žinomos funkcijos. Labai dažnai $a(\phi) = \phi/p$; čia p yra žinomas skaičius, vadinamas aprioriniu svoriu. Daugeliu atvejų $p = 1$. Parametrai θ ir ϕ yra susiję su Y vidurkiu ir dispersija:

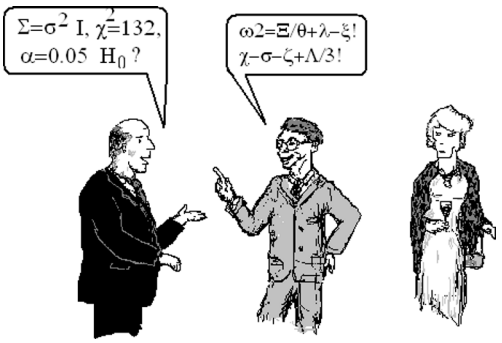
$$\mathbf{E}Y = \mu = b'(\theta), \qquad \mathbf{D}Y = \sigma^2 = b''(\theta)a(\phi). \tag{1.2.5}$$

Užrašas (1.2.4) yra pakankamai bendras, todėl eksponentinių skirstinių šeima labai didelė. Iš tikrųjų jai priklauso normalusis, binominis, Puasono, gama ir dar daug kitų

skirstinių. Vieningas daugelio skirstinių traktavimas leidžia sukurti *visiems* jiems tinkamus įverčių skaičiavimo ir hipotezių tikrinimo metodus. Taigi ir įvairių regresijos modelių parametrų įverčiams skaičiuoti tinka *viena* programa.

Sakome, kad turime kanoninę jungties funkciją, jeigu $\theta = g(\mu)$.

Didelė simbolių ir santrumpų gausa knygoje gali baigtis tuo, kad parašytą tekstą supras tik autoriai (ir tai tik tol, kol jį rašys). Todėl stengsimės tokiais užrašais nepiktnaudžiauti ir į smulkias kiekvieno parametro interpretacijas bendruoju atveju nesivelti.



1.2.1 pavyzdys. Įsitikinsime, kad normalųjį skirstinį galima priskirti prie eksponentinių skirstinių. Tarkime, kad $Y \sim \mathcal{N}(\mu, \sigma^2)$. Žinoma, kad $\mathbf{E}Y = \mu$, o tankis

$$f\big(y; \mu; \sigma^2\big) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\Big\{\frac{-(y-\mu)^2}{2\sigma^2}\Big\}.$$

Perrašykime tankio funkciją taip:

$$\begin{aligned} f\big(y; \mu; \sigma^2\big) &= \exp\Big\{-\ln\big((2\pi\sigma^2)^{1/2}\big)\Big\} \exp\Big\{\frac{-(y-\mu)^2}{2\sigma^2}\Big\} \\ &= \exp\Big\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} - \frac{y^2}{2\sigma^2}\Big\}. \end{aligned}$$

Palyginę ją su (1.2.4) formule, matome, kad normaliojo dydžio $\theta = \mu$, $\phi = \sigma^2$, $b(\mu) = \mu^2/2$, $a(\phi) = a(\sigma^2) = \sigma^2$ ir

$$c(y, \phi) = -\frac{\ln(2\pi\sigma^2)}{2} - \frac{y^2}{2\sigma^2}.$$

Taigi kanoninė jungties funkcija yra $g(\mu) = \mu$, o modelį nusakanti lygtis $\mu = \eta(\boldsymbol{\beta})$ sutampa su LNM modelį aprašančia lygtimi.

Jeigu Y yra normalusis kintamasis, tai GLM = LNM.

Priklausomo kintamojo Y skirstinys lemia galimą jungties funkcijos, tuo pačiu ir modelio, pavidalą.

1.2.2 pavyzdys. Įsitikinsime, kad Puasono skirstinys yra eksponentinių skirstinių šeimos atstovas. Žinome, kad $Y \sim \mathcal{P}(\mu)$ vidurkis $\mathbf{E}Y = \mu$, o tikimybė įgyti konkrečią reikšmę y yra

$$P(Y = y) = \frac{\mu^y}{y!}\mathrm{e}^{-\mu}, \quad y = 0, 1, 2, \ldots.$$

Perrašykime šią tikimybę taip:

$$P(Y=y)=\exp\Big\{\ln\frac{\mu^y\mathrm{e}^{-\mu}}{y!}\Big\}=\exp\Big\{\frac{y\ln\mu-\mu}{1}-\ln(y!)\Big\}=\exp\Big\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y)\Big\};$$

čia $\theta=\ln\mu$, $b(\theta)=\mathrm{e}^\theta$, $a(\phi)=1$, $c(y,\phi)=-\ln(y!)$, $\phi=1$. Matome, kad kanoninė jungties funkcija $g(\mu)=\ln\mu$. Dabar jau galime užrašyti apibendrintąjį tiesinį modelį Puasono skirstinio atveju:

$$\ln\mu=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_KX_K. \tag{1.2.6}$$

Analogiškai randamos ir kitų skirstinių jungties funkcijos. Kelių svarbesnių skirstinių kanoninių jungties funkcijų pavyzdžiai pateikiami 1.2.1 lentelėje. Paskutiniame lentelės stulpelyje parodyta, kaip rasti vidurkio įvertį, kai turimi koeficientų įverčiai $\hat{\boldsymbol{\beta}}$ ir žinomos nepriklausomų kintamųjų (regresorių) reikšmės $\mathbf{X}$.

1.2.1 lentelė. Kanoninės jungties funkcijos

Skirstinys	Funkcijos pavadinimas	Jungties funkcija	Modelis	Vidurkio įvertinys
Normalusis	Identiška	$g(\mu)=\mu$	$\mu=\mathbf{X}\boldsymbol{\beta}$	$\hat\mu=\mathbf{X}\hat{\boldsymbol{\beta}}$
Eksponentinis	Atvirkštinė	$g(\mu)=\mu^{-1}$	$\mu^{-1}=\mathbf{X}\boldsymbol{\beta}$	$\hat\mu=(\mathbf{X}\hat{\boldsymbol{\beta}})^{-1}$
Puasono	Log	$g(\mu)=\ln\mu$	$\ln\mu=\mathbf{X}\boldsymbol{\beta}$	$\hat\mu=\exp\{\mathbf{X}\hat{\boldsymbol{\beta}}\}$
Binominis	Logit	$g(\mu)=\ln\frac{\mu}{1-\mu}$	$\ln\frac{\mu}{1-\mu}=\mathbf{X}\boldsymbol{\beta}$	$\hat\mu=\frac{\exp\{\mathbf{X}\hat{\boldsymbol{\beta}}\}}{1+\exp\{\mathbf{X}\hat{\boldsymbol{\beta}}\}}$

2.3. GLM formuluočių pavyzdžiai

Medicina žmones susargdina, matematika juos liūdina, o teologija – paverčia nuodėmingais.
M. Liuteris

Pateiksime keletą pavyzdžių, kada pravartu taikyti apibendrintuosius tiesinius modelius. Modelių užrašuose praleisime indeksus, palikdami tik esmines priklausomybes atskleidžiančius dydžius. Pradėsime nuo vieno populiaresnių GLM atvejų – klasikinio logistinės regresijos modelio.

1.2.3 pavyzdys. Norime įvertinti pakartotinio susirgimo skrandžio vėžiu tikimybę, praėjus 6 metams po operacijos. Atsižvelgiame į respondento amžių (*AMŽIUS*), išoperuoto auglio didumą (*AUGLYS*) ir lytį (*LYTIS*). Priklausomas kintamasis Y įgyja tik dvi reikšmes: 0 (jeigu respondentas nesuserga) ir 1 (jeigu respondentas suserga). Nesunku suprasti, kad $\mathbf{E}Y=P(Y=1)=P(\textit{respondentas susirgs})$. Kintamojo Y skirstinys yra paprasčiausias binominio skirstinio atvejis, todėl naudodamiesi 1.2.1 lentele galime suformuluoti tokį modelį:

$$\ln\frac{P(Y=1)}{P(Y=0)}=\beta_0+\beta_1AMŽIUS+\beta_2AUGLYS+\beta_3LYTIS.$$

Jeigu statistinė modelio analizė parodytų, kad šis modelis gerai tinka duomenims, tai respondento susirgimo tikimybę galima būtų įvertinti naudojantis formule

$$P(Y=1)=\frac{\exp\{\hat\beta_0+\hat\beta_1AMŽIUS+\hat\beta_2AUGLYS+\hat\beta_3LYTIS\}}{1+\exp\{\hat\beta_0+\hat\beta_1AMŽIUS+\hat\beta_2AUGLYS+\hat\beta_3LYTIS\}};$$

čia $\hat\beta_0,\hat\beta_1,\hat\beta_2,\hat\beta_3$ yra modelio koeficientų įverčiai.

Logistinė regresija taikoma, kai modeliuojama, kokia yra tikimybė išlaikyti vairavimo egzaminą; kai norima išsiaiškinti, ar bus perkama nauja prekė, ar numatomas firmos bankrotas; kai reikia įvertinti blogos banko paskolos tikimybę ir pan. Išsamiai ši logistinė regresija nagrinėjama [3] knygoje.

Logistinės regresijos modelis, apibrėžtas tikimybių santykio logaritmu (funkcija logit) nėra vienintelis modelis, kurį galima taikyti, kai priklausomas kintamasis įgyja tik dvi reikšmes. Suformuluosime vadinamąjį *probit* modelį.

1.2.4 pavyzdys. Ankstesniojo pavyzdžio duomenims pritaikysime tokį modelį:

$$P(Y = 1) = \Phi(\beta_0 + \beta_1 AMŽIUS + \beta_2 AUGLYS + \beta_3 LYTIS);$$

čia $\Phi(\cdot)$ yra standartinio normaliojo kintamojo pasiskirstymo funkcija. *Probit* modelio jungties funkcija yra $g(\cdot) = \Phi^{-1}(\cdot)$. Norėdami pritaikyti šį modelį prognozėms, turėtume tiesiog į jo išraišką įstatyti koeficientų įverčius $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$. Taikant *probit* ir *logit* modelius, dažniausiai gaunamos labai panašios statistinės išvados. Modelis *logit* geriau tinka, kai aiškinamieji kintamieji nėra normalieji.

Dar viena apibendrintųjų tiesinių modelių klasė – Puasono regresijos modeliai. Jeigu stebime *retus* įvykius, prielaida, kad kintamieji yra normalieji, nėra natūrali. Viena iš alternatyvų – vadinamoji Puasono regresija, kuri yra atskiras GLM atvejis. Puasono regresija taikoma, kai modeliuojamas retų įvykių *skaičius*.

Išvardysime dar keletą situacijų, kada natūralu išbandyti Puasono regresiją:

1. Modeliuojamas respondento patirtų infarktų skaičius. Atsižvelgiama į jo amžių, lytį ir antsvorį.
2. Modeliuojamas moters vaikų skaičius. Atsižvelgiama į jos amžių ir išsilavinimą.
3. Modeliuojams vaikų, gimusių po dirbtinio apvaisinimo, skaičius. Atsižvelgiama į bendrą gimimų skaičių.
4. Modeliuojamas praleistų statistikos paskaitų[16] skaičius. Atsižvelgiama į respondento pažymius.
5. Modeliuojamas kalinio režimo pažeidimų per mėnesį skaičius. Atsižvelgiama į kalinio amžių ir įkalinimo įstaigose praleistų metų skaičių per ankstesnius teistumus.

[16] Pasvajokim, kad tai būtų retas įvykis.

1.2.5 pavyzdys. Stebime, kiek kartų (kintamasis Y) per trejus metus į draudimo įmonę kreipėsi vairuotojai. Žinome vairuotojų vairavimo stažą (kintamasis S), lytį (kintamasis L), automobilio markę (M) ir automobilio eksploatavimo trukmę (T) periodo pradžioje. Stebime retus įvykius, todėl galima išbandyti Puasono modelį, t. y. tarti, kad $Y \sim \mathcal{P}(\mu)$. Kaip matyti iš 1.2.1 lentelės, reikia sudaryti vidutinio kreipimosi skaičiaus logaritmo tiesinį modelį:

$$\ln \mu = \beta_0 + \beta_1 S + \beta_1 M + \beta_2 T + \beta_3 L.$$

Įvertinus parametrus β_j ir įsitikinus, kad modelis duomenims tinka, jį galima naudoti prognozėms. Vidutinį kreipimųsi į draudimo firmą skaičių galima prognozuoti pagal formulę

$$\mu = \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 S + \hat{\beta}_1 M + \hat{\beta}_2 T + \hat{\beta}_3 L\right\}.$$

Viena iš Puasono skirstinio savybių – dispersija lygi vidurkiui. Jeigu dispersija daug didesnė už vidurkį, tai Puasono regresija nėra tinkamiausias modelis. Tuo atveju sakoma, kad susidūrėme su per didelės dispersijos problema[17]. Dažniausiai tokiems duomenims taikoma *neigiamai binominė* regresija. Išsamiai neigiamas binominis skirstinys aprašytas [2] knygoje. Čia pateiksime jo užrašą, kuris naudojamas GLM. Kintamasis Y turi neigiamą binominį skirstinį, jeigu jo tikimybė skaičiuojama pagal formulę

$$P(Y = y) = \frac{\Gamma(y + 1/\kappa)}{y!\Gamma(1/\kappa)}\left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^{y}, \quad y = 0, 1, 2, \ldots.$$

Be to,

$$\mathbf{E}Y = \mu, \qquad \mathbf{D}Y = \mu(1 + \kappa\mu).$$

Matome, kad κ parodo, kiek daug kintamasis skiriasi nuo Puasono kintamojo. Neigiamas binominis skirstinys priskiriamas prie eksponentinių skirstinių šeimos, todėl galima rasti jo kanoninę jungties funkciją $g(\mu) = \ln \kappa\mu - \ln(1+\kappa\mu)$. Vis dėlto modeliuojant dažniausiai naudojama jungties funkcija $g(\mu) = \ln \mu$. Būtent tokia jungties funkcija taikoma ir daugelyje statistinių programų.

Neigiamai binominės regresijos modelio užrašas niekuo nesiskiria nuo Puasono regresijos užrašo. Gali skirtis tik parametrų įverčių skaičiavimo metodai ir statistinės išvados.

Atkreipiame dėmesį, kad visuose modeliuose gali būti naudojamos kintamųjų transformacijos, pavyzdžiui, galima tirti ir tokius 1.2.5 pavyzdžio problemai spręsti skirtus modelius:

$$\ln \mu = \beta_0 + \beta_1 \ln S + \beta_1 M + \beta_2 \ln T + \beta_3 L$$

arba

$$\ln \mu = \beta_0 + \beta_1 S + \beta_2 S^2 + \beta_3 S \ln T.$$

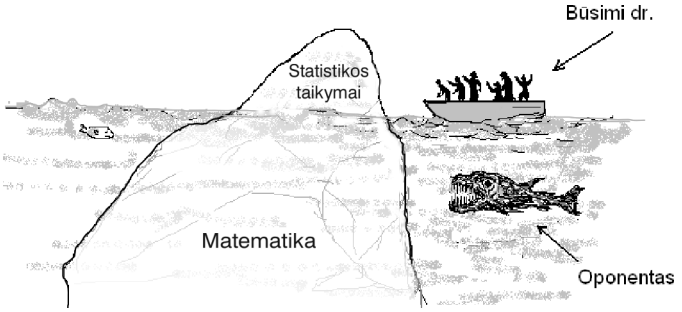
Tiriant tokius modelius statistinėmis programomis, dažniausiai sukuriami nauji kintamieji, pavyzdžiui, $X_1 = \ln T$ ir pan., kurie po to įtraukiami į GLM.

[17] Angl. *Overdispersion*.

2.4. GLM parametrų įverčių skaičiavimas

Auksinė matematikos dėstytojų taisyklė: turi sakyti tiesą, vien tik tiesą, bet ne visą tiesą.

Žinant parametrų $\boldsymbol{\beta}$ įverčius, galima įvertinti ir visą modelį, prognozuoti priklausomo kintmojo Y vidurkio reikšmes, interpretuoti regresorių X_j įtaką. Parametrų $\boldsymbol{\beta}$ įverčiai apskaičiuojami pritaikius didžiausiojo tikėtinumo metodą. Praktiškai naudojami įverčiai, gauti statistinėmis programomis. Vis dėlto norintiems geriau suprasti įverčių skaičiavimo metodus, pateiksime keletą formulių.



Tarsime, kad $Y_1, Y_2, \ldots, Y_n$ yra kintamieji, turintys eksponentinius skirstinius. Didžiausiojo tikėtinumo funkcija yra jų tankių sandauga:

$$L = \exp\left\{\sum_{i=1}^{n}\left[\frac{[y_i\theta_i - b(\theta_i)]}{a_i(\phi)} + c(y_i, \phi)\right]\right\}.$$

Pažymėkime

$$\ell_i(y_i, \mu_i, \phi) = \frac{[y_i\theta_i - b(\theta_i)]}{a_i(\phi)} + c(y_i, \phi).$$

Tada didžiausiojo tikėtinumo funkcijos logaritmas (logtikėtinumo funkcija) užrašoma taip:

$$\ln L = \sum_{i=1}^{n} \ell_i.$$

Taikydami didžiausiojo tikėtinumo metodą, apskaičiuojame šios funkcijos dalines išvestines pagal β_j ir prilyginame nuliui:

$$\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0, \quad j = 1, 2, \ldots, K. \tag{1.2.7}$$

1.2.6 pavyzdys. Prisiminkime pirmojo skyriaus pavyzdį apie draudimo firmos trijų atsitiktinai parinktų mėnesių išmokas (tūkst. litų).

Mėnuo	1	1	2	3	3
Išmoka	5	7	7	4	6

Taikydami vienfaktorės dispersinės analizės modelį, teigiame kad išmokos pasiskirsčiusios pagal normalųjį dėsnį. Ši prielaida nėra natūrali. Labiau pagrįsta manyti, kad išmoka – retas įvykis, tiksliau, kad išmokos didumas pasiskirstęs pagal Puasono dėsnį. Užrašysime puasoninės regresijos modelį ir įvertinsime jo parametrus. Naudodamiesi 1.1.2 skyrelio trečiojo modelio užrašu $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, formaliai užrašome $\ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$; čia

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}. \tag{1.2.8}$$

Užrašykime modelio lygtis nematriciniu pavidalu:

$$\begin{cases} \ln \mu_1 = \beta_1, \\ \ln \mu_1 = \beta_1, \\ \ln \mu_2 = \beta_2, \\ \ln \mu_3 = \beta_3, \\ \ln \mu_3 = \beta_3, \end{cases} \quad \text{arba} \quad \begin{cases} \mu_1 = \mathrm{e}^{\beta_1}, \\ \mu_1 = \mathrm{e}^{\beta_1}, \\ \mu_2 = \mathrm{e}^{\beta_2}, \\ \mu_3 = \mathrm{e}^{\beta_3}, \\ \mu_3 = \mathrm{e}^{\beta_3}. \end{cases}$$

Toks modelio užrašas parodo, kaip išmokų vidurkiai susiję su parametrais $\beta_1, \beta_2, \beta_3$. Užrašysime didžiausiojo tikėtinumo funkciją L. Iš 1.2.2 pavyzdžio matyti, kad

$$\ell_i = Y_i \ln \mu_i - \mu_i - \ln(Y_i!).$$

Todėl

$$\begin{aligned} \ln L = {} & 5\beta_1 - \mathrm{e}^{\beta_1} - \ln(5!) + 7\beta_1 - \mathrm{e}_1^{\beta} - \ln(7!) \\ & + 7\beta_2 - \mathrm{e}^{\beta_2} - \ln(7!) + 4\beta_3 - \mathrm{e}^{\beta_3} - \ln(4!) + 6\beta_3 - \mathrm{e}^{\beta_3} - \ln(6!). \end{aligned}$$

Apskaičiuojame dalines išvestines:

$$\frac{\partial \ln L}{\partial \beta_1} = 12 - 2\mathrm{e}^{\beta_1}, \qquad \frac{\partial \ln L}{\partial \beta_2} = 7 - \mathrm{e}^{\beta_2}, \qquad \frac{\partial \ln L}{\partial \beta_3} = 5 - 2\mathrm{e}^{\beta_3}.$$

Prilyginę šias išvestines nuliui, gauname įverčius:

$$\hat{\beta}_1 = \ln 6, \qquad \hat{\beta}_2 = \ln 7, \qquad \hat{\beta}_3 = \ln 5.$$

Todėl

$$\hat{\mu}_1 = \mathrm{e}^{\hat{\beta}_1} = 6, \qquad \hat{\mu}_2 = \mathrm{e}^{\hat{\beta}_2} = 7, \qquad \hat{\mu}_3 = \mathrm{e}^{\hat{\beta}_3} = 5.$$

Taigi vidurkių įverčiai liko tokie pat, kaip ir LNM atveju. Puasoninės regresijos prielaida įverčių nepakeitė. Taip įvyko todėl, kad modelyje yra tik vienas kategorinis aiškinamasis kintamasis.

Iš pateiktų pavyzdžių matyti, kad nėra sunku rasti įverčius, kai yra tik kategoriniai regresoriai. Kai modelyje atsiranda intervaliniai regresoriai, parametrų įverčiai skaičiuojami iteraciniu būdu. Užrašysime didžiausiojo tikėtinumo funkcijos logaritmą iteraciniam skaičiavimui tinkama forma.

Dalinė išvestinė skaičiuojama pagal tokią formulę:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i}.$$

Kadangi (žr. (1.2.5)) $\mathbf{E}Y_i = \mu_i = b'(\theta_i)$, $\mathbf{D}Y_i = b''(\theta_i)a_i(\phi)$, tai

$$\frac{\partial \ell_i}{\partial \mu_i} = \frac{\partial \ell_i}{\partial \theta_i} : \frac{\partial \mu_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)b''(\theta_i)} = \frac{y_i - \mu_i}{\sigma_i^2}.$$

Viena vertus, $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK}$, todėl $\partial\eta_i/\partial\beta_j = X_{ij}$. Kita vertus, $\eta_i = g(\mu_i)$, todėl

$$\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i).$$

Pažymėję *svorį*

$$W_i = \frac{1}{(g'(\mu_i))^2\sigma_i^2},$$

(1.2.7) lygtį užrašome

$$\sum_{i=1}^{n} W_i(y_i - \mu_i)g'(\mu_i)X_{ij} = 0, \quad j = 1, 2, \ldots, K.$$

Šiose lygtyse sunku įžiūrėti β_j. Tačiau reikia prisiminti, kad $\mu_i = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK})$.

Lygčių sistema sprendžiama iteraciškai – taikant Niutono–Rafsono metodą. Beje, lygčių sistemą galima užrašyti ir matricine forma:

$$\mathbf{X}'\mathbf{W}\mathbf{G}(\mathbf{y} - \boldsymbol{\mu}) = 0; \tag{1.2.9}$$

čia $\mathbf{W}$ yra įstrižainė matrica, kurios pagrindinėje įstrižainėje yra W_i, o $\mathbf{G}$ – įstrižainė matrica, kurios pagrindinėje įstrižainėje yra $g'(\mu_i)$

1.2.7 pavyzdys. Turime 572 Vilniaus universiteto Matematikos ir informatikos fakulteto studentų duomenis: pirmojo matematinės analizės egzamino pažymį (*MA*), mokyklos baigimo atestato vidurkį (*ATVID*), lytį (*LYTIS*, 1 – vyr., 0 – mot.), studijų programos eilės numerį (*NR*, stojant prioritetų sąraše, kai nurodoma 20 pageidavimų) ir informaciją apie bakalauro studijų baigimą (*BID*, 1 – baigė, 0 – nebaigė). Studijuojamos programos prioritetą galima traktuoti kaip tam tikrą motyvacijos rodiklį.

Labai tikėtina, kad respondentai studijuojamą specialybę nurodė viena pirmųjų. Kintamojo *NR* skirstinys nėra simetrinis. Todėl *NR* vargiai gali būti laikomas normaliuoju kintamuoju. Pritaikysime šiam kintamajam neigiamos binominės regresijos modelį, regresoriais imdami atestato vidurkį, matematinės analizės pažymį ir respondento lytį. Mūsų modelis yra toks:

$$\ln(\mathbf{E}NR) = \beta_0 + \beta_1 LYTIS + \beta_2 ATVID + \beta_3 MA.$$

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	4,281	,8128	2,688	5,874	27,738	1	,000
[LYTIS=0]	-,457	,1083	-,669	-,245	17,808	1	,000
[LYTIS=1]	0[a]	.	.	.	.	.	.
MA	-,069	,0326	-,133	-,005	4,409	1	,036
ATVID	-,281	,0864	-,450	-,111	10,550	1	,001
(Scale)	1[b]						
(Negative binomial)	1						

Dependent Variable: NR

Model: (Intercept), LYTIS, MA, ATVID

1.2.1 pav. GLM parametrų įverčiai

Įvertinus parametrus SPSS programa, gautos reikšmės pateiktos 1.2.1 paveiksle. Matome, kad yra nurodytos dvi kintamojo *LYTIS* reikšmės. Kai $LYTIS = 1$ (t. y. vyrai), modelio lygtis yra (suapvalinus įverčių reikšmes)

$$\ln(\mathbf{E}NR)_{\text{vyr}} = 4{,}28 - 0{,}28ATVID - 0{,}07MA.$$

Analogiškai moterims skirta lygtis

$$\ln(\mathbf{E}NR)_{\text{mot}} = (4{,}28 - 0{,}46) - 0{,}28ATVID - 0{,}07MA = 3{,}82 - 0{,}28ATVID - 0{,}07MA.$$

Iš šių lygčių galima gauti eilės numerio vidurkių išraiškas. Gautąsias lygtis jau galima naudoti prognozėms. Tarkime, kad norime įvertinti, kokią vidutiniškai studijuojamos specialybės poziciją prioritetų sąraše nurodė respondentės, kurios išlaikė analizę 7 balais, o jų atestato vidurkis yra 9 balai. Prognozę apskaičiuojame, įstatę šiuos skaičius į formulę:

$$\text{Vidutinis Nr.} = \exp\{3{,}82 - 0{,}28ATVID - 0{,}07MA\} = \exp\{3{,}82 - 0{,}28 \cdot 9 - 0{,}07 \cdot 7\} = 2{,}247\ldots.$$

Analogiškas prognozes darytume respondentams vyrams.

2.5. Hipotezių tikrinimas nagrinėjant GLM

Yra sadistų mokslininkų, kurie skuba ieškoti klaidų, užuot bandę nustatyti tiesą.

M. Kiuri

Nagrinėjant apibendrintuosius tiesinius modelius, kaip ir LNM atveju, galima tikrinti tokias hipotezes: a) ar parametras β_j lygus nuliui, b) ar nuliui lygūs parametrų tiesiniai dariniai. Visur šiame skyrelyje tariama, kad priklausomas kintamasis turi eksponentinės šeimos skirstinį, o jo koeficientų įverčiai apskaičiuoti naudojantis (1.2.9) formule. Norime atkreipti dėmesį, kad praktiškai hipotezės tikrinamos naudojantis statistinėmis programomis.

Hipotezių apie atskirų parametrų lygybę nuliui tikrinimas. Hipotezę, ar β_j lygus nuliui, galima patikrinti naudojantis daugeliu statistinių programų. Ją galima interpretuoti kaip tyrimą, ar reikalingas modelyje kintamasis X_j. Statistikoje šios hipotezės tikrinimas vadinamas Voldo testu.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \beta_j = 0, \\ H_1\colon\ \beta_j \neq 0. \end{cases}$$

2 *Kriterijaus statistika.* Apskaičiuojame

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)};$$

čia $\hat{\beta}_j$ yra β_j įvertis, o $SE(\hat{\beta}_j)$ yra $\hat{\beta}_j$ standartinė paklaida. Ji lygi kvadratinei šakniai iš matricos $a(\phi)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ pagrindinės įstrižainės j-ojo elemento.

3 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo yra α. Nulinę hipotezę atmetame ir darome išvadą, kad $g(\mu)$ statistiškai reikšmingai priklauso nuo X_j, jeigu $|t| > t_\alpha(n - K - 1)$; čia $t_\alpha(n - K - 1)$ yra Stjudento skirstinio su $n - K - 1$ laisvės laipsnių α lygmens kritinė reikšmė.

4 *Sprendimo priėmimo, atsižvelgiant į p reikšmę, taisyklė.* Nulinė hipotezė atmetama (daroma išvada, kad kintamasis X_j modelyje statistiškai reikšmingas), jeigu $p < \alpha$.

Jeigu nulinė hipotezė neatmetama, tai reikėtų pabandyti kintamąjį X_j iš modelio pašalinti.

1.2.8 pavyzdys. Prisiminkime 1.2.7 pavyzdį. Tarkime, kad reikšmingumo lygmuo yra $\alpha = 0{,}05$. Tad, kaip matyti iš 1.2.1 paveikslo, visi modelio kintamieji statistiškai reikšmingi.

Hipotezės apie parametrų tiesinių darinių lygybę skaičiui. Šios hipotezės tikrinamos kaip ir LNM atveju. Tarkime, kad turime kontrastų matricą $\mathbf{L}$.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \mathbf{L}\boldsymbol{\beta} = \mathbf{c}, \\ H_1\colon\ \mathbf{L}\boldsymbol{\beta} \neq \mathbf{c}; \end{cases}$$

čia $\mathbf{c}$ – iš anksto pasirinktas atsižvelgiant į norimą tikrinti hipotezę konstantų vektorius. Dažniausiai $\mathbf{c} = \mathbf{0}$.

2 *Kriterijaus statistika.* Apskaičiuojame

$$T = (\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'\left[a(\phi)\mathbf{L}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}).$$

3 *Sprendimo priėmimo taisyklė.* Nulinė hipotezė atmetama, jeigu $T > \chi^2_\alpha(q)$; čia $\chi^2_\alpha(q)$ yra chi kvadrato skirstinio su q laisvės laipsnių α lygmens kritinė reikšmė, α – pasirinktas reikšmingumo lygmuo, o q – matricos $\mathbf{L}$ eilučių skaičius.

4 *Sprendimo priėmimo, atsižvelgiant į p reikšmę, taisyklė.* Nulinė hipotezė atmetama, jeigu $p < \alpha$.

1.2.9 pavyzdys. Prisiminkime 1.2.7 pavyzdžio modelį:

$$\ln(\mathbf{E}NR) = \beta_0 + \beta_1 LYTIS + \beta_2 ATVID + \beta_3 MA.$$

Patikrinsime hipotezę, kad atestato vidurkis ir analizės egzamino pažymys vienodai stipriai susiję su respondento motyvacija. Tiksliau, tikrinsime hipotezę, kad $\beta_2 = \beta_3$. Kaip matyti iš 1.2.2 paveiksle pateiktų SAS programa gautų rezultatų, ši hipotezė atmetama ($p = 0{,}0025 < 0{,}01$).

Contrast Results

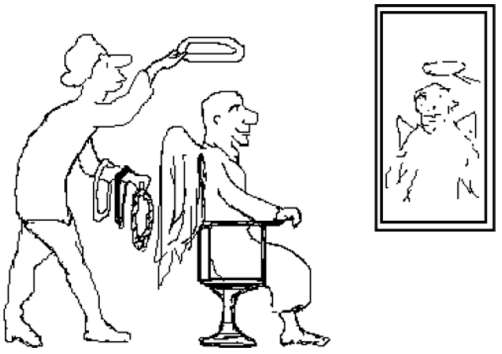
Contrast	DF	Chi-Square	Pr > ChiSq	Type
beta2=beta3	1	9.17	0.0025	Wald

1.2.2 pav. Hipotezė apie koeficientų lygybę

2.6. Modelių tinkamumas

Dauguma žmonių skiria daugiau laiko ir energijos, siekdami išvengti problemų nei jas spręsti.
H. Fordas

Vienas iš svarbiausių modelių analizės uždavinių yra nustatyti, ar pasirinktas modelis gerai tinka duomenims. LNM tinkamumas vertinamas atsižvelgiant į liekamąsias paklaidas, t. y. į skirtumus tarp kintamojo Y_j stebėjimų (turimų reikšmių) ir reikšmių, apskaičiuotų naudojantis tiriamu modeliu. GLM atveju toks tinkamumo matas ne visada natūralus. GLM užrašomas Y vidurkiui. Jeigu Y įgyja tik dvi reikšmes (0 ir 1), tai modelis užrašomas vieneto įgijimo tikimybei (šiuo paprastu atveju vidurkis ir tikimybė sutampa). Šią tikimybę galima naudoti prognozuojant, kokias reikšmes įgis Y, tačiau skirtumas tarp tikrosios ir prognozuojamos reikšmių visada bus lygus 0 arba 1. Norint apibendrintuosius tiesinius modelius palyginti ir įvertinti, ar jie tinka duomenims, naudojama *deviacija* arba normuotoji deviacija[18]. Pabandysime paaiškinti deviacijos idėją. Tarkime, kad turime sudėtingą, bet labai gerai duomenims tinkantį modelį. Siekiame skaitiškai įvertinti kiekvieno modelio skirtumą nuo šio bazinio modelio. Didelis skirtumas (deviacija) rodo, kad modelis nelabai tinka duomenims. Vienas iš GLM analizės tikslų – rasti paprastos struktūros modelį su maža deviacija, t. y. modelį, daug paprastesnį už pilnąjį, bet tinkantį duomenims beveik taip pat gerai.



Dabar apibrėšime deviaciją tiksliau. *Pilnuoju*[19] (baziniu, pertekliniu, prisotintuoju) vadinamas toks modelis, kai nežinomų parametrų yra tiek pat, kiek ir stebėjimų. Prisiminkime, kad parametrų įverčiai gaunami maksimizuojant didžiausiojo tikėtinumo funkcijos logaritmą. Pažymėkime pilnojo modelio didžiausiojo tikėtinumo funkcijos logaritmą (logtikėtinumo funkciją) $\ln L_f$. Tarkime, kad nagrinėjame modelį, kuriame yra K nepriklausomų kintamųjų $X_1, \ldots, X_K$, o modelio logtikėtinumo funkcija yra $\ln L$. Modelio normuotoji deviacija apibrėžiama taip:

$$\Delta = 2(\ln L_f - \ln L). \tag{1.2.10}$$

[18] Angl. *Scaled Deviance*.

[19] Angl. *Saturated*.

Nagrinėsime eksponentinei šeimai priklausančius Y_j, kurių $a_i(\phi) = \phi/p$. Primename, kad tada GLM parametrų įverčiai gaunami maksimizuojant logtikėtinumo funkciją

$$\ln L = \sum_{i=1}^{n} \left[\frac{p[y_i\theta_i - b(\theta_i)]}{\phi} + c(y_i, \phi) \right] \tag{1.2.11}$$

(žr. 2.4 skyrelį).

Modelio deviacija vadinamas dydis $\phi\Delta$. Nagrinėjant Puasono ar neigiamai binominę regresiją, $\phi = 1$. Todėl šiais atvejais normuotoji deviacija sutampa su deviacija.

Pastebėsime, kad, be pilnojo modelio, yra ir vadinamasis nulinis modelis. Tai toks modelis, kai Y nepriklauso nuo jokių regresorių X_j. Visi kiti apibendrintieji modeliai yra sudėtingesni už nulinį ir paprastesni už pilnąjį.

Taigi faktiškai deviacija palygina nagrinėjamą modelį su modeliu, kuris tiksliausiai aprašo duomenis, bet yra sudėtingiausios struktūros. Konkreti tiriamo modelio Δ reikšmė nėra labai informatyvi, nors ir žinoma, kad modelis tuo geriau tinka duomenims, kuo jo normuotoji deviacija mažesnė.

Jeigu eksponentinės šeimos tankio parametras ϕ žinomas, tai $\Delta \approx \chi^2_{n-K}$, t. y. apytiksliai turi chi kvadrato skirstinį su $n - K$ laisvės laipsnių. (Primename, kad n yra imties didumas, o K – modelio nepriklausomų parametrų skaičius.) Galima patikrinti statistinę hipotezę, ar normuotoji deviacija statistiškai reikšmingai skiriasi nuo nulio. Tarkime, pasirinktas reikšmingumo lygmuo yra $\alpha = 0{,}05$.

GLM tinkamumas duomenims yra pakankamas, jeigu jo normuotoji deviacija Δ statistiškai nereikšminga, t. y. p reikšmė $p \geqslant 0{,}05$.

Esant didelėms imtims, tokia statistiniu reikšmingumu grįsta taisyklė ne itin tinkama. Iš bendrosios hipotezių tikrinimo teorijos žinome, kad esant didelėms imtims net maži skirtumai gali būti statistiškai reikšmingi. Todėl šiuo atveju normuotąją deviaciją siūloma vertinti „iš akies“: modelis laikomas tinkamu duomenims, jeigu

$$\frac{\Delta}{n - K - 1} \approx 1.$$

Žinoma, nereikia šios taisyklės suabsoliutinti. Kur kas dažniau sakoma, kad modelis pakankamai gerai tinka duomenims, jeigu minėtas santykis nedaug skiriasi nuo vieneto. Pavyzdžiui, jeigu minėtas santykis lygus 1,23, tai galima manyti, kad modelis dar gana neblogai aprašo duomenis.

1.2.10 pavyzdys. Prisiminkime 1.2.7 pavyzdyje nagrinėtą neigiamai binominės regresijos modelį apie respondentų studijų motyvaciją. Naudodamiesi SAS programa, apskaičiuojame šio modelio deviacijos ir laisvės laipsnių santykį: $\Delta/(572 - 4) = 0{,}828$ (žr. 1.2.3 paveikslą). Tiems patiems duomenims taikydami Puasono regresijos modelį, gauname didesnę deviaciją: $\Delta/568 = 1{,}5$. Todėl darome išvadą, kad duomenims labiau tinka neigiamai binominės regresijos modelis. Beje, abijų modelių deviacijos statistiškai reikšmingai skiriasi nuo nulio. Kadangi respondentų buvo daug (per 500), tai šiuo atveju labiau reikėtų pasitikėti vertinimu „iš akies“.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	568	470.2996	0.8280
Scaled Deviance	568	470.2996	0.8280
Pearson Chi-Square	568	762.3760	1.3422
Scaled Pearson X2	568	762.3760	1.3422
Log Likelihood		-39.2533	
Full Log Likelihood		-1086.0932	
AIC (smaller is better)		2182.1863	
AICC (smaller is better)		2182.2923	
BIC (smaller is better)		2203.9320	

1.2.3 pav. SAS programos rezultatai

Normuotųjų deviacijų išraiškos, esant įvairiems apibendrintiems tiesiniams modeliams, pateiktos 1.2.2 lentelėje.

1.2.2 lentelė. Normuotosios deviacijos

Skirstinys	Normuotoji deviacija
Normalusis	$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2$
Puasono	$2\sum_{i=1}^{n}\{Y_i(\ln Y_i-\ln\hat{\mu}_i)-(Y_i-\hat{\mu}_i)\}$
Neigiamas binominis	$2\sum_{i=1}^{n}\left\{Y_i\ln\frac{Y_i}{\hat{\mu}_i}-(Y_i+1/\kappa)\ln\frac{Y_i+1/\kappa}{\hat{\mu}_i+1/\kappa}\right\}$

Deviaciją galima pritaikyti ir dviejų modelių palyginimui. Tarkime, turime dalinį GLM (t. y. dalis pradinio modelio $\boldsymbol{\beta}$ koordinačių laikomos lygiomis 0, tuo pačiu nepriklausomų kintamųjų skaičius $K' < K$). Statistika $\Delta(K, K') = 2(\ln L_K - \ln L_{K'})$ apytiksliai turi chi kvadrato skirstinį su $K - K'$ laisvės laipsnių; čia L_K ir $L_{K'}$ yra abiejų modelių tikėtinumo funkcijos. Sakoma, kad modeliai statistiškai reikšmingai skiriasi, jeigu $\Delta(K, K')$ statistiškai reikšmingai skiriasi nuo nulio.

Jei ϕ yra nežinomas eksponentinės šeimos tankio parametras, tai jo įvertinys užrašomas taip:

$$\hat{\phi} = \frac{\Delta}{n-K-1}.$$

Modelio tinkamumą duomenims rodo ir mažos liekanos (liekamosios paklaidos). GLM liekanos apibrėžiamos panašiai, kaip ir regresijos modelyje. Jei $\hat{\mu}_i$ yra vidurkio μ_i įvertinys, tai liekana vadinamas skirtumas $Y_i - \hat{\mu}_i$. Liekana, padalyta iš įvertinio $\hat{\mu}_i$ dispersijos reikšmės, vadinama normalizuotąja, arba Pirsono, liekana:

$$\frac{Y_i-\hat{\mu}_i}{\sqrt{\mathbf{D}\hat{\mu}_i}}.$$

Parametro ϕ įvertinio išraišką galima užrašyti naudojant Pirsono liekanas:

$$\hat{\phi} = \frac{1}{n-K-1}\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{\mathbf{D}\hat{\mu}_i}.$$

Esant kai kuriems skirstiniams, ϕ vertinti nereikia. Jau ne kartą minėjome, kad Puasono ir neigiamo binominio skirstinio $\phi = 1$.

2.7. Apibendrintųjų tiesinių modelių analizė naudojantis SPSS, SAS ir R programomis

Klysti žmogiška, bet, norint iš tikro viską supainioti, reikalingas kompiuteris.

P. Erlichas

Užrašysime, kaip atrodo neigiamai binominės regresijos modelio, nagrinėto 1.2.7 pavyzdyje, SPSS programa:

```
GENLIN NR BY LYTIS WITH ATVID MA
/MODEL LYTIS MA ATVID INTERCEPT=YES
DISTRIBUTION=NEGBIN(1) LINK=LOG
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Kaip ir tiriant LNM, programa užrašoma panašiai. Pradžioje nurodomas priklausomas kintamasis (*NR*), po to rašoma BY ir išvardijami kategoriniai aiškinamieji kintamieji (*LYTIS*). Tada rašoma WITH ir išvardijami intervaliniai aiškinamieji kintamieji (*ATVID, MA*). Sakinyje MODEL norodoma, ar modelyje bus laisvasis narys (INTERCEPT=YES), ir išvardijami visi kintamieji. Jeigu norėtume ištirti kintamųjų sąveikas, tai čia turėtume įtraukti atitinkamų kintamųjų sandaugas. Tolesnėje eilutėje nurodomas modelis.

R programoje apibendrintųjų modelių analizei naudojama funkcija *glm*. Štai kaip atrodo R programa, kai 1.2.7 pavyzdžio duomenims taikomas Puasono regresijos modelis:

```
GLMmod=glm(NR ATVID+MA+LYTIS,data=p,family="poisson")
summary(GLMmod)
library(MASS)
mnb=glm.nb(NR~ATVID+MA+LYTIS)
summary(mnb)
library(lmtest)
waldtest(GLMmod, test="Chisq")
library(sandwich)
coeftest(GLMmod, vcov=sandwich)
residuals(GLMmod, type="pearson") # Pirsono liekanų reikšmės
plot(GLMmod$fitted.values,GLMmod$residuals) # prognozuojamų reikšmių – liekanų grafikas
predict(GLMmod,type="response") # prognozuojamos reikšmės
```

```
Call:
glm(formula = NUMERIS ~ ATESTVID + IVERTINIM + LYTIS, family = "poisson",
   data = p)

Deviance Residuals:
   Min      1Q   Median      3Q     Max
-1.8882  -0.9026  -0.5132   0.1392   6.8417

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.57520    0.37762   9.468  < 2e-16 ***
ATESTVID    -0.26139    0.04248  -6.153 7.63e-10 ***
IVERTINIM   -0.05887    0.01745  -3.373 0.000744 ***
LYTIS        0.47298    0.05830   8.113 4.92e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 967.28  on 571  degrees of freedom
Residual deviance: 860.69  on 568  degrees of freedom
AIC: 2313.8

Number of Fisher Scoring iterations: 5
```

1.2.4 pav. R programos rezultatai taikant Puasono regresijos modelį

Šios programos rezultatai pateikiami 1.2.4 paveiksle.

Parametrui *formula* galioja tos pačios reikšmės užrašymo taisyklės, kaip ir taikant funkciją *lm*; *family* – modelio jungties funkcijos ir priklausomo kintamojo skirstinio parametras. Pateikiame tik keletą dažniau naudojamų parametro *family* reikšmių (atkreipiame dėmesį, kad vienam ir tam pačiam priklausomo kintamojo skirstiniui galima parinkti kelias skirtingas jungties funkcijas).

binomial(link = "logit" "probit" "log" "cloglog"')
gaussian(link ="identity")
Gamma(link = "inverse" "identity" "log")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log" "identity" "sqrt")

Sakinys *summary(fit)* išspausdina parametrų įverčius ir deviacijos reikšmes. Išsamesnė informacija apie modelį pateikiama funkcijos *glm* aprašyme R dokumentacijoje. Susipažinsime tik su keletu naudingų sakinių:

residuals(GLMmod, type="pearson") – išveda modelio kiekvieno stebėjimo Pirsono liekanų reikšmes;

plot(fit$fitted.values,fit$residuals) – išveda prognuozuojamų reikšmių-liekanų grafiką;

predict(GLMmod,type="response") – išveda prognozuojamas reikšmes (prognozuojami vidurkiai μ_i).

SAS programoje tiesinių modelių analizei skirtos kelios procedūros: PROC GLM, PROC GENMOD, PROC MIXED. Procedūrą PROC GLM aprašėme anksčiau, PROC MIXED aprašyta antroje šios knygos dalyje. Todėl apsiribosime procedūra PROC GENMOD.

Pastebėsime, kad CLASS, CONTRAST, ESTIMATE, LSMEANS sakinių paskirtis yra ta pati, kaip ir procedūros PROC GLM. Pateiksime sakinio MODEL pavyzdžių, nes apibendrintųjų modelių užrašymo forma skiriasi nuo LNM. Bendroji sakinio MODEL išraiška yra tokia:

```
MODEL <response(s)> = <variable list> / DIST = <dist>
LINK = <link><options>;
```

Prisiminus logistinės regresijos 1.2.3 pavyzdį apie skrandžio vėžio atsinaujinimą, *MODEL* išraiška yra tokia:

```
MODEL ATSINAUJINO(EVENT=LAST) = AMZIUS AUGLYS LYTIS NR
/ DIST=BINARY LINK=LOGIT;
```

Pastebėsime, kad *LINK*, *DIST* suteikia galimybę nurodyti jungties funkciją ir pradinių duomenų struktūrą. Pateiktas sakinio MODEL užrašas tiktų tokiems duomenims:

ATSINAUJINO	AMŽIUS	AUGLYS	LYTIS
1	40	7	1
0	50	6,8	2
.	.	.	.
1	60	7,2	1

Kartais nurodomi priklausomo kintamojo dažnių duomenys. Pavyzdžiui, žinome, kiek buvo atitinkamos lyties, tam tikro amžiaus pacientų, turinčių vienodo didumo auglį, ir kiek iš jų susirgo:

ATSINAUJINO	IŠ VISO	AMŽIUS	AUGLYS	LYTIS
2	8	40	7	1
1	4	50	6,8	2
.	.	.	.	.
3	7	60	7,2	1

Iš 8 keturiasdešimtmečių vyrų, kurių auglys buvo 7 mm skersmens, susirgo 2; iš 4 penkiasdešimtmečių moterų, kurių auglys buvo 6,8 mm skersmens, susirgo 1 ir t. t. Tokiems duomenims *MODEL* užrašomas taip:

```
MODEL ATSINAUJINO/VISO = AMZIUS AUGLYS LYTIS / DIST = BIN
LINK = LOGIT;
```

Pateikiame 1.2.10 pavyzdžio analizės, taikant neigiamai binominės regresijos modelį, SAS programą.

```
PROC GENMOD DATA = POISSONREG; CLASS LYTIS;
MODEL NR = ATVID MA LYTIS /DIST=NEGBIN;
RUN;
```

Kontrastai koeficientų lygybei tikrinti aprašomi panašiai, kaip ir LNM atveju. Prisiminkime 1.2.9 pavyzdį. Tikrinant hipotezę, kad $\beta_2 = \beta_3$, SAS programa užrašoma taip:

```
PROC GENMOD DATA = POISSONREG; CLASS LYTIS;
MODEL NR = ATVID MA LYTIS /DIST=NEGBIN;
CONTRAST "beta1=beta2" ATESTVID 1 IVERTINIM -1 /Wald;
RUN;
```

deviacija	jungtis	pilnasis modelis

UŽDAVINIAI

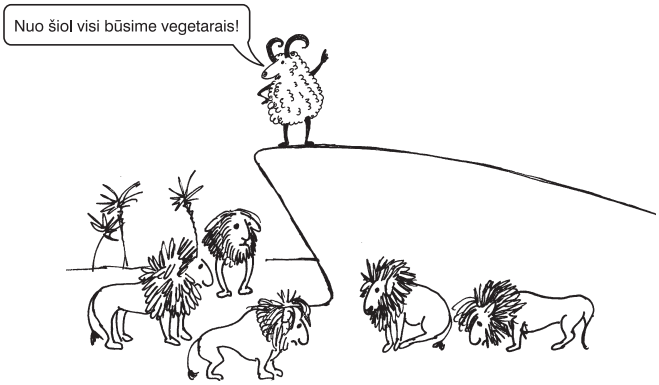
1. Įrodykite, kad eksponentinių skirstinių šeimai priklauso:
   a) Pareto skirstinys $f(y;\theta) = \theta y^{-\theta-1}$;
   b) eksponentinis skirstinys $f(y;\theta) = \theta \exp\{-y\theta\}$.
2. Įsitikinkite, kad binominis skirstinys priklauso eksponentinių skirstinių šeimai.
3. Įrodykite, kad binominio skirstinio atveju deviaciją galima užrašyti taip:

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{p}_i}\right)\right];$$

   čia $Y_i \sim Bi(n_i, p_i)$, $\hat{\mu}_i := n_i\hat{p}_i$, o $\hat{p}_i$ yra p_i didžiausiojo tikėtinumo įvertis.
4. Prisiminkite 1 skyriaus 6 uždavinį apie svogūnų derlingumą. Užrašykite Puasono regresijos modelį ir apskaičiuokite vidurkių įverčius.
5. Faile **mif2003ma.sav** pateikiami duomenys, aprašyti 1.2.7 pavyzdyje. Sudarykite logistinės regresijos modelį, padedantį įvertinti bakalauro studijų baigimo tikimybę.

2

HIERARCHINIAI MODELIAI



Geriau avinai, vadovaujami liūto, nei liūtai, vadovaujami avino.

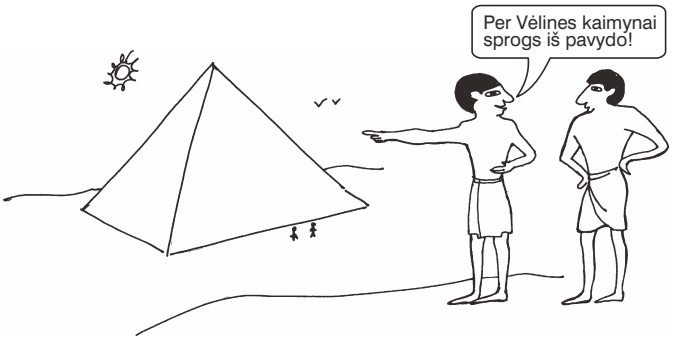
Arabų patarlė

1. TIESINIAI HIERARCHINIAI MODELIAI

Hierarchinėje sistemoje kiekvienas kyla karjeros laiptais tol, kol pagaliau nebesusitvarko su savo pareigomis.

Piterio principas

Kaip teigia legendos, Vytautas sugebėjo savo arklius pagirdyti Juodosios jūros vandeniu[1], Cheopso vadovaujami egiptiečiai, statydami antkapį, pasiekė Gineso rekordą, o štai Mozė savo bendražygiams parinko tokį maršrutą, kad jie nuo Egipto iki Judėjos ėjo keturiasdešimt metų. Visi minėtieji pavyzdžiai iliustruoja lengvai suvokiamą tiesą, kad rezultatai priklauso ne tik nuo pačių pavaldinių (arklių, darbuotojų, bendražygių), bet ir nuo jų vadovo. Kasdienybėje apstu analogiškų pavyzdžių. Kai jūsų vaikas per egzaminą gauna prastą pažymį, kas *labiausiai* kaltas? Teisingai, dėstytojas. (Tai kaimyno vaikas pats žioplas, kai neišlaiko.)



Aišku, kad kurdami pavaldinio pasiekimų statistinį modelį turėtume atsižvelgti ir į vadovo indėlį. Deja, paprastas regresinis modelis čia netinka. Kodėl – aptarsime vėliau. Dabar paminėsime tik pagrindinę priežastį – regresijoje reikalaujama, kad skirtingų respondentų rezultatai būtų nepriklausomi. Tuo tarpu iš to paties mokytojo mokinių rezultatų matyti „mokytojo ranka“. Todėl stebėjimai grupėse panašesni, nei turėtų būti nagrinėjant negrupuotus duomenis. Iš pačių duomenų aiškiai matyti dviejų lygmenų hierarchinė struktūra: pirmasis lygmuo – mokiniai, antrasis lygmuo – klasės. Panašios situacijos kyla, kai:

- medikas tiria kvėpavimo takų susirgimų dažnio priklausomybę nuo ligonio savybių (I lygmuo) ir individo gyvenamosios vietos ypatybių (II lygmuo);
- kriminologas tiria jaunimo polinkio į nusikalstamumą priklausomybę nuo respondento šeimos socialinių ir ekonominių parametrų (I lygmuo), respondento ir jo draugų gyvenamojo rajono kriminogeniškumo (II lygmuo) ir nuo miesto didumo (III lygmuo);
- sveikatos priežiūros įstaigas kontroliuojanti organizacija aiškinasi reabilitacijos trukmės po širdies operacijų priklausomybę nuo ligonių (I lygmuo), gydytojų (II lygmuo) ir ligoninių (III lygmuo) ypatumų.

[1] Priešingai patarlei, jog yra daug būdų atvesti arklį prie vandens, bet nėra būdo priversti jį gerti.

Visais minėtais atvejais tyrimams tikslinga taikyti tiesinį hierarchinį modeliavimą (toliau – HLM[2]). HLM, kuris kartais dar vadinamas atsitiktinių koeficientų (RC[3]) modeliavimu, yra hierarchinės regresinės analizės forma, išplėtota dvidešimtojo amžiaus devintajame dešimtmetyje. Šiuo metu HLM labai plačiai taikomas daugelyje sričių, ypač edukologijoje, psichologijoje, biologijoje, medicinoje ir ekonomikoje. HLM populiarumą lemia ir tai, kad sukurta daug programinių priemonių jam atlikti.

1.1. Hierarchiniai duomenys

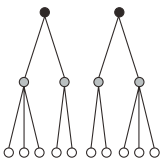
Pristatydami HLM užsiminėme, kad jis taikomas hierarchiniams duomenims. Kokie tai duomenys? Tiriamos aibės pradiniai elementai sudaro grupes. Tos grupės laikomos aukštesnio lygmens elementais ir savo ruožtu sudaro grupes, kurios yra dar aukštesnio lygmens elementai ir t. t. Jei tiriami duomenys yra tokios struktūros, tai jie vadinami hierarchiniais. Apskritai, kažin ar pasaulyje yra nehierarchinių duomenų, tik ne visada į tai svarbu atsižvelgti. Jei atsitiktinai parenkami vilniečiai paaugliai ir tiriami tik jų rūkymo įpročiai, tai duomenys dar nėra hierarchiniai. Tuo tarpu, jei kreipiamas dėmesys į paauglių gyvenamųjų rajonų ypatybes, – jau turime hierarchinius duomenis. Šiame skyriuje terminas „elementas“ yra vartojamas vietoje statistikos termino „objektas“. Taip yra todėl, kad „kapstomės po sistemotyros daržą“ ir vartojame sistemotyrai įprastesnę terminiją. Statistikoje būtų įprasčiau kalbėti apie populiaciją, jos objektus ir pan.

Hierarchiniai duomenys yra neatsiejami nuo daugialygio atrankos būdo. Dažniausiai jie gaunami, kai parenkamos sluoksninės arba lizdinės imtys. Pavyzdžiui, atsitiktinai parenkame kokio nors rajono mokyklas, tada atsitiktinai parenkame šių mokyklų mokinius.

Klasikinis hierarchinės struktūros pavyzdys: mokinys → klasė → mokykla. Tokį žymėjimą reikėtų suprasti taip: mokinys (pirmojo lygmens elementas) yra klasės (antrojo lygmens elemento) sudėtinė dalis, o ši savo ruožtu yra mokyklos (trečiojo lygmens elemento) sudėtinė dalis (žr. 2.1.1 pav.). Lygmenų gali būti ir keli, ir keliolika. Ką laikyti pirmojo lygmens elementais, sprendžia tyrėjas. Pavyzdžiui, jis gali tirti tik klases (pirmasis lygmuo) ir mokyklas (antrasis lygmuo). Tipiniai dviejų lygmenų elementų pavyzdžiai: mokiniai → klasės, rajonai → apskritys, tarnautojai → įstaigos, pacientai → ligoninės, vaikai → šeimos.

Skirtingų lygmenų elementų charakteristikos priklausomos. Juk žinoma, kad net ir vidutinių gabumų mokinio, atėjusio į pažangių mokinių klasę, rezultatai pagerėja, t. y. aplinka (klasė) irgi veikia individą.

Lygmuo	Lygmens elementai	Lygmens kintamieji
3	Mokyklos	Rajono išsivystymas
2	Klasės	Mokytojo patirtis
1	Mokiniai	Tėvų išsilavinimas



2.1.1 pav. Hierarchinės duomenų struktūros pavyzdys

[2] HLM – angl. *Hierarchical Linear Modeling*.

[3] RC – angl. *Random Coefficients*.

Pirmasis lygmuo yra žemiausias.
Elementai apibrėžiami kiekviename lygmenyje.

Išsiaiškinus hierarchinę duomenų struktūrą, lengviau sudaryti teisingą statistinį modelį. Dydžiai, apibūdinantys tiriamųjų elementų matuojamas savybes, toliau vadinami kintamaisiais. Pavyzdžiui, tarkime, kad nagrinėjame pradinių klasių mokinių raštingumą. Tėvų išsilavinimas – pirmojo lygmens kintamasis; mokytojo patirtis – antrojo, o rajoną, kuriame yra mokykla, apibūdinantys kintamieji – trečiojo (žr. 2.1.1 pav.). Norėtume atkreipti dėmesį, kad turint hierarchinius duomenis galima nagrinėti įvairių lygmenų kintamuosius.

1.2. Kodėl taikomas HLM

Pasaulyje gyvena dviejų rūšių žmonės: tie, kurie nesupranta matematikos, ir tie, kurie jais rūpinasi.

Hierarchinių duomenų analizė atliekama jau seniai, tik tyrimui nebuvo naudojamas hierarchinis tiesinis modeliavimas. Anksčiau hierarchiniai duomenys būdavo *agreguojami* (jungiami) arba *deagreguojami* (išskaidomi), o po to taikomas tradicinis tiesinės regresijos modelis.

Duomenų agregavimas reiškia žemesnio lygmens kintamojo reikšmių suvidurkinimą ir priskyrimą aukštesnio lygmens elementui. Tarkime, kad tiriama įvairių klasių mokinių matematikos testo rezultatų priklausomybė nuo jų socialinio ir ekonominio statuso. Klasės rezultatas – joje besimokančių mokinių testo rezultatų vidurkis. Analogiškai galima suvidurkinti ir klasei priskirti kitus mokinių rodiklius. Po to jau standartiniais statistiniais metodais tiriamos klasės. Toks tyrimas nėra nekorektiškas, tačiau:

1 sujungus duomenis, gaunama kur kas mažiau tiriamų elementų; pavyzdžiui, klasių yra daug mažiau nei mokinių;

2 elementai, atstovaujantys tik nedaugeliui stebėjimų, gali daryti neproporcingai didelę įtaką; pavyzdžiui, trisdešimties mokinių pažymių vidurkis laikomas lygiareikšmiu šešių mokinių pažymių vidurkiui;

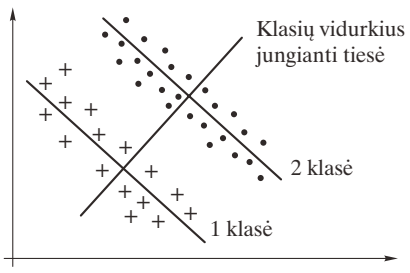
3 prarandama galimybė įvertinti įvairių lygmenų elementų tarpusavio įtaką; pavyzdžiui, nepavyks nustatyti klasės ir mokinių savybių priklausomybės;

4 prarandama (vidinė) grupinė informacija, dėl to gali būti stipriai iškraipyti kintamųjų sąryšiai; pavyzdžiui, tiriant mokinius ir klases ir vertinant dviejų kintamųjų koreliaciją, galima gauti labai skirtingus rezultatus.

Deagregavimas (išskaidymas) – tai atvirkštinė agregavimui procedūra, kai kiekvienam žemesniojo lygmens elementui priskiriama *ta pati* aukštesniojo lygmens kintamojo reikšmė. Pavyzdžiui, klasės didumas traktuojamas kaip mokinį apibūdinantis kintamasis. Visi klasikiniai statistiniai kriterijai remiasi prielaida, kad imtis sudaro nepriklausomi stebėjimai, t. y. kiekvienas imties stebėjimas suteikia naujos informacijos. Tačiau šiuo atveju informacija kartojasi. Ignoruojant šį faktą, gaunamos mažesnės kintamųjų charakteristikų standartinės paklaidos, nei turėtų būti. Dėl to taikant statistinius kriterijus galimos per daug liberalios išvados, t. y. nulinė hipotezė priimama dažniau nei derėtų, o pasikliautinieji intervalai yra nepagrįstai trumpi.

Agregavimo ir deagregavimo tendencingumas:
tai, kas būdinga grupėms, yra būdinga ir individams.

Pateiksime dar vieną pavyzdį, iliustruojantį, kad į duomenų struktūrą būtina atsižvelgti. Tarkime, kad tirta dviejų klasių mokinių matematinio testo klaidų skaičiaus priklausomybė nuo testo trukmės. Jeigu kiekvienos klasės duomenys tiriami atskirai, tai gaunama natūrali priklausomybė – kuo ilgiau daromas testas, tuo mažiau klaidų (atitinkamas tendencijas iliustruojančias dvi tieses žr. 2.1.2 pav.). Kita vertus, jeigu imami klasių vidurkiai, tai juos jungianti tiesė rodo visiškai priešingą (ir, beje, prieštaraujančią sveikam protui) priklausomybę.

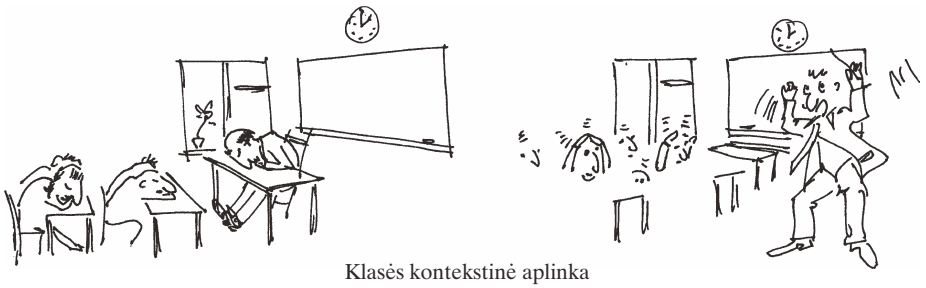


2.1.2 pav. Klaidų skaičiaus priklausomybė nuo testo trukmės

Tad kyla natūralus klausimas – kodėl nenagrinėti kiekvienos klasės atskirai? Iš tikrųjų tai būtų visiškai geras sprendimas, bet tik tuo atveju, jei: a) mokinių klasėse būtų daug, b) mūsų nedomintų klasės įtaka mokiniams. Žinome, kad statistinių išvadų patikimumas labai priklauso nuo tiriamųjų respondentų skaičiaus.

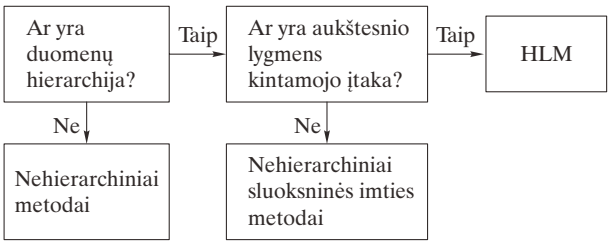
Tarkime, kad tiriamos šešios klasės po 30 mokinių. Tirinat kiekvieną klasę atskirai, tyrimo išvados grindžiamos 30 mokinių duomenimis, o taikant HLM modeliavimą, atsižvelgiama į daugiau mokinių. Aišku, kad antruoju atveju išvados yra daug patikimesnės. Žinoma, jei grupėse bus tik po kelis elementus, dalinių charakteristikų skaičiavimui nepadės ir HLM. Dar svarbesnė HLM savybė – galimybė nustatyti vadinamąją *kontekstinės aplinkos* įtaką, t. y. atsižvelgti į antrojo (trečiojo ir t. t.) lygmens kintamųjų įtaką. Tiriant

mokinius, kartu nustatoma, *kokios* klasės charakteristikos svarbios mokinių pasiektiems rezultatams ir *kaip* jos tuos rezultatus veikia.



Klasės kontekstinė aplinka

Skyrelio pabaigoje pateikiame schemą (žr. 2.1.3 pav.), kuri turėtų padėti skaitytojui orientuotis, kada verta taikyti HLM.



2.1.3 pav. HLM taikymo schema

1.3. Besąlyginis HLM modelis

Įstaigos aritmetika:
Protingas viršininkas + protingas pavaldinys = pelnas.
Protingas viršininkas + kvailas pavaldinys = nuostoliai.
Kvailas viršininkas + protingas pavaldinys = paaukštinimas.
Kvailas viršininkas + kvailas pavaldinys = viršvalandžiai.

Pirmiausia panagrinėsime hierarchinį tiesinį modelį, kuris vadinamas *besąlyginiu*, arba nuliniu, modeliu. Kuo jis svarbus? Besąlyginio modelio analizė padeda nuspręsti, ar apskritai duomenims taikytinas hierarchinis modeliavimas. Be to, kiti hierarchiniai modeliai su juo lyginami tikrinant, ar juos praplėtus jie geriau tinka duomenims.

Modelio formuluotė. Besąlyginį HLM modelį aprašysime remdamiesi pavyzdžiu, kuris mums pasitarnaus ir vėliau nagrinėjant HLM modelių sudarymo, tyrimo ir interpretavimo problemas. Naudosimės TIMSS[4] tyrimo Lietuvoje duomenimis. Vaizdesnei HLM iliustracijai pasirinktas duomenų poaibis – kaimo ir Vilniaus miesto mokyklų VIII klasės mokinių tyrimo rezultatai. Turime 559 mokinių iš 27 mokyklų matematikos testo rezultatus ir įvairius mokyklą bei mokinį apibūdinančius duomenis.

[4] TIMSS – angl. *Trends in International Mathematics and Science Study 2003*, žr. http://timss.bc.edu/.

Besąlyginį HLM modelį sudaro abiejų – mokinio ir mokyklos – lygmenų lygtys. Aprašydami mokinio pasiekimus, vadovausimės tokia logika: mokinio rezultatų skirtumą nuo visos mokyklos vidurkio skaičiuosime atsižvelgdami į tai, kas būdinga tik tam mokiniui – jo gabumai, namų aplinka, repetitoriaus patirtis ir pan.

Mokinio rezultatas = mokyklos vidurkis + individualūs skirtumai nuo vidurkio.

Analogiškai aprašome mokyklos lygmenį:

Mokyklos vidurkis = visų mokyklų vidurkis + šios mokyklos skirtumas nuo kitų mokyklų.

Gauta dviejų priklausomybių schema iliustruoja besąlyginio modelio idėją – visa, kas gali turėti įtakos mokinio ar mokyklos pasiekimams, pateikiama kaip atsitiktinės paklaidos. Aprašysime besąlyginį HLM modelį matematiškai. Simboliu Y_{ij} pažymėkime i-ojo mokinio iš j-osios mokyklos matematikos testo rezultatus. Tarkime, kad mokyklų yra J, t. y. $j = 1, 2, \ldots, J$, o mokinių j-oje mokykloje yra n_j, t. y. $i = 1, 2, \ldots, n_j$.

Pirmasis (mokinio) lygmuo:

$$Y_{ij} = \beta_{0j} + e_{ij}, \quad i = 1, \ldots, n_j, \; j = 1, \ldots, J. \tag{2.1.1}$$

Antrasis (mokyklos) lygmuo:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad j = 1, \ldots, J. \tag{2.1.2}$$

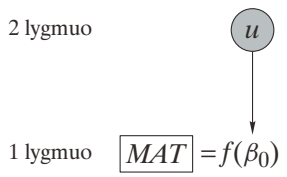
Modelio struktūra: pirmojo lygmens lygtį sudaro *atsitiktinis* koeficientas β_{0j} ir atsitiktinė paklaida e_{ij}. Antrojo lygmens lygtį sudaro *konstanta* γ_{00} ir atsitiktinė paklaida u_{0j}. Kaip ir klasikiniuose regresiniuose modeliuose, tariama, kad visos atsitiktinės paklaidos e_{ij} nepriklausomos, normaliai pasiskirsčiusios ir turi tas pačias dispersijas σ^2. Taip pat tariama, kad mokyklos indėlį rodanti paklaida u_{0j} turi dispersiją τ_{00} ir nekoreliuoja su e_{ij}.

Modelio prielaidos:

1) $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ nepriklausomi visiems $i = 1, \ldots, n_j$, $j = 1, \ldots, J$;
2) $u_{0j} \sim \mathcal{N}(0, \tau_{00})$ nepriklausomi visiems $j = 1, \ldots, J$;
3) e_{ij} ir u_{0j} nekoreliuoja.

Paklaidos yra atsitiktiniai dydžiai, todėl norint apibūdinti jų didumą vertinamos dispersijos σ^2 ir τ_{00}. Didelės dispersijų reikšmės rodo, kad atsitiktiniai dydžiai įgyja labai skirtingas reikšmes. Taigi, norėdami įvertinti besąlyginį modelį, vertiname tris parametrus (konstantas), t. y. γ_{00}, σ^2 ir τ_{00}.

γ_{00} vadinamas fiksuoto poveikio parametru,
τ_{00}, σ^2 vadinami atsitiktinio poveikio parametrais.



2.1.4 pav. Besąlyginio modelio idėja

Skyrelio pradžioje pateikto pavyzdžio parametrai yra tokie:

- β_{0j} yra j-osios mokyklos rezultatų vidurkis;
- e_{ij} – i-ojo mokinio rezultatų skirtumai nuo j-osios mokyklos vidurkio;
- σ^2 – dispersija; kuo jos reikšmė didesnė, tuo mokinių rezultatai labiau skiriasi;
- γ_{00} – bendrasis visų mokyklų mokinių rezultatų vidurkis;
- u_{0j} – j-osios mokyklos atsitiktinė paklaida, kuri apima visa, dėl ko j-osios mokyklos rezultatai geresni (blogesni) negu kitų mokyklų;
- τ_{00} – dispersija; kuo jos reikšmė didesnė, tuo mokyklų rezultatai labiau skiriasi.

Grafiškai besąlyginio modelio idėja pavaizduota 2.1.4 paveiksle. Visuose šio skyriaus paveiksluose užrašas $f(\ldots)$ tereiškia „tiesiškai priklauso nuo skliaustuose nurodytų dydžių“. Taip daroma sąmoningai, kad lengviau būtų galima suvokti esmines priklausomybes.

Besąlyginį modelį galima aprašyti ir viena jungtine lygtimi, įstačius (2.1.1) išraišką į (2.1.2) formulę:

$$Y_{ij} = \gamma_{00} + [u_{0j} + e_{ij}]. \tag{2.1.3}$$

Atkreipiame dėmesį, kad jungtinio besąlyginio (2.1.3) modelio atsitiktinė paklaida turi du dėmenis: u_{0j} – mokyklos indėlį (bendrą visiems tos mokyklos mokiniams) ir individualią paklaidą e_{ij}. Kurį modelio pavidalą rinktis – dviejų lygčių ar vienos? Tai priklauso nuo modelio analizei naudojamos programinės įrangos. SPSS, SAS ir R programų sintaksė tokia, kad patogiau atsižvelgti į jungtinę lygtį, HLM programoje – į abiejų lygmenų lygtis.

Mūsų pavyzdyje besąlyginis modelis padės nuspręsti, ar mokyklų rezultatai skiriasi. Nesunku pastebėti, kad suformuluotoji problema – tipiškas dispersinės analizės (prisiminkime trumpinį ANOVA) uždavinys. Iš tikrųjų pateiktasis modelis ekvivalentus atsitiktinių poveikių dispersinės analizės modeliui, tik kitokie vartojami žymenys.

Parametrų įverčiai ir statistinės hipotezės. Atliekant statistinį tyrimą, galima rasti nežinomų parametrų įverčius ir patikrinti statistines hipotezes apie parametrų reikšmes. Besąlyginio modelio analizė padės:

- gauti γ_{00}, σ^2 ir τ_{00} įverčius;
- patikrinti statistines hipotezes:

$$A\colon \begin{cases} H_0\colon\ \gamma_{00} = 0, \\ H_1\colon\ \gamma_{00} \neq 0, \end{cases} \qquad B\colon \begin{cases} H_0\colon\ \sigma^2 = 0, \\ H_1\colon\ \sigma^2 > 0, \end{cases} \qquad C\colon \begin{cases} H_0\colon\ \tau_{00} = 0, \\ H_1\colon\ \tau_{00} > 0. \end{cases}$$

Prisiminus nagrinėjamą pavyzdį, galima konstatuoti, kad faktiškai sprendžiamos šios problemos:

- Bandoma skaitiškai įvertinti visų mokinių rezultatų vidurkį, skirtumus tarp mokinių ir tarp mokyklų.
- Tikrinamos trys hipotezės: A (ar bendrasis mokinių testo rezultatas nėra lygus nuliui), B (ar tarp mokinių rezultatų yra ne mokyklų nulemtų skirtumų), C (ar mokyklų įtaka rezultatams skiriasi).

Kaip apskaičiuoti parametrų įverčius ir patikrinti hipotezes naudojantis SPSS, SAS ir R programomis, smulkiau aprašyta kituose skyreliuose. Dabar tik pateikiame SPSS programa gautus rezultatus:

1 Vidutinio visų mokinių matematikos testo γ_{00} įvertis $\hat{\gamma}_{00} = 499{,}47$ ir jo 95 % pasikliautinasis intervalas yra [479,15; 519,79].

2 Hipotezės A (apie γ_{00} lygybę nuliui) p reikšmė $p < 0{,}05$. Tarkime, kad pasirinktas reikšmingumo lygmuo $\alpha = 0{,}05$. Tada nulinė hipotezė atmetama ir daroma išvada, kad $\gamma_{00} \neq 0$. Tai reikštų, kad bendrasis visų mokinių testo rezultatų vidurkis nėra lygus nuliui. Ši išvada nėra itin vertinga (kažin ar buvo galima tikėtis, kad visi mokiniai – absoliutūs nemokšos), daug svarbesnis to vidurkio skaitinis įvertis.

3 Mokinių (pirmojo lygmens elementų) rezultatų skirtumus apibūdinančios dispersijos σ^2 įvertis $\hat{\sigma}^2 = 5117{,}19$ ir jo 95 % pasikliautinasis intervalas yra [4537,53; 5770,89].

4 Mokyklų (antrojo lygmens elementų) skirtumus apibūdinančios dispersijos τ_{00} įvertis $\hat{\tau}_{00} = 2356{,}11$ ir jo 95 % pasikliautinasis intervalas yra [1276,40; 4349,12].

5 Hipotezės B (H_0: $\sigma^2 = 0$) p reikšmė $p < 0{,}05$. Nulinė hipotezė atmetama ir daroma išvada, kad $\sigma^2 > 0$. Tai interpretuojama kaip įrodymas, jog yra statistiškai reikšmingų mokinių rezultatų skirtumų, priklausančių nuo mokinių individualių savybių, bet ne nuo mokyklos. Didelis σ^2 įvertis šiuo atveju informatyvesnis nei hipotezės išvada. Jis rodo, kad mokinių rezultatai labai skiriasi. Be to, vertėtų paieškoti pirmojo lygmens kintamųjų, galinčių paaiškinti šiuos skirtumus.

6 Hipotezės C (H_0: $\tau_{00} = 0$) p reikšmė $p < 0{,}05$. Nusprendžiame, kad statistiškai reikšmingai $\tau_{00} > 0$. Tai leidžia padaryti vieną iš svarbiausių išvadų – mokyklų rezultatai statistiškai reikšmingai skiriasi. Didelė dispersijos įverčio reikšmė taip pat rodo, kad reikėtų paieškoti antrojo lygmens kintamųjų, galinčių paaiškinti vidutinių mokyklų rezultatų skirtumus.

Taigi matome, kad besąlyginio modelio įverčiai dažniausiai informatyvesni nei atmestos arba neatmestos statistinės hipotezės.

Kitos skaitinės modelio charakteristikos. Kaip ir dispersinėje analizėje, galima apskaičiuoti tarpklasinės koreliacijos koeficientą *ICC*[5]. Jis parodo, kaip stipriai skiriasi grupių (mokyklų) rezultatai, palyginti su rezultatais grupėse. Tikslesnis *ICC* apibrėžimas,

[5] *ICC* – angl. *Intraclass Correlation Coefficient*.

vartojant šio skyrelio žymenis, yra toks:

$$ICC = \frac{\tau_{00}}{\sigma^2 + \tau_{00}}.$$

Kaip naudojamas *ICC* hierarchinių modelių analizėje? Jeigu būtų $ICC = 0$ (tada $\tau_{00} = 0$), tai gautume, kad nėra grupių įtakos ir nereikia jokių hierarchinių modelių, užtenka paprasčiausios regresijos. Taigi kuo *ICC* didesnis, tuo labiau galima manyti, kad reikia atsižvelgti į hierarchinę struktūrą. Nėra griežtos taisyklės, ką laikyti dideliu skaičiumi. Čia jau teks pasikliauti sveiku protu. Mūsų nagrinėtame pavyzdyje gaunamas toks *ICC* įvertis:

$$ICC = \frac{2356{,}11}{2356{,}11 + 5117{,}19} = 0{,}315.$$

Tokią *ICC* reikšmę galima interpretuoti kaip indikatorių, kad per 31 % mokinių rezultatų skirtumų lemia mokyklos.

Yra ir kita koeficiento *ICC* interpretacija. Nesunku įrodyti, kad *ICC* atspindi dviejų tos pačios mokyklos mokinių rezultatų koreliaciją, kai $i \neq i'$:

$$ICC = \mathrm{Cor}(Y_{i'j}Y_{ij}) = \frac{\mathbb{C}\mathrm{ov}(Y_{i'j}Y_{ij})}{\sqrt{\mathbf{D}(Y_{ij})\mathbf{D}(Y_{i'j})}} = \frac{\tau_{00}}{\sqrt{\tau_{00} + \sigma^2}\sqrt{\tau_{00} + \sigma^2}} = \frac{\tau_{00}}{\sigma^2 + \tau_{00}}.$$

Taigi *ICC* galima interpretuoti kaip stebėjimų grupėse priklausomumo matą. Bet kuriuo atveju didelė *ICC* reikšmė – tai rodiklis, kad verta tirti HLM modelius.

ICC yra apibendrinantis visų mokyklų įtakos įvertis. Norint įvertinti konkrečios (tarkime, j-osios) mokyklos įtaką tos mokyklos mokinių rezultatams, reikia įvertinti β_{0j}. Primename, kad β_{0j} reikšmė interpretuojama kaip j-osios mokyklos matematikos testo rezultatų vidurkis. Galima tiesiog ignoruoti visas likusias mokyklas ir mąstyti taip: j-ojoje mokykloje rezultatai aprašomi (2.1.1) lygtimi, todėl imame visų tos mokyklos rezultatų vidurkį

$$\overline{Y}_{\cdot j} = \frac{Y_{1j} + Y_{2j} + \cdots + Y_{n_j,j}}{n_j}.$$

Toks įvertinys yra nepaslinktasis, jo dispersija lygi σ^2/n_j. Ši dispersija gali būti labai didelė mažiems n_j (pavyzdžiui, n_j – mažytės kaimo mokyklėlės mokinių skaičius). Primename, kad didelė įvertinio dispersija – blogas rezultatas, nes praktiškai gauti parametro įverčiai dėl imties atsitiktinumo gali būti labai nepanašūs į tikrąją parametro reikšmę.

Kita galimybė – parametro β_{0j} įvertinį pasirinkti $\hat{\gamma}_{00}$, t. y. pasinaudoti visų mokyklų rezultatais. Aišku, kad toks įvertinys taip pat nėra geras, nes jis visų mokyklų vienas ir tas pats. Be to, tai yra paslinktasis β_{0j} įvertinys, nors poslinkio paklaida dėl didelio mokinių skaičiaus pakankamai maža.

Išeitis – susieti abu įvertinius. Naudojamas vadinamasis svertinis (sutraukiantysis[6]) koeficiento β_{0j} įvertinys, kuris yra dviejų minėtų įvertinių tiesinis darinys.

Svertinis įvertinys $\beta^*_{0j} = \lambda_j \overline{Y}_{\cdot j} + (1 - \lambda_j)\hat{\gamma}_{00}$; čia $\lambda_j = \dfrac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2/n_j}$.

[6] Angl. *shrinkage*.

Parametras λ_j vadinamas įvertinio $\overline{Y}_{\cdot j}$ *patikimumu*, t. y. jis rodo, kiek patikimas yra $\overline{Y}_{\cdot j}$, kaip parametro β_{0j} įvertinys. Iš tikrųjų, jei λ_j didelis, tai β^*_{0j} apytiksliai lygus $\overline{Y}_{\cdot j}$. Kita vertus, jei λ_j – mažas, tai β^*_{0j} „susitraukia“ iki $\hat{\gamma}_{00}$. Mūsų nagrinėjamo pavyzdžio svertinio įvertinio idėja tokia: jei tiriama daug j-osios mokyklos mokinių, tai natūralu, kad visos mokyklos vidurkis turi būti panašus į turimos imties duomenų vidurkį. Kita vertus, jei mokinių yra nedaug, tai žiūrima, kokie yra kitų mokyklų vidurkiai. Taigi svertinis įvertinys – patikslintas įvertinys. Plačiau įvertinių skaičiavimo metodus aptarsime 1.7 skyrelyje. Dabar tik pateiksime svertinio įvertinio pavyzdį penktosios iš nagrinėtų mokyklų. Tyrime dalyvavo $n_5 = 11$ jos mokinių, jų matematikos testo rezultatų vidurkis buvo $\overline{Y}_{\cdot 5} = 467{,}12$. Primename, kad jau gavome įverčius $\hat{\sigma}^2 = 5117{,}19$, $\hat{\tau}_{00} = 2356{,}11$, $\hat{\gamma}_{00} = 499{,}47$. Randame $\overline{Y}_{\cdot 5}$ patikimumą:

$$\lambda_5 = \frac{2356{,}11}{2356{,}11 + 5447{,}19/11} = 0{,}83.$$

Svertinis mokyklos testo rezultatų vidurkio įvertis yra

$$\beta^*_{05} = 0{,}83 \cdot 467{,}12 + 0{,}16 \cdot 499{,}47 = 472{,}46.$$

Norime atkreipti dėmesį, kad sudėtingesnių HLM modelių koeficiento β^*_{0j} interpretacija gali keistis.

Pakartosime svarbiausią šio skyrelio informaciją.

Besąlyginio hierarchinio modelio analizė parodo, ar reikia atsižvelgti į hierarchinę duomenų struktūrą. Požymiai, kad atsižvelgti reikia, yra tokie:

a) didelė *ICC* reikšmė,
b) statistiškai reikšmingas patvirtinimas, kad $\tau_{00} > 0$.

Besąlyginis modelis naudojamas palyginimui su kitais sudėtingesniais modeliais.

Didelės įverčių $\hat{\sigma}^2$ ir $\hat{\tau}_{00}$ reikšmės rodo, kad reikia nagrinėti HLM modelius su didesniu skaičiumi kintamųjų.

Trumpai apibūdinsime gautas išvadas apie mokinių matematikos testo rezultatų besąlyginį modelį.

Nustatėme statistiškai reikšmingus mokinių rezultatų skirtumus ir statistiškai reikšmingą mokyklų įtaką rezultatams. Mokyklų įtaką parodė ir nemaža koeficiento *ICC* reikšmė. Taigi į hierarchinę duomenų struktūrą reikia atsižvelgti. Kita vertus, didelės $\hat{\sigma}^2$ ir $\hat{\tau}_{00}$ reikšmės rodo, kad vertėtų paieškoti kintamųjų, lemiančių mokinių ir mokyklų skirtumus.

1.4. Bendrasis HLM modelis

Matematikas netiki jokia teorija, kol ji neįrodyta. Fizikas tiki bet kokia teorija, kol ji nepaneigta. Filosofui – nesvarbu.

Panagrinėkime nesudėtingą bendrojo HLM modelio pavyzdį. Tarkime, kad kintamasis Y priklauso nuo kintamųjų X_1, X_2, X_3.

- *Pirmojo lygmens priklausomybė*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e.$$

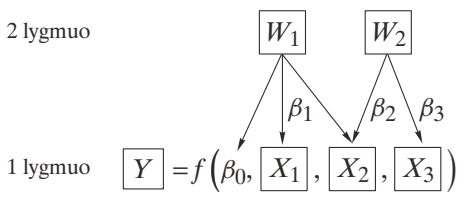
- *Antrojo lygmens priklausomybės*:

$$\begin{aligned}
\beta_0 &= \gamma_{00} + \gamma_{01} W_1 + u_0,\\
\beta_1 &= \gamma_{10} + \gamma_{11} W_1 + u_1,\\
\beta_2 &= \gamma_{20} + \gamma_{21} W_1 + \gamma_{22} W_2 + u_2,\\
\beta_3 &= \gamma_{30} + \gamma_{32} W_2 + u_3.
\end{aligned}$$

Matome, kad kintamojo Y reikšmės priklauso ir nuo pirmojo lygmens kintamųjų X_1, X_2, X_3, ir nuo antrojo lygmens kintamųjų W_1, W_2. Tai dar akivaizdžiau, kai antrojo lygmens koeficientai įstatomi į pirmojo lygmens lygtį ir gaunamas *jungtinis modelis*:

$$\begin{aligned}
Y = {} & \gamma_{00} + \gamma_{01} W_1 + \gamma_{10} X_1 + \gamma_{11} W_1 X_1 + \gamma_{20} X_2 + \gamma_{21} W_1 X_1 + \gamma_{22} W_2 X_2\\
& + \gamma_{30} X_3 + \gamma_{32} W_2 X_3 + [u_0 + X_1 u_1 + X_2 u_2 + X_3 u_3 + e].
\end{aligned} \tag{2.1.4}$$

Laužtiniais skliaustais pažymėjome modelio atsitiktinę dalį. Modelyje aprašytas priklausomybes galima pavaizduoti grafiškai (žr. 2.1.5 pav.).



2.1.5 pav. Pavyzdžio priklausomybių schema

Paveiksle nėra nurodyta tikslių kintamųjų indeksų, čia tik pateikiamos esminės priklausomybės – koks kintamasis nuo kokio priklauso. Dažniausiai tokio nesudėtingo užrašo visiškai pakanka, norint sumodeliuoti spėjamas priklausomybes. Matematiniu požiūriu jis nėra tikslus, labiau atitinka šnekamąją kalbą. Sakome: spėjame, kad mokinio rezultatai priklauso nuo išklausytų pamokų skaičiaus ir nuo mokytojo darbo stažo. Tikslią matematinę formuluotę atitiktų pasakymas – j-ojo mokytojo i-ojo mokinio rezultatas priklauso nuo jo paties išmoktų pamokų skaičiaus ir nuo mokytojo darbo stažo.

Bendrąjį dviejų lygmenų hierarchinį tiesinį modelį užrašysime tiksliai matematiškai, indeksais nurodydami priklausomybę grupei ir individui. Tarkime, kad turime $j = 1, 2, \ldots, J$ grupių. Kiekvienos grupės individų rezultatai priklauso nuo kintamųjų X_k, $k = 1, 2, \ldots, K$.

Pirmojo lygmens lygtis yra

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \cdots + \beta_{Kj} X_{Kij} + e_{ij}; \tag{2.1.5}$$

čia $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$.

Aišku, kad X_{kij} žymi j-osios grupės i-ojo individo k-ojo kintamojo reikšmę.

Kiekvienas iš $K+1$ pirmojo lygmens modelio parametrų (β) antrajame lygmenyje traktuojamas kaip priklausomas kintamasis.

Antrojo lygmens lygtys:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \cdots + \gamma_{0S} W_{Sj} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + \cdots + \gamma_{1S} W_{Sj} + u_{1j}, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \beta_{Kj} = \gamma_{K0} + \gamma_{K1} W_{1j} + \cdots + \gamma_{KS} W_{Sj} + u_{Kj}. \end{cases} \tag{2.1.6}$$

Vengdami indeksų gausos, aprašėme šiek tiek suprastintą bendrojo modelio variantą, laikydami visus β priklausančiais nuo tų pačių kintamųjų W_{sj}. Sudarant modelius, nebūtina į kiekvieną lygtį įtraukti *visus* W_{sj}. Pavyzdžiui, galime sudaryti modelį, kuriame β_{0j} priklausys tik nuo W_{1j}, o β_{1j} – tik nuo W_{2j}. Be to, dalį kintamųjų galima u_{ij} laikyti lygiais nuliui, t. y. spėti, kad nusakyta absoliučiai tiksli priklausomybė ir jokios paklaidos nėra.

Bendrojo modelio prielaidos:

1) $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ ir e_{ij} tarpusavyje yra nepriklausomi;
2) $u_{kj} \sim \mathcal{N}(0, \tau_{kk})$;
3) $\mathbb{C}\mathrm{ov}(u_{kj}, u_{mj}) = \tau_{km}$;
4) u_{kj} ir u_{ml} su visais k, m yra nepriklausomi, jei tik $j \neq l$;
5) u_{ml} ir e_{ij} yra nepriklausomi;
6) kintamasis Y yra intervalinis;
7) kintamieji $X_1, X_2, \dots, X_K, W_1, \dots, W_S$ yra intervaliniai arba pseudokintamieji.

Norime pabrėžti, kad suformulavome klasikinį bendrąjį dviejų lygmenų HLM modelį. Keisdami modelio prielaidas, gautume kitus, dar bendresnius, modelius. Apie kai kuriuos iš jų pakalbėsime atskirame skyrelyje.

Kaip matome iš modelio prielaidų, X ir W nėra atsitiktiniai. Kaip ir regresijoje, tai yra standartinis modelio reikalavimas. Galima būtų jo atsisakyti, tačiau tada tektų visus reikalavimus perrašyti naudojantis sąlyginiais vidurkiais ir užrašai taptų gerokai sudėtingesni. Nors parametrų γ skaičiavimo formulės nepasikeistų. Primename, kad HLM modeliuose priklausomo kintamojo Y stebėjimai koreliuoja. Pirmasis prielaidų reikalavimas sutampa su tiesinės regresijos reikalavimu, kad atsitiktinės paklaidos turi būti nepriklausomos ir normaliai pasiskirsčiusios. Antrasis reikalavimas panašus – antrojo lygmens paklaidos taip pat normaliai pasiskirsčiusios. Atkreipiame dėmesį, kad priklausomybė tarp antrojo lygmens kintamųjų galima tik toje pačioje grupėje (trečiasis ir ketvirtasis reikalavimai). Pavyzdžiui, natūralu manyti, kad tą pačią mokyklą apibūdinantys kintamieji yra susiję, o skirtingų mokyklų kintamieji jau nepriklausomi. Penktasis reikalavimas – antrojo ir pirmojo lygmenų kintamųjų paklaidos turi būti nepriklausomos. Šeštasis ir septintasis yra tipiniai regresijos reikalavimai. Pseudokintamasis yra dvireikšmis (0 ir 1) kintamasis. Pseudokintamųjų neturi būti pernelyg daug.

Yra nusistovėję tradiciniai parametrų pavadinimai.

Visus γ vadinsime fiksuoto poveikio parametrais, o σ^2, τ_{ij} – atsitiktinio poveikio parametrais.

Įstatę (2.1.6) išraiškas į (2.1.5) lygtį, gauname vadinamąjį *jungtinį*, arba *mišrųjį, modelį*:

$$Y_{ij} = \gamma_{00} + \sum_{s=1}^{S} \gamma_{0s} W_{sj} + \sum_{k=1}^{K} \sum_{s=1}^{S} \gamma_{ks} W_{sj} X_{kij} + \left[u_{0j} + \sum_{k=1}^{K} X_{kij} u_{kj} + e_{ij} \right]. \qquad (2.1.7)$$

Kai jungtiniame modelyje yra pirmojo ir antrojo lygmenų kintamųjų sandauga, tai sakoma, kad yra atitinkamų kintamųjų *sąveika* (interakcija). Pavyzdžiui, (2.4) modelyje buvo W_1 ir X_1, W_1 ir X_2, W_2 ir X_3 sąveikos. Jungtinis modelio užrašas leidžia geriau įsivaizduoti, nuo ko priklauso Y.

Kintamųjų ir jų priklausomybių parinkimas sudarant bendrąjį modelį daug kur remiasi logika. Pirmiausia nusprendžiama, nuo kokių aiškinamųjų kintamųjų (regresorių) gali priklausyti pirmojo lygmens paaiškinamojo (regresuojamo) kintamojo reikšmės. Tada mąstoma, kokie antrojo lygmens kintamieji gali būti svarbūs ir per kokius pirmojo lygmens regresorius jų įtaka pasireikš. Viena iš problemų, kurias reikia išspręsti šiame etape, – ar naudoti centruotus kintamuosius. Iš pradinio kintamojo atėmus jo reikšmių vidurkį, gaunamas vadinamasis *centruotas* kintamasis.

Kam reikia centravimo? Dažniausiai tinkamai centruoti kintamieji palengvina parametrų įverčių interpretavimą. Palyginkime du modelius, kai mokinio testo rezultatas priklauso nuo jo socialinio ir ekonominio statuso. Pirmuoju atveju naudojamas pats socialinis ir ekonominis statusas (*SES*), antruoju – *CSES*, kuris kiekvienam mokiniui apskaičiuojamas iš jo *SES* atėmus visų tos mokyklos mokinių *SES* vidurkį:

$$\text{rezultatas}_{ij} = \beta_{0j} + \beta_{1j} SES_{ij} + e_{ij} \quad \text{ir} \quad \text{rezultatas}_{ij} = \beta_{0j} + \beta_{1j} CSES_{ij} + e_{ij}.$$

Pirmuoju atveju β_{0j} žymi mokinio, kurio socialinio ir ekonominio statuso reikšmė lygi 0, testo rezultatų vidurkį. Realiai gyvenime žmonių su $SES = 0$ nėra, taigi β_{0j} interpretacija šiuo atveju gana abejotina. Antruoju atveju β_{0j} žymi j-osios mokyklos mokinių su vidutiniu *SES* matematikos testo rezultatų vidurkį. Taigi šiuo atveju koeficiento β_{0j} interpretavimas kur kas natūralesnis. Beje, galimi ir kitokie centravimai. Pavyzdžiui, galima iš mokinio rezultatų atimti ne jo mokyklos, o visų mokyklų mokinių rezultatų vidurkį. Tada β_{0j} bus interpretuojamas kaip bendrasis visų tiriamų mokinių rezultatų vidurkis.

Centruoti yra tokie kintamieji, iš kurių atimtas jų vidurkis.
Kintamųjų centravimas parenkamas atsižvelgiant į norimą rezultatų interpretaciją.

Realių tyrimų duomenys centruojami jų empiriniu vidurkiu. Vadinasi, papildžius duomenų imtį naujais stebėjimais, reikia iš naujo apskaičiuoti vidurkį ir vėl centruoti visus stebėjimus.

Kaip praktiškai yra parenkamas tinkamiausias modelis? Hierarchinių duomenų (kaip, beje, ir nehierarchinių) analizė pradedama nuo paprasčiausio besąlyginio modelio sudarymo. Modelis specifikuojamas tolydžio – pradedant nuo 1-ojo lygmens modelio. Tuomet konstruojamas sudėtingesnis modelis, kuris lyginamas su prieš tai buvusiu, ir t. t. Atsižvelgiama į γ, σ^2 ir τ_{ij} įverčius. Plačiau tai panagrinėsime 1.7 skyrelyje.

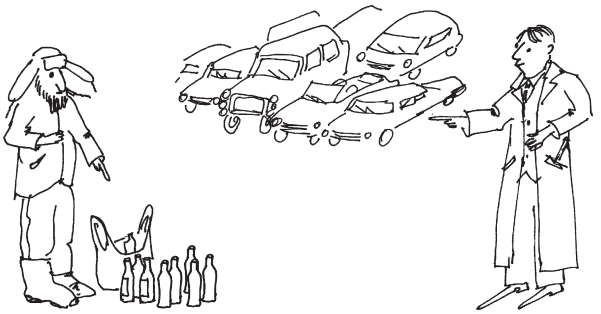
1.5. Dviejų lygmenų modelių pavyzdžiai

Džiunglių kaimelio vadas pasiuntė tinginį sūnų mokytis į misionierių mokyklą. Po pirmo semestro vadas buvo maloniai nustebintas gerais sūnaus pažymiais. Vis dėlto jis norėtų sužinoti, kas taip įkvėpė sūnui norą mokytis.
– Žinai, – aiškina sūnus, – pačią pirmą dieną nuėjęs į matematikos kabinetą pamačiau ant sienos prie pliuso ženklo prikaltą vyruką ir iškart supratau, kad jie čia nejuokauja.

Pateiksime keletą dviejų lygmenų modelių. Vartosime tokius žymenis:

$CSES_{ij}$ – mokyklos vidurkiu centruotas i-ojo mokinio socialinis ir ekonominis statusas j-ojoje mokykloje;
$MSES_j$ – j-osios mokyklos socialinio ir ekonominio statuso vidurkis;
VK_j – mokyklos tipas (0 – Vilniaus mokykla, 1 – kaimo mokykla);
MAT_{ij} – matematikos testo rezultatai.

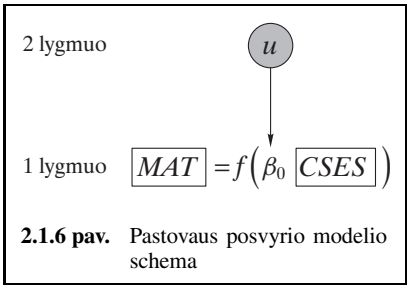
Socialinio ir ekonominio statuso įvertinimas (*SES*) yra daugelio požymių konstrukcija. Nenorėdami apsunkinti skaitytojo (ypač needukologo) detaliais aprašymais, tik paminėsime, kad autorių sudarytasis *SES* įvertinimas buvo pagrįstas tėvų išsilavinimo įvertinimu, turimų namuose knygų skaičiumi ir pan.



SES ir matematika

Pastovaus posvyrio modelis. Tarkime, kad kiekvieno mokinio matematikos testo rezultatai *MAT* priklauso tik nuo *CSES* ir ta priklausomybė visose mokyklose tokia pat, tačiau gali skirtis papildomas mokyklos indėlis (žr. 2.1.6 pav.).

Primename, kad visi nagrinėjami modeliai yra tiesiniai, o paveikslėliuose nurodomas simbolinis užrašas $f(\ldots)$ tik vengiant daugiklių gausos. Modelio pirmojo lygmens lygtis



2.1.6 pav. Pastovaus posvyrio modelio schema

yra

$$MAT_{ij} = \beta_{0j} + \beta_{1j} CSES_{ij} + e_{ij};$$

čia $j = 1, \ldots, J$, $i = 1, \ldots, n_j$. Antrojo lygmens lygtys yra

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j}, \\ \beta_{1j} = \gamma_{10}; \end{cases}$$

čia $j = 1, \ldots, J$. Antroji lygtis parodo, kad visų mokyklų β_{1j} yra ta pati konstanta. Kitaip tariant, matematikos testo rezultato priklausomybė nuo *CSES* visose mokyklose vienoda – nėra taip, kad vienose mokyklose socialinis ir ekonominis statusas svarbesnis nei kitose. Jeigu turėdami tokio modelio duomenis braižytume *MAT* priklausomybės nuo *CSES* grafikus, tai visų mokyklų jie būtų vienodi ir skirtųsi tik postūmiu (žr. 2.1.7).

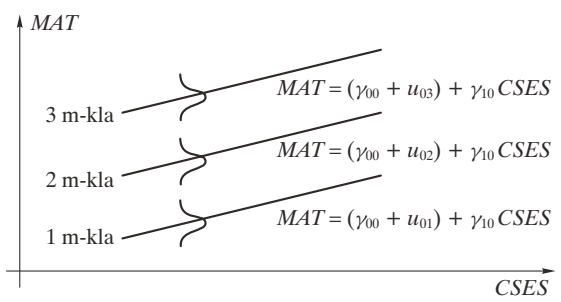
Kadangi u_{0j} yra atsitiktiniai dydžiai, grafike tai akcentuojama pasuktais normaliojo tankio grafikais, o tiesės iliustruoja tik priklausomybę konkrečioms imtims. Primename, kad tiesės lygtyje $y = a + bx$ postūmio parametras a parodo, kaip aukštai virš ašies x bus tiesė, o posvyrio parametras b parodo, kiek tiesė bus pasvirusi virš ašies x.

Įsistatę antrojo lygmens lygtis į pirmojo lygmens lygtį, gauname jungtinio modelio užrašą:

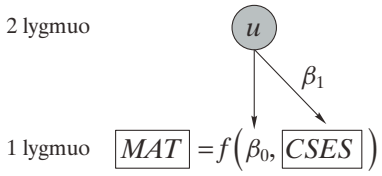
$$MAT_{ij} = \gamma_{00} + \gamma_{10} CSES_{ij} + [u_{0j} + e_{ij}]. \tag{2.1.8}$$

Laužtiniais skliaustais pažymėjome atsitiktinę paklaidą.

Atsitiktinio postūmio ir posvyrio modelis. Tarkime, kad kiekvieno mokinio matematikos testo rezultatai priklauso tik nuo *CSES*, bet ta priklausomybė mokyklose gali skirtis, t. y. vienose mokyklose socialinis ir ekonominis statusas daro labai menką įtaką mokinių rezultatams, o kitose jis labai svarbus (žr. 2.1.8 pav.).



2.1.7 pav. *MAT* priklausomybės nuo *CSES* grafikai



2.1.8 pav. Atsitiktinio postūmio ir posvyrio modelio priklausomybių schema

Modelio pirmojo lygmens lygtis lieka ta pati:

$$MAT_{ij} = \beta_{0j} + \beta_{1j} CSES_{ij} + e_{ij}.$$

Antrojo lygmens lygtys yra tokios:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + u_{1j}. \end{cases}$$

Antroji lygtis reiškia, kad β_{1j} skirtingose mokyklose gali būti skirtinga (žr. 2.1.9 pav.).

Įsistatę antrojo lygmens lygtis į pirmojo lygmens lygtį, gauname jungtinę lygtį

$$MAT_{ij} = \gamma_{00} + \gamma_{10} CSES_{ij} + [u_{1j} CSES_{ij} + u_{0j} + e_{ij}]. \qquad (2.1.9)$$

Palyginę (2.1.8) ir (2.1.9) formules, matome, kad abi lygtys skiriasi tik paklaidomis.

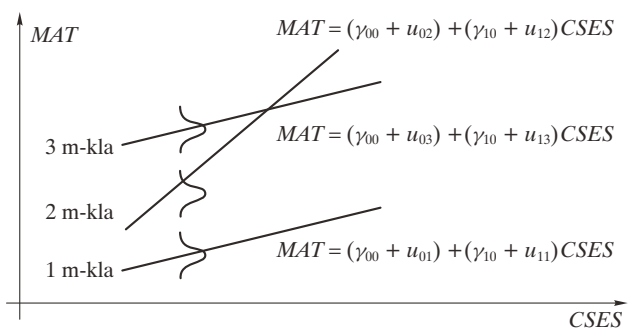
Modelis su antrojo lygmens intervaliniu kintamuoju. Tariame, kad kiekvieno mokinio matematikos testo rezultatai priklauso tik nuo *CSES*, o mokyklos įtaka priklauso nuo *MSES* (žr. 2.1.10 pav.).

Modelio pirmojo lygmens lygtis lieka ta pati:

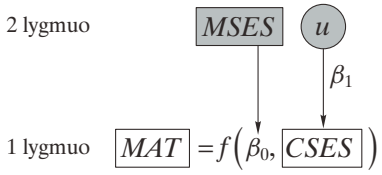
$$MAT_{ij} = \beta_{0j} + \beta_{1j} CSES_{ij} + e_{ij}.$$

Antrojo lygmens lygtys yra tokios:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} MSES_j + u_{0j}, \\ \beta_{1j} = \gamma_{10} + u_{1j}. \end{cases}$$



2.1.9 pav. Posvyrio parametro vaidmuo



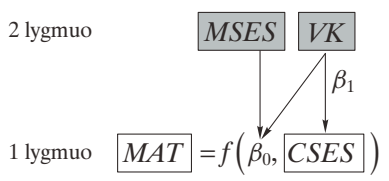
2.1.10 pav. Modelio su antrojo lygmens intervaliniu kintamuoju schema

Įsistatę antrojo lygmens lygtis į pirmojo lygmens lygtį, gauname bendrąjį užrašą:

$$MAT_{ij} = \gamma_{00} + \gamma_{01} MSES_j + \gamma_{10} CSES_{ij} + [u_{1j} CSES_{ij} + u_{0j} + e_{ij}]. \qquad (2.1.10)$$

Šiame pavyzdyje tarėme, kad mokyklos socialinis ir ekonominis statusas turi įtakos tik postūmio parametrui. Taigi manome, kad visų konkrečios mokyklos mokinių rezultatai padidėja (sumažėja) vienodai, neatsižvelgiant į pačių vaikų socialinį ir ekonominį statusą. Jeigu manytume, kad mokyklos *SES* turi įtakos ir posvyriui (yra tarpusavio priklausomybė tarp mokinio *CSES* ir mokyklos *SES*), tai turėtume įtraukti *MSES* ir į β_{1j} aprašančią lygtį.

Modelis su kategoriniu antrojo lygmens kintamuoju. Tarkime, kad kiekvieno mokinio matematinio testo rezultatai priklauso tik nuo *CSES*, o mokyklos įtaka priklauso nuo *MSES* ir nuo vietovės, t. y. nuo regiono kintamojo *VK* (žr. 2.1.11 pav.).



2.1.11 pav. Modelio su kategoriniu antrojo lygmens kintamuoju schema

Modelio pirmojo lygmens lygtis yra

$$MAT_{ij} = \beta_{0j} + \beta_{1j} CSES_{ij} + e_{ij}.$$

Antrojo lygmens lygtys yra tokios:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} MSES_j + \gamma_{02} VK_j + u_{0j}, \\ \beta_{1j} = \gamma_{10} + \gamma_{12} VK_j + u_{1j}. \end{cases}$$

Įsistatę antrojo lygmens lygtis į pirmojo lygmens lygtį, gauname bendrąjį užrašą:

$$\begin{aligned} MAT_{ij} = {} & \gamma_{00} + \gamma_{01} MSES_j + \gamma_{02} VK_j \\ & + \gamma_{10} CSES_{ij} + \gamma_{12} VK_j CSES_{ij} \\ & + [u_{1j} CSES_{ij} + u_{0j} + e_{ij}]. \end{aligned} \qquad (2.1.11)$$

Atkreipiame dėmesį, kad kintamasis *VK* yra vadintinas pseudokintamuoju, nes įgyja tik dvi reikšmes (nėra intervalinis).

1.6. Matricinė modelio formuluotė

Paklausiau docento patyrusio,
Kuo plyta nuo matricos skiriasi.
Žinau, kad ir ką besakys,
Vis tiek jis meluos per akis.

VU matematikų „krikštynų giesmė“, 1978 m.

Matricinė HML modelio formuluotė leidžia geriau įsivaizduoti modelio kintamųjų priklausomybės struktūrą. Be to, matricinis modelio užrašas leidžia trumpai paaiškinti matematines idėjas, naudojamas darant statistines išvadas. Kita vertus, praktiškai sudarant modelius, matricinis užrašas visai nebūtinas. Todėl norime atkreipti dėmesį, kad skaitytojai, kuriuos domina tik taikomieji HLM aspektai, šį skyrelį gali praleisti[7].

2.1.1 pavyzdys. Visų pirma parodysime, kaip matricomis užrašomas paprastas dviejų lygmenų modelis. Pirmojo lygmens lygtis yra

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij},$$

čia $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$, $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ ir e_{ij} visi yra nepriklausomi.

Antrojo lygmens lygtys yra tokios:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + u_{1j}, \end{cases}$$

čia visi $u_{kj} \sim \mathcal{N}(0, \tau_{kk})$, taip pat visi u_{kj} ir e_{ij} yra tarpusavyje nepriklausomi.

Įvesime tokius žymenis:

$$\mathbf{Z}_{ij} = \begin{pmatrix} 1 & X_{ij} \end{pmatrix}, \qquad \boldsymbol{\beta}_j = \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix}, \tag{2.1.12}$$

$$\mathbf{W}_j = \begin{pmatrix} 1 & W_{1j} & 0 & 0 \\ 0 & 0 & 1 & W_{1j} \end{pmatrix}, \qquad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix}, \qquad \mathbf{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix}. \tag{2.1.13}$$

Tada abi antrojo lygmens lygtys užrašomos viena matricine lygybe

$$\boldsymbol{\beta}_j = \mathbf{W}_j \boldsymbol{\gamma} + \mathbf{u}_j,$$

o jungtinį modelį galima užrašyti taip:

$$\begin{aligned} Y_{ij} &= \begin{pmatrix} 1 & X_{ij} \end{pmatrix} \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} + e_{ij} = \mathbf{Z}_{ij} \boldsymbol{\beta}_j + e_{ij} \\ &= \mathbf{Z}_{ij} (\mathbf{W}_j \boldsymbol{\gamma} + \mathbf{u}_j) + e_{ij} = \mathbf{Z}_{ij} \mathbf{W}_j \boldsymbol{\gamma} + \mathbf{Z}_{ij} \mathbf{u}_j + e_{ij}. \end{aligned}$$

Pažymėjus

$$\mathbf{Y}_j = \begin{pmatrix} Y_{1j} \\ Y_{2j} \\ \ldots \\ Y_{n_j j} \end{pmatrix}, \qquad \mathbf{X}_j = \begin{pmatrix} \mathbf{Z}_{1j} \mathbf{W}_j \\ \mathbf{Z}_{2j} \mathbf{W}_j \\ \ldots \\ \mathbf{Z}_{n_j j} \mathbf{W}_j \end{pmatrix}, \qquad \mathbf{Z}_j = \begin{pmatrix} \mathbf{Z}_{1j} \\ \mathbf{Z}_{2j} \\ \ldots \\ \mathbf{Z}_{n_j j} \end{pmatrix}, \qquad \mathbf{e}_j = \begin{pmatrix} e_{1j} \\ e_{2j} \\ \ldots \\ e_{n_j j} \end{pmatrix},$$

gaunamas matricinis užrašas kiekvienam $j = 1, 2, \ldots, J$:

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\gamma} + \mathbf{Z}_j \mathbf{u}_j + \mathbf{e}_j.$$

Galų gale visą modelį irgi galima užrašyti vien matricomis. Pažymėkime:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \ldots \\ \mathbf{Y}_J \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \ldots \\ \mathbf{X}_J \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{Z}_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \mathbf{Z}_J \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \ldots \\ \mathbf{u}_J \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \ldots \\ \mathbf{e}_J \end{pmatrix}.$$

[7] Užteks perskaityti „krikštynų giesmę“.

Tada visas HLM modelis užrašomas taip:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \mathbf{e}. \tag{2.1.14}$$

Kai kintamųjų yra daugiau, aprašytame (2.1.5) ir (2.1.6) lygtimis modelyje (2.1.12) ir (2.1.13) žymenis reikia pakeisti tokiais:

$$\mathbf{Z}_{ij} = \begin{pmatrix} 1 & X_{1ij} & X_{2ij} & \dots & X_{Kij} \end{pmatrix}, \qquad \boldsymbol{\beta}_j = \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \dots \\ \beta_{Kij} \end{pmatrix}, \qquad \mathbf{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \\ \dots \\ u_{Kj} \end{pmatrix},$$

$$\mathbf{W}_j = \begin{pmatrix} 1 & W_{1j} & \dots & W_{Sj} & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & W_{1j} & \dots & W_{Sj} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & W_{Sj} \end{pmatrix}, \qquad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{00} \\ \dots \\ \gamma_{0S} \\ \gamma_{10} \\ \dots \\ \gamma_{1S} \\ \dots \\ \gamma_{KS} \end{pmatrix}.$$

Nesunku pastebėti, kad visi likę pažymėjimai lieka nepakitę. Taigi pirmojo lygmens priklausomybes galima užrašyti taip:

$$\mathbf{Y}_j = \mathbf{Z}_j\boldsymbol{\beta}_j + \mathbf{e}_j. \tag{2.1.15}$$

Antrojo lygmens priklausomybės užrašomos taip:

$$\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j. \tag{2.1.16}$$

Pasinaudoję ankstesniais žymenimis, įsitikiname, kad (2.1.14) lygybė taip pat galioja.

Matricinis bendrojo HLM modelio (MBM) užrašas yra

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \mathbf{e};$$

čia: $\mathbf{X}$ – fiksuoto poveikio kintamieji,
$\boldsymbol{\gamma}$ – fiksuoto poveikio parametrai,
$\mathbf{Z}$ – atsitiktinio poveikio kintamieji,
$\mathbf{u}$ – antrojo lygmens atsitiktinės paklaidos,
$\mathbf{e}$ – pirmojo lygmens atsitiktinės paklaidos.

Matome, kad modelis išsiskaido į tris dėmenis, kuriuos sąlygiškai galima įvardyti, kaip fiksuotų poveikių dėmenį, atsitiktinių poveikių dėmenį ir atsitiktinę paklaidą. Nagrinėjant bendrąjį HLM modelį, mums reiks įvertinti ne tik nežinomus parametrus, bet ir atsitiktinių paklaidų įtaką. Visos paklaidos (e_{ij} ir u_j) yra normaliai pasiskirsčiusios ir turi nulinius vidurkius, todėl informaciją apie jų elgesį ir galimas priklausomybes nusako kovariacijų matricos. Pažymėkime vektoriaus $\mathbf{u}$ kovariacijų matricą simboliu $\mathbf{T}$, vektoriaus $\mathbf{e}$ kovariacijų matricą simboliu $\mathbf{R}$, o vektoriaus $\mathbf{Y}$ kovariacijų matricą – simboliu $\mathbf{V}$.

Visos paklaidos u_{ij}, e_{ij} yra nepriklausomos, t. y.

$$\mathbb{E}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \qquad \mathbb{C}\mathrm{ov}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}.$$

Be to, $\mathbf{u}_j$ nepriklauso nuo $\mathbf{u}_m$, o $\mathbf{e}_j$, nuo $\mathbf{e}_m$, kai $j \neq m$. Todėl $\mathbf{T}$ ir $\mathbf{R}$ turi blokines struktūras. Pažymėkime, $\mathbf{u}_j$ kovariacijų matricą simboliu $\mathbf{T}_j$, o $\mathbf{e}_j$ kovariacijų matricą – simboliu $\mathbf{R}_j$. Tada

$$\mathbf{T} = \begin{pmatrix}\mathbf{T}_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \mathbf{T}_2 & \mathbf{0} & \dots & \mathbf{0}\\ \dots & \dots & \dots & \dots & \dots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{T}_J\end{pmatrix}, \qquad \mathbf{T}_j = \begin{pmatrix}\tau_{00} & \tau_{01} & \dots & \tau_{0K}\\ \tau_{10} & \tau_{11} & \dots & \tau_{1K}\\ \dots & \dots & \dots & \dots\\ \tau_{K0} & \tau_{K1} & \dots & \tau_{KK}\end{pmatrix} \tag{2.1.17}$$

ir

$$\mathbf{R} = \begin{pmatrix}\mathbf{R}_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_2 & \mathbf{0} & \dots & \mathbf{0}\\ \dots & \dots & \dots & \dots & \dots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{R}_J\end{pmatrix}. \tag{2.1.18}$$

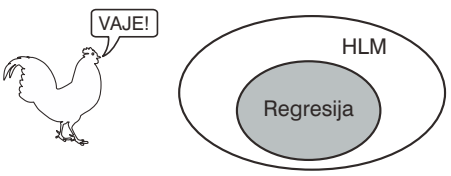
Iki šiol nagrinėtuose modeliuose $\mathbf{R}_j = \sigma^2\mathbf{I}$, čia $\mathbf{I}$ yra vienetinė matrica.

Galima įsitikinti, kad $\mathbf{Y}$ kovariacijų matrica lygi

$$\mathbf{V} = \mathbf{ZTZ}' + \mathbf{R}. \tag{2.1.19}$$

Ši lygybė išlieka teisinga ir tada, kai e_{ij} priklauso viena nuo kitos, t. y. $\mathbf{R}_j$ turi sudėtingesnę struktūrą nei $\sigma^2\mathbf{I}$. Parinkdami vienokią ar kitokią matricų $\mathbf{T}$ ir $\mathbf{R}$ struktūrą, galime kompaktiškai užrašyti pačius įvairiausius modelius. Ateityje matysime, kad jie gali būti daug sudėtingesni nei standartiniu atveju, kai $\mathbf{R} = \sigma^2\mathbf{I}$.

Kai $\mathbf{Z} = \mathbf{0}$, tai bendrasis HLM modelis užrašomas $\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{e}$. Tai ne kas kita kaip tiesinės regresijos modelis, kurį jau nagrinėjome pirmoje dalyje. Taigi dar kartą įsitikiname, kad HLM yra tiesinės regresijos modelio apibendrinimas.



Skyrelio pabaigoje norime atkreipti dėmesį į literatūroje vartojamus matricinius HLM žymėjimus. Nagrinėjant paprastąją regresiją, įprasta rašyti $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Dažnai aptinkamas hierarchinio modelio užrašas yra $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$. Logika atrodytų lyg ir graži – prie regresijos modelio pridedamas atsitiktinio poveikio dėmuo. Vis dėlto tokie žymenys nieko bendra neturi su nematriciniame modelyje vartotais žymenimis. Manome, kad nelogiška vektorių, sudarytą tik iš γ, žymėti simboliu $\boldsymbol{\beta}$, todėl palikome simbolį $\boldsymbol{\gamma}$. Kita vertus, literatūroje įprasto kintamųjų prie $\mathbf{u}$ žymens $\mathbf{Z}$ nepakeitėme.

1.7. Parametrų įverčių skaičiavimas ir modelio charakteristikos

– Ar žinai, kad statistikos profesoriaus santuoka žlunga!
– Nieko nuostabaus, jis įpratęs prie regresijos modelių, o jo žmona – neprognozuojama.

Šiame skyrelyje aptariama, kaip skaičiuojami bendrojo HLM modelio parametrų įverčiai. Vengiant indeksų gausos, pavyzdžiams naudojamas paprastas dviejų lygmenų modelis:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij},$$

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + u_{1j}; \end{cases}$$

čia $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$, $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ ir e_{ij} visi yra nepriklausomi. Be to, e_{ij} nepriklauso nuo $u_{kj} \sim \mathcal{N}(0, \tau_{kk})$. Antrojo lygmens paklaidos gali koreliuoti: $\mathbb{C}\mathrm{ov}(u_{0j}, u_{1j}) = \tau_{01} = \tau_{10}$. Atkreipiame dėmesį, kad kintamuosius e_{ij} ir u_{sj} apibūdina jų dispersijos σ^2 ir τ_{ss}, o priklausomybę tarp u_{sj} ir u_{kj} – jų kovariacija τ_{sk}.

Skaičiuojami parametrų įverčiai:

1) fiksuoto poveikio parametrų įverčiai $\hat{\gamma}_{kj}$;
2) atsitiktinio poveikio parametrų įverčiai $\hat{\sigma}^2$, $\hat{\tau}_{kj}$;
3) pirmojo lygmens atsitiktinių koeficientų įverčiai $\hat{\beta}_{mj}$;
4) antrojo lygmens liekamųjų paklaidų įverčiai $\hat{u}_{kj}$.

Modelio tinkamumą duomenims rodo informaciniai indeksai *AIC*, *AICC*, *CAIS*, *BIC*.

Lygindami šiuos įverčius su skaičiuotais besąlyginio modelio atveju, matome, kad kai kurie iš jų nauji, tačiau nebeliko koreliacijos koeficiento *ICC*. Nors apskritai *ICC* galima skaičiuoti, tačiau šio rodiklio interpretacija tampa ne itin prasminga. Prisiminkime besąlyginio modelio pavyzdį apie mokyklas ir mokinių matematikos testo rezultatus. Jokių kitų aiškinamųjų kintamųjų modelyje nebuvo, todėl *ICC* parodė, kokį procentą mokinių skirtumų galima paaiškinti mokyklų įtaka. Bendrajame modelyje yra aiškinamieji kintamieji. Todėl tegalime sakyti, kad *ICC* parodo mokyklų įtaką tada, kai tie kintamieji įgyja nulines reikšmes. Toks pasakymas dažniausiai nelabai prasmingas.

Trumpai aptarsime, kokios matematinės idėjos naudojamos įvertiniams gauti. Prisiminkime bendrąjį HLM modelio užrašą:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + [\mathbf{Z}\mathbf{u} + \mathbf{e}], \qquad \mathbb{C}\mathrm{ov}\,\mathbf{Y} = \mathbb{C}\mathrm{ov}[\mathbf{Z}\mathbf{u} + \mathbf{e}] = \mathbf{V}.$$

Taikant apibendrintąjį mažiausiųjų kvadratų metodą, gaunami fiksuoto poveikio parametrų įvertiniai[8]:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{2.1.20}$$

[8] Angliškos įvertinių išraiškos gautos LNM atveju (žr. 1.1 skyrelį).

Prisiminkime (2.1.15) *lygtį:*

$$\mathbf{Y}_j = \mathbf{Z}_j\boldsymbol{\beta}_j + \mathbf{e}_j, \qquad \mathbb{C}\mathrm{ov}\,\mathbf{e}_j = \mathbf{R}_j.$$

Vėl pasinaudojus apibendrintuoju mažiausiųjų kvadratų metodu, gaunami atsitiktinio poveikio parametrų įvertiniai:

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{Z}_j'\mathbf{R}_j^{-1}\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\mathbf{R}_j^{-1}\mathbf{Y}_j. \tag{2.1.21}$$

Kadangi $\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j$, *tai įstačius įvertinių* $\hat{\boldsymbol{\gamma}}$ *ir* $\hat{\boldsymbol{\beta}}_j$ *išraiškas gaunamas atsitiktinės paklaidos* $\mathbf{u}_j$ *įvertinys*:

$$\hat{\mathbf{u}}_j = \hat{\boldsymbol{\beta}}_j - \mathbf{W}_j\hat{\boldsymbol{\gamma}}. \tag{2.1.22}$$

Pažymėtina, kad atsitiktiniams koeficientams $\boldsymbol{\beta}$ *rasti naudojami ne tik* (2.1.21) *įvertiniai, bet ir svertiniai įvertiniai:*

$$\boldsymbol{\beta}_j^* = \boldsymbol{\Lambda}_j\hat{\boldsymbol{\beta}}_j + (\mathbf{I} - \boldsymbol{\Lambda}_j)\mathbf{W}_j\hat{\boldsymbol{\gamma}}, \quad \boldsymbol{\Lambda}_j = \mathbf{T}_j(\mathbf{T}_j + \mathbf{B}_j)^{-1}; \tag{2.1.23}$$

čia $\mathbf{B}_j$ *yra* $\hat{\boldsymbol{\beta}}_j$ *dispersija. Kai* $\mathbf{e}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, *tai* $\mathbf{B}_j = \sigma^2(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}$. *Nesunku suprasti, kad pirmoji* (2.1.23) *lygybė yra tiesinis dviejų* $\boldsymbol{\beta}_j$ *įvertinių darinys. Jeigu* j*-osios grupės imtyje yra daug duomenų, tai* $\boldsymbol{\beta}_j^*$ *tampa panašus į* $\hat{\boldsymbol{\beta}}_j$, *o jei stebėjimų mažai į* $\mathbf{W}_j\hat{\boldsymbol{\gamma}}$. *Antrasis įvertis gautas tiesiog iš* (2.1.16) *formulės. Svertinio įverčio sudarymo logika tokia – kai grupėje turima daug duomenų, tai jie pakankamai gerai leidžia įvertinti nežinomą parametrą. Tuo atveju, kai duomenų mažai, atsižvelgiama į įvertinį, gautą iš visų duomenų* $\boldsymbol{\gamma}$. *Svertiniai įvertiniai jau buvo minėti* 1.3 *skyrelyje. Yra ir kitokių įverčių skaičiavimo metodų, kuriais galima gauti, pavyzdžiui, tokį* $\mathbf{u}$ *įvertinį:* $\hat{\mathbf{u}} = \mathbf{TZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

Visiems įvertiniams $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$ būdinga tai, kad reikia žinoti kovariacijų matricas $\mathbf{T}_j$ ir $\mathbf{R}_j$ (prisiminkime, kad $\mathbf{V} = \mathbf{ZTZ}' + \mathbf{R}$). Dažniausiai tokios informacijos neturima, todėl vietoj pačių matricų imami šių matricų įverčiai $\widehat{\mathbf{T}}_j$ ir $\widehat{\mathbf{R}}_j$. Yra keliolika skirtingų metodų šiems įverčiams skaičiuoti. Dažniausiai naudojamas didžiausio tikėtinumo metodas (ML) ir apribotojo didžiausio tikėtinumo metodas (REML[9]). REML metodas skiriasi nuo ML metodo tuo, kad optimizuojamas ne didžiausio tikėtinumo funkcijos logaritmas $\ln L$, o REML funkcijos logaritmas

$$\ln L_R = \ln L - \frac{1}{2}\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|.$$

ML ir REML metodų palyginimas

1. REML atsižvelgia į fiksuoto poveikio parametrų skaičių modelyje ir atitinkamai sumažina modelio laisvės laipsnių skaičių.
2. REML efektyvesnis už ML (mažesnės įverčių dispersijos), kai yra stipri stebėjimų koreliacija.
3. Kaip ir ML, REML užtikrina įvertinių suderinamumą ir asimptotinį normalumą.
4. REML koreguoja ML įvertinių kairįjį paslinktumą.
5. REML įverčiai konservatyvesni, t. y. kovariacijų matricos komponenčių įverčių reikšmės didesnės, nei gautosios ML metodu. Skirtumai didėja didėjant fiksuotų parametrų skaičiui modelyje ir grupių skaičiui.

[9] REML – angl. *Restricted Maximum Likelihood*.

Atrodytų, kad REML metodas yra kur kas pranašesnis už ML metodą. Tačiau lyginti dviejų HLM modelių parametrus galima tik tada, kai jų įverčiai skaičiuojami naudojant ML metodą.

Pastabos apie parametrų įverčius. Parametrų įverčių pavyzdžių rasite 1.9 ir 1.11 skyreliuose. Čia pateiksime tik keletą pastabų.

1 Ne visos statistinės programos skaičiuoja įverčius $\hat{\boldsymbol{\beta}}$ ir $\hat{\mathbf{u}}$. Iš tikrųjų pakankamai tiksliai modelio analizei pakanka ir $\hat{\boldsymbol{\gamma}}$ ir $\hat{\sigma}^2$, $\hat{\tau}$ įverčių.

2 Turint įverčius $\hat{\boldsymbol{\gamma}}$, galima užrašyti prognozės funkciją. Skyrelio pradžioje suformuluoto pavyzdžio jungtinis modelis yra

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_{1j} + \gamma_{10} X_{ij} + \gamma_{11} W_{1j} X_{ij} + [u_{0j} + X_{ij} u_{1j} + e_{ij}].$$

Todėl j-os grupės mokinio prognozės funkcija yra

$$Y = \hat{\gamma}_{00} + \hat{\gamma}_{01} W_{1j} + \hat{\gamma}_{10} X + \hat{\gamma}_{11} W_{1j} X.$$

Įstatę konkrečią pirmojo lygmens kintamojo X reikšmę, gausime to mokiniuo pirmojo lygmens tikėtiną Y reikšmę.

3 Didelė $\hat{\sigma}^2$ reiškmė rodo, kad modelyje ne visą Y elgesį paaiškina kintamasis X. Dažnai, kai lyginami du modeliai, kartu pažiūrima, kiek sumažėja naujasis σ^2 įvertis. Tarkime, kad anksščiau minėtame modelyje gauta $\hat{\sigma}^2 = 30$, o įtraukus į pirmojo lygmens lygtį kokį nors naują kintamąjį $\widetilde{X}$, gauto modelio liekanos įvertis yra $\hat{\sigma}^2 = 24$. Tuomet apskaičiuojamas santykinis įverčio pokytis:

$$\frac{30 - 24}{30} = 0{,}20.$$

Gautą reikšmę galima paaiškinti taip: įtraukus į modelį papildomą kintamąjį, 20 % geriau paaiškinami skirtumai tarp mokinių. Aišku, kuo labiau sumažėja dispersijos įvertis, tuo naujasis modelis geresnis. Vis dėlto reikia turėti omenyje, kad modelis turi būti prasmingas. Įtraukus stipriai koreliuotą pirmojo lygmens kintamąjį, antrojo lygmens kintamieji gali būti nebereikalingi.

4 Didelė $\hat{\tau}_{kk}$ reikšmė parodo, kad antrojo lygmens kintamieji ne itin gerai paaiškina atitinkamo atsitiktinio koeficiento β elgesį. Įtraukus papildomus kintamuosius į antrojo lygmens modelius, santykinis įverčio $\hat{\tau}_{kk}$ sumažėjimas skaičiuojamas ir interpretuojamas analogiškai ankstesnei pastabai.

5 Koeficientus β galima įvertinti naudojantis pirmojo lygmens lygtimi (vienas įvertis) ir antrojo lygmens lygtimis (antrasis įvertis). Optimaliu laikomas šių dviejų įvertinių tiesinis darinys – svertinis (2.1.23) įvertinys.

6 Liekamojo nario įvertinys gaunamas į atitinkamą antrojo lygmens lygtį įstačius β ir γ įvertinius. Pavyzdžiui, $\hat{u}_{0j} = \hat{\beta}_{0j} - \hat{\gamma}_{00} - \hat{\gamma}_{01} W_{1j}$.

7 Be pačių įverčių, statistinėse programose dažniausiai skaičiuojami ir jų 95 % pasikliautinieji intervalai.

Informaciniai indeksai. HLM modelio sudarymas – iteracinis procesas. Dažniausiai imamas besąlyginis modelis ir tobulinamas pridedant naujus kintamuosius. Jau minėta, kaip įvertinti dispersijų įverčių pokyčius. Dviem modeliams palyginti taip pat taikomi informaciniai indeksai (informacinių kriterijų reikšmės). Jų reikšmės susijusios su didžiausio tikėtinumo funkcijos logaritmo maksimumu ℓ. Susipažinsime su keletu iš jų tuo atveju, kai ℓ apskaičiuota REML metodu. Visose formulėse n lygus pirmojo lygmens duomenų ir fiksuoto poveikio parametrų skaičiaus skirtumui, d yra atsitiktinio poveikio parametrų skaičius.

- Dvigubas didžiausio tikėtinumo funkcijos logaritmas: -2ℓ;
- Akaikės informacinis kriterijus $AIC = -2\ell + 2d$;
- Hurvičiaus–Cajo indeksas $AICC = -2\ell + 2dn/(n - d - 1)$;
- Bozdogano kriterijus $CAIS = -2\ell + d[\ln n + 1]$;
- Švarco–Bajeso kriterijus $BIC = -2\ell + d \ln n$.

Skyrelio pradžioje lygtimis aprašytame modelyje $d = 4$, nes yra keturi atsitiktinio poveikio parametrai: σ^2, τ_{00}, τ_{11}, τ_{01}.

Kuo kiekvieno indekso reikšmė mažesnė, tuo modelis geriau suderintas su duomenimis.[10] Informaciniai indeksai – santykiniai. Vien pažiūrėję į indekso reikšmę dar negalime nuspręsti ar modelis gerai suderintas su duomenimis. Todėl minėtieji indeksai naudojami lyginant du modelius.

1.8. Hipotezių tikrinimas

Profesorius pateikia statistinių išvadų pavyzdžių.
– Kaip? – stebisi studentas. – Ištyrę tik kelis tūkstančius, darome išvadą apie tris milijonus?
Profesorius replikuoja: – Žinoma, galima tirti ir visą populiaciją. Kai kitą kartą poliklinikoje darysitės kraujo tyrimus, pareikalaukite tyrimui paimti VISĄ kraują...

[10] Kartais indeksai apibrėžiami truputį kitaip. Tuo atveju *didesnė* indekso reikšmė rodo geresnį modelio tinkamumą.

Kaip ir ankstesniame skyrelyje, hipotezės formuluojamos nagrinėjant bendrąjį HLM modelį, aprašomą (2.1.5)–(2.1.6) lygtimis, o kai kurios jų iliustruojamos paprastu dviejų lygmenų modeliu. Primename pastarąjį modelį:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij},$$

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + u_{1j}; \end{cases} \tag{2.1.24}$$

čia $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, $u_{kj} \sim \mathcal{N}(0, \tau_{kk})$ ir $\mathbb{C}\mathrm{ov}(u_{0j}, u_{1j}) = \tau_{01} = \tau_{10}$. Be to, visi e_{ij} yra tarpusavyje nepriklausomi ir nepriklauso nuo u_{kj}.

Statistinės hipotezės formuluojamos nagrinėjant:

1) fiksuoto poveikio parametrus γ_{ks};
2) atsitiktinio poveikio parametrus σ^2 ir τ_{jk};
3) atsitiktinius koeficientus β_{kj};
4) dviejų modelių lygybę.

Hipotezes galima formuluoti apie: a) vieną parametrą ir b) keletą parametrų (parametrų tiesinį darinį). Tyrimuose vyrauja pirmojo tipo hipotezės, todėl joms ir skirsime daugiausia dėmesio. Atkreipsime dėmesį, kad vektorinių hipotezių tikrinimas realizuotas ne kiekvienoje statistinėje programoje.

Panašiai, kaip ir tiesinėje regresijoje, reikia nepamiršti, kad esant didelėms duomenų imtims, nulinės hipotezės dažnai būna atmetamos. Kita vertus, neatmesta nulinė hipotezė apie parametro lygybę nuliui tėra ženklas, kad vertėtų ištirti naują modelį, vietoj to parametro įrašius nulį. Ar verta pereiti prie tokio paprastesnio (redukuotojo) modelio, nusprendžiama atsižvelgus į tai, kaip pasikeitė kitų parametrų įverčiai ir informacinių indeksų (*AIC*, *AICC* ir pan.) reikšmės. Dar kartą pabrėžiame – neatmetus nulinės hipotezės, tam tikras parametras tampa tik „kandidatu“ į nulį.

Jeigu redukuotasis modelis, palyginti su pradiniu, labai suprastėja, tai to parametro keisti nuliu neverta.

Praktiniuose skaičiavimuose esant tokiai situacijai reikia ieškoti naujų aiškinamųjų kintamųjų (o gal net ir naujų modelių) arba prilyginti nuliui kitą parametrą.

Hipotezė apie fiksuoto poveikio parametrą γ_{ks}. Kiekvienam fiskuoto poveikio parametrui γ_{ks} tikrinama hipotezė, ar jis nelygus nuliui.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \gamma_{ks} = 0, \\ H_1\colon\ \gamma_{ks} \neq 0. \end{cases}$$

2 *Sprendimo priėmimo taisyklė*. Turint statistikos p reikšmę, nulinė hipotezė atmetama (t. y. nusprendžiama, kad γ_{ks} statistiškai reikšmingai skiriasi nuo 0), jei $p < \alpha$. Čia α – pasirinktasis reikšmingumo lygmuo. Pavyzdžiui, jei $\alpha = 0{,}05$, tai nulinė hipotezė atmetama, kai $p < 0{,}05$.

Hipotezei tikrinti naudojamasi tuo, kad statistika $t = \hat{\gamma}_{kj}/\sqrt{\mathbf{D}(\hat{\gamma}_{kj})}$, kai H_0 teisinga, turi asimptotinį Stjudento skirstinį su $J - S - 1$ laisvės laipsnių. Čia $\mathbf{D}(\hat{\gamma}_{kj})$ yra ML ar REML metodu gautas $\hat{\gamma}_{kj}$ dispersijos įvertis, o S – antrojo lygmens kintamųjų skaičius.

Ką parodo neatmesta nulinė hipotezė? Tiesiog tuomet galima įtarti, kad iš modelio reikėtų pašalinti kintamąjį, kurio daugiklio reikšmingumą tikrinome. Tarkime, kad nagrinėdami skyrelio pradžioje pavyzdį neatmetėme hipotezės H_0: $\gamma_{01} = 0$. Tada, įstatę į (2.1.24) išraiškas vietoj γ_{10} nulį, gauname naujas antrojo lygmens lygtis:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + u_{1j}. \end{cases}$$

Taigi nors tikrinama hipotezė apie parametro reikšmę, realiai tikrinama, ar kintamasis W_1 reikalingas (ar kintamojo įtaka reikšminga) atsitiktinio postūmio β_{0j} lygtyje. Reikėtų palyginti redukuotojo (be W_{1j}) ir neredukuotojo modelių skaitines charakteristikas ir tik tada nuspręsti, kuris jų geriau tinka imties duomenims.

Vektorinės hipotezės apie parametrą $\boldsymbol{\gamma}$. Vektorinė yra tokia hipotezė, kai tikrinama, ar parametrų γ_{ks}, $k = 1, \ldots, K$, tiesinių darinių vektorius lygus nuliniam vektoriui. Primename, kad parametrų tiesinis darinys gaunamas juos padauginus iš pasirinktų daugiklių ir sudėjus. Kokius daugiklius parinkti? Dažniausiai naudojami standartiniai tokių daugiklių rinkiniai, vadinamieji ortogonalūs kontrastai. Taigi vektorinė hipotezė – tai hipotezė, formuluojama iškart apie keletą parametrų γ_{ks}.

Dažniausiai vektorinių hipotezių prireikia tada, kai norime įsitikinti, kad visose antrojo lygmens lygtyse kuris nors kintamasis nereikalingas. Tuomet formuluojame hipotezę, kad visose lygtyse daugikliai γ_{ks}, $s = 1, 2, \ldots$, prie to kintamojo lygūs nuliui.

Panagrinėkime skyrelio pradžioje suformuluotą modelį. Tarkime, kad norime patikrinti hipotezę, jog kintamasis W_{1j} modelyje nereikalingas, t. y. β_{0j} ir β_{1j} nepriklauso nuo W_{1j}, $j = 1, \ldots, J$. Formalus vektorinis nulinės hipotezės užrašas su tiesiniais dariniais atrodytų taip:

$$H_0\colon \begin{matrix} 0 \cdot \gamma_{00} + 1 \cdot \gamma_{01} = 0, \\ 0 \cdot \gamma_{10} + 1 \cdot \gamma_{11} = 0, \end{matrix} \quad \text{arba} \quad H_0\colon \begin{matrix} \gamma_{01} = 0, \\ \gamma_{11} = 0. \end{matrix} \tag{2.1.25}$$

Kadangi abudu parametrai lygūs nuliui, tai šias išraiškas galima užrašyti kompaktiškiau.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \gamma_{01} = \gamma_{11} = 0, \\ H_1\colon\ \text{taip nėra.} \end{cases}$$

2 *Sprendimo priėmimo taisyklė.* Turint statistikos p reikšmę, nulinė hipotezė atmetama (t. y. nusprendžiama, kad bent vienas iš parametrų γ_{01} arba γ_{11} statistiškai reikšmingai skiriasi nuo 0), jei $p < \alpha$; čia α yra pasirinktasis reikšmingumo lygmuo.

Neatmesta nulinė hipotezė būtų rodiklis, kad reikia pabandyti redukuotą modelį be W_1, t. y. modelį su

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + u_{1j}. \end{cases}$$

Atmesta nulinė hipotezė reikštų, kad visiškai iš modelio pašalinti kintamojo W_1 negalima.

Kontrastus irgi galima užrašyti vektoriniu pavidalu. Naudojantis (2.1.13) matriciniais žymenimis (2.1.25) nulinę hipotezę galima užrašyti tokiu būdu:

$$\mathbf{C}\boldsymbol{\gamma} = \mathbf{0}, \qquad \mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix}, \qquad \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Analogiškai nulinė hipotezė užrašoma ir bendruoju atveju, tik tada $\mathbf{C}$ turi daugiau eilučių ir stulpelių. Hipotezei tikrinti naudojama statistika:

$$H = \hat{\boldsymbol{\gamma}}'\mathbf{C}'(\mathbf{C}\,\mathbf{D}(\hat{\boldsymbol{\gamma}})\mathbf{C}')^{-1}\mathbf{C}\hat{\boldsymbol{\gamma}},$$

kuri esant nulinei hipotezei turi χ^2 skirstinį su tiek laisvės laipsnių, kiek yra kontrastų (t. y. tiek, kiek $\mathbf{C}$ turi eilučių). Čia $\mathbf{D}(\hat{\boldsymbol{\gamma}})$ yra $\hat{\boldsymbol{\gamma}}$ dispersijos įvertis.

Hipotezė apie atsitiktinio poveikio parametrą σ^2. Tikrinama, ar paklaidų dispersija lygi nuliui.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \sigma^2 = 0, \\ H_1\colon\ \sigma^2 > 0. \end{cases}$$

2 *Sprendimo priėmimo taisyklė.* Turėdami statistikos p reikšmę, nulinę hipotezę atmetame (t. y. nusprendžiame, kad σ^2 statistiškai reikšmingai skiriasi nuo 0), jei $p < \alpha$; čia α yra pasirinktasis reikšmingumo lygmuo. Jei $\alpha = 0{,}05$, tai nulinę hipotezę atmesime, kai $p < 0{,}05$.

Hipotezei tikrinti taikomas tikėtinumų santykių kriterijus. Kaip interpretuojama neatmesta nulinė hipotezė? Visų pirma reikia patikrinti, ar duomenų buvo pakankamai daug. Žinoma, kad esant mažoms duomenų imtims nulinės hipotezės atmetamos retai. Jeigu duomenų daug, tai neatmesta nulinė hipotezė gali reikšti vieną iš dviejų dalykų. Galbūt pirmojo lygmens modelis visiškai tiksliai aprašo respondento gautą rezultatą, t. y. nebėra jokių kitų nežinomųjų, kurie dar darytų įtaką rezultatų skirtumams. O gal į modelį įtrauktas kintamasis X labai koreliuoja su Y, nors nėra su juo susijęs priežastiniais ryšiais.

Hipotezės apie atsitiktinio poveikio parametrus τ_{sk}. Visiškai analogiškai formuluojamos ir interpretuojamos hipotezės antrojo lygmens paklaidų dispersijoms τ_{kk}.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \tau_{kk} = 0, \\ H_1\colon\ \tau_{kk} > 0. \end{cases}$$

2 *Sprendimo priėmimo taisyklė.* Turint statistikos p reikšmę, nulinė hipotezė atmetama (t. y. nusprendžiama, kad τ statistiškai reikšmingai skiriasi nuo 0), jei $p < \alpha$; čia α yra pasirinktasis reikšmingumo lygmuo. Dažnai $\alpha = 0{,}05$.

Bendriausiu atveju antrojo lygmens paklaidos gali koreliuoti, todėl dar tikrinamos hipotezės dėl jų koreliuotumo. Nulinė hipotezė – paklaidos nekoreliuoja. Suformuluosime

tokios statistinės hipotezės pavyzdį, imdami skyrelio pradžioje aprašytą modelį. Primename, kad, esant fiksuotiems kintamiesiems W_{1j}, $\mathbb{C}\mathrm{ov}(\beta_{0j}, \beta_{1j}) = \mathbb{C}\mathrm{ov}(u_{0j}, u_{1j}) = \tau_{01} = \tau_{10}$.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \tau_{01} = 0, \\ H_1\colon\ \tau_{01} \neq 0. \end{cases}$$

2 *Sprendimo priėmimo taisyklė.* Turint statistikos p reikšmę, nulinė hipotezė atmetama (t. y. nusprendžiama, kad β_{0j} ir β_{1j} statistiškai reikšmingai koreliuoja), jei $p < \alpha$; čia α yra pasirinktasis reikšmingumo lygmuo.

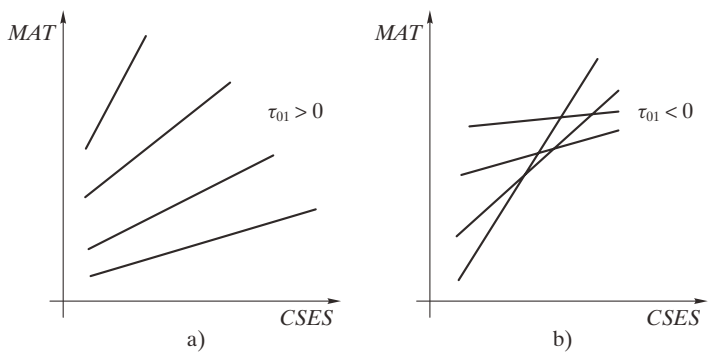
Tikrinant hipotezę naudojamasi tuo, kad statistika $z = \hat{\tau}_{kk}/\sqrt{\mathbf{D}(\hat{\tau}_{kk})}$ turi asimptotiškai normalųjį skirtinį. Čia $\mathbf{D}(\hat{\tau}_{kk})$ yra $\hat{\tau}_{kk}$ dispersijos įvertis. Beje, tikrinant hipotezes apie kovariacijų lygybę nuliui, galimos ir vienpusės alternatyvos, pavyzdžiui, $\tau_{01} > 0$.

Kaip interpretuojama atmesta nulinė hipotezė? Reikia atsižvelgti į parametrų prasmę. Tarkime, kad turime pirmojo lygmens priklausomybę $MAT_{ij} = \beta_{0j} + \beta_{1j}CSES$, čia *MAT* žymi matematikos testo rezultatą, *CSES* – centruotą socialinį ir ekonominį mokinio statusą, o β_{0j} – vidutinį mokyklos matematikos testo rezultatą. Jeigu, patikrinę statistinę hipotezę, nusprendėme, kad $\tau_{01} > 0$, tai galima teigti, kad tose mokyklose, kuriose testo vidurkis aukštesnis, mokinių rezultatų priklausomybė nuo *CSES* yra stipresnė (žr. 2.1.12 pav., a). Jeigu būtų $\tau_{01} < 0$, tai priklausomybė būtų priešinga. Mokyklose, kuriose vidutinis testo rezultatas didesnis, didėjant *CSES* mokinių rezultatai augtų lėčiau, nei mokyklose su mažesniu vidutiniu rezultatu (žr. 2.1.12 pav., b).

Vektorinė hipotezė apie atsitiktinio poveikio parametrus reikštų, kad visą liekamųjų paklaidų kovariacijų matricą lyginame su kokia nors konkrečia kovariacijų matrica. Pavyzdžiui, nulinė hipotezė gali būti užrašoma taip:

$$H_0\colon \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Vektorinių hipotezių tikrinimui naudojami ML arba REML metodu gauti parametrų įvertčiai.



2.1.12 pav. Koreliuotųjų parametrų pavyzdys

Hipotezės apie atsitiktinius koeficientus β. Hipotezės formuluojamos ir priimamos, atsižvelgiant į p reikšmes, kaip ir fiksuoto poveikio parametrams γ.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \beta_{kj} = 0, \\ H_1\colon\ \beta_{kj} \neq 0. \end{cases}$$

2 *Sprendimo priėmimo taisyklė.* Turint statistikos p reikšmę, nulinė hipotezė atmetama (t. y. nusprendžiama, kad β_{kj} statistiškai reikšmingai skiriasi nuo 0), jei $p < \alpha$; čia α yra pasirinktasis reikšmingumo lygmuo.

Hipotezei tikrinti naudojama statistika $t = \hat{\beta}_{kj}/\sqrt{\widehat{B}_{kj}}$, kuri turi asimptotinį Stjudento skirstinį. Čia $\widehat{B}_{kj}$ yra $\hat{\beta}_{kj}$ dispersijos įvertis. Hipotezės apie atsitiktinių koeficientų reikšmes realizuotos ne visose statistinėse programose.

Nors hipotezė formuluojama parametrui β, iš tiesų tikrinama, ar atitinkami parametrai γ ir atsitiktinio poveikio dispersija τ_{kj} lygūs nuliui.

Neatmesta nulinė hipotezė yra indikatorius, kad modelio pirmojo lygmens lygtyje nereikalingas kintamasis X_k. Skyrelio pradžios pavyzdžio atveju, neatmesta nulinė hipotezė $\beta_{1j} = 0$ rodytų, kad galima spėti, jog Y ir X_1 nesusiję, ir reikia nagrinėti redukuotąjį modelį su tokia pirmojo lygmens lygtimi:

$$Y_{ij} = \beta_{0j} + e_{ij}.$$

Reikia atkreipti dėmesį, kad hipotezes apie atsitiktinių koeficientų reikšmes prasminga formuluoti tik tada, kai turima pakankamai daug antrojo lygmens elementų. Pavyzdžiui, tiriant schemą mokinys $\rightarrow$ mokykla, hipotezes apie β (mokyklos lygmens kintamuosius) beprasmiška tikrinti, jei tų mokyklų yra tik penkios.

Analogiškai fiksuotų parametrų atvejui galima formuluoti ir vektorines hipotezes, pavyzdžiui, hipotezę, kad dalis β koeficientų lygūs nuliui. Tiktai naudojama statistika $H = \hat{\boldsymbol{\beta}}'\mathbf{C}'(\mathbf{C}\widehat{\mathbf{B}}\mathbf{C}')^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$; čia $\widehat{\mathbf{B}}$ yra $\boldsymbol{\beta}$ kovariacijų matricos įvertinys, o $\mathbf{C}$ – kontrastas.

Dviejų modelių palyginimas. Tarkime, vienas modelis (toliau vadinamas *redukuotuoju*) yra kito modelio (toliau vadinamo *pilnuoju*) atskiras (dalinis) atvejis. Pilnojo modelio didžiausio tikėtinumo funkcijos logaritmo maksimumą (apskaičiuotą ML metodu) pažymėkime ℓ_p, o redukuotojo modelio – ℓ_r. Kuo didesnis skirtumas tarp ℓ_p ir ℓ_r, tuo labiau tikėtina, kad redukuotas modelis yra nepakankamas.

1 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ \text{modeliai nesiskiria}, \\ H_1\colon\ \text{modeliai skiriasi}. \end{cases}$$

2 *Sprendimo priėmimo taisyklė.* Kriterijaus statistika yra $(-2\ell_r) - (-2\ell_p)$. Jei ši reikšmė didesnė nei χ^2 skirstinio su k laisvės laipsnių α lygmens kritinė reikšmė, tai nulinę hipotezę atmetame. Čia k yra pilnojo ir redukuotojo modelių vertinamų parametrų skaičių skirtumas, o α – pasirinktasis reikšmingumo lygmuo.

Primename, kad dažniausiai $\alpha = 0{,}05$. Žinoma, jeigu turime p reikšmę, tai nulinę hipotezę atmetame, kai p reikšmė mažesnė už α. Paprastai tikrinant hipotezę visus skaičiavimus tenka atlikti patiems. Laimei, -2ℓ skaičiuoja beveik visos statistinės programos.

Neatmesta nulinė hipotezė rodo, kad pilnasis modelis nėra geresnis už redukuotąjį, t. y. papildomi kintamieji statistiškai reikšmingai modelio nepagerino.

Pastebėsime, kad HLM paprasčiausias yra besąlyginis (nulinis) modelis (žr. 1.3 skyrelį). Jį galima laikyti bet kurio kito HLM modelio redukuotuoju modeliu. Lygindami bet kurį HLM modelį su besąlyginiu, sužinome, ar jis „vertas dėmesio“. Lyginimo su besąlyginiu modeliu testas vadinamas *nulinio modelio tikėtinumo santykių testu*[11]. Galima sakyti, kad šiuo testu tikrinama, ar yra bent vienas nenulinis fiksuoto poveikio parametras γ_{ks}. Nulinio modelio tikėtinumų santykių testas yra realizuotas daugelyje statistinių programų. Jo taikymas – pirmasis HLM modelių tyrimo žingsnis. Sprendimas turint statistikos p reikšmę atrodo taip: nulinė hipotezė atmetama (t. y. nusprendžiama, kad nagrinėjamas modelis skiriasi nuo nulinio modelio), jei $p < \alpha$; čia α pasirinktasis reikšmingumo lygmuo.

1.9. HLM modelių tyrimas naudojantis SPSS programa

Pasaulyje yra tik dviejų rūšių kompiuteriai: tie, kurie gaišina jūsų brangų laiką, ir tie, kurie tai daro sparčiau.

Bendrosios pastabos. Darbą SPSS programa galima aprašyti dvejopai: nurodant meniu parinktis arba aprašant programų sintaksę. Pirmasis būdas aprašytas tinklalapyje ***http://stat.vadoveliai.lt***. Šiame skyrelyje susipažinsime su antruoju būdu.

Hierarchinių modelių analizė SPSS programa atliekama taikant procedūrą MIXED. SPSS programos kintamieji, kurie patenka į atsitiktinę jungtinio modelio dalį (dauginami iš u_{ij}), vadinami atsitiktiniais (*RANDOM*). Kintamieji, patekę į neatsitiktinę modelio dalį, vadinami fiksuotaisiais (*FIXED*). Tai nėra vienareikšmis kintamųjų priskyrimas vienam iš dviejų tipų. Tas pats kintamasis gali patekti ir į fiksuotųjų, ir į atsitiktinių kintamųjų sąrašus. Jeigu modelyje yra postūmio liekamoji paklaida u_0, tai postūmis taip pat priklauso atsitiktiniams kintamiesiems. Tarkime, kad turime tokią bendrojo modelio lygtį (dėl paprastumo dalį indeksų praleidžiame):

$$Y = \gamma_{00} + \gamma_{01}W + \gamma_{02}V + \gamma_{10}X + \gamma_{11}WX + [u_0 + Xu_1 + e].$$

Fiksuotieji kintamieji yra X, W, V. Atsitiktiniai kintamieji – postūmis ir X. Ši terminologija atsispindi ir procedūros MIXED sintaksėje. Tegul kintamasis W yra tolydus, o kintamasis V – kategorinis (pvz., nurodo respondento tautybę). Be to, tarkime, kad kintamasis *ID* norodo, kokiam antrojo lygmens elementui priklauso duomenys.

Rašydami modelį, priklausomybę konkrečiam antrojo lygmens elementui nurodome indeksais, o statistinėse programose tam skirtas atskiras – grupavimo – kintamasis.

[11] Angl. *Null Model Likelihood Ratio Test*.

Šį modelį SPSS programoje aprašanti procedūra MIXED yra tokia:

```
MIXED Y BY V WITH X W
/CRITERIA = CIN(95) MXITER(100) MXSTEP(5) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE)
LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED = X W V | SSTYPE(3)
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT X | SUBJECT(ID) COVTYPE(UN) .
```

2.1.1 lentelė. MIXED procedūros sintaksės paaiškinimai

1	MIXED Y BY V WITH X W	Procedūra – MIXED, priklausomas kintamasis – Y, po BY išvardijami kategoriniai kintamieji, vadinamieji faktoriai (V), po WITH išvardijami tolydūs kintamieji, vadinamosios kovariantės
2	/METHOD=REML	Įverčių skaičiavimo metodas (METHOD) gali būti REML arba ML. REML yra numatytasis
3	/PRINT=SOLUTION TESTCOV	SOLUTION nurodo, kad bus pateikti fiksuoto poveikio (FIXED) parametrų įverčiai. TESTCOV nurodo, kad kovariacijų kriterijai bus pateikiami kartu su atsitiktinių poveikių parametrų įverčiais
4	/FIXED= X W V | SSTYPE(3)	Po FIXED išvardijami fiksuotieji kintamieji. SSTYPE(3) – kvadratų sumų reikšmingumo testas
5	/RANDOM=INTERCEPT X | SUBJECT(ID) COVTYPE(UN)	Po RANDOM išvardijami atsitiktiniai kintamieji. Grupavimo kintamasis nurodomas skliaustuose po SUBJECT. Kovariacijų matricos pavidalą nurodo parametras COVTYPE. COVTYPE(UN) parodo, kad įvertinamos visos kintamųjų kovariacijos ir dispersijos

Sintaksės paaiškinimai pateikiami 2.1.1 lentelėje.

Užrašas /CRITERIA = CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) nurodo, kokie konvergavimo kriterijai bus naudojami skaičiuojant parametrų įverčius. Parametrų įverčiai skaičiuojami žingsnis po žingsnio juos tikslinant. Liaujamasi, kai patikslintas įvertis nuo prieš tai buvusiojo skiriasi mažiau nei nurodyta reikšmė. Šie parametrai SPSS programoje yra numatytieji. Pavyzdžiui, nurodyta, kad tikslinimų (iteracijų) bus iki 100. Ši sintaksės dalis keičiama retai, todėl ją aprašėme ne lentelėje, o atskirai. Toliau aprašydami procedūras jos nebenurodysime.

Parodysime, kaip SPSS programa atliekama analizė pasirinkus 1.5 skyrelyje suformuluotus modelius. Primename, kad buvo tirti aštuntokų matematikos pasiekimai ir vartoti tokie žymenys:

CSES – mokyklos vidurkiu centruotas mokinio socialinis ir ekonominis statusas;
MSES – mokyklos socialinio ir ekonominio statuso vidurkis;
IDMOK – kintamasis, kuriuo nurodomas mokyklos indeksas;
VK – mokyklos tipas (0 – Vilniaus mokykla, 1 – kaimo mokykla);
MAT – matematikos testo rezultatai.

Indeksus praleidome, nes naudojant SPSS programos sintaksę jie nėra svarbūs. Kita vertus, svarbus grupavimo kintamasis *IDMOK*, pagal kurį nustatome, kuriai mokyklai priklausė tiriamasis. Visuose pavyzdžiuose tikrindami hipotezes tariame, kad reikšmingumo lygmuo lygus 0,05.

Besąlyginis modelis. Visų pirma parodysime, kaip SPSS programa tiriamas besąlyginis (nulinis) modelis. Jis išsamiai aptartas 1.3 skyrelyje, todėl čia apsiribosime tik svarbiausiais tyrimo momentais.

Mokinio lygmuo:

$$MAT = \beta_0 + e.$$

Mokyklos lygmuo:

$$\beta_0 = \gamma_{00} + u_0.$$

Jungtinis modelis:

$$MAT = \gamma_{00} + u_0 + e.$$

Parametras γ_{00} nusako visų mokyklų matematikos testo rezultatų vidurkį. Mokyklų skirtumus rodo kintamojo u_0 dispersija τ_{00}, mokinių individualius skirtumus – kintamojo e dispersija σ^2. Kuo dispersijos didesnės, tuo didesni ir atitinkami skirtumai.

SPSS programos MIXED procedūra yra tokia:

```
MIXED MAT
/CRITERIA = ...
/FIXED = | SSTYPE(3)
/METHOD = REML /PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT | SUBJECT(IDMOK) COVTYPE(UN) .
```

Aptarsime gautus rezultatus (žr. 2.1.13 pav). Pirmojoje lentelėje pateiktos informacinių indeksų reikšmės. SPSS programoje mažesnė indekso reikšmė rodo geresnį modelio tinkamumą. Reikšmės mažumas – reliatyvi sąvoka. Todėl informaciniai indeksai naudojami dviem modeliams palyginti. Modelis, kurio informacinių indeksų reikšmės mažesnės, geriau tinka duomenims. Nagrinėjame nulinį modelį, todėl kol kas indeksų reikšmių nėra su kuo lyginti.

Antrojoje ir trečiojoje lentelėse pateiktos statistinės hipotezės H_0: $\gamma = 0$, H_1: $\gamma \neq 0$ p reikšmės (žr. 2.1.13 pav.). Nors trečiojoje lentelėje naudojama Voldo statistika, turinti Stjudento skirstinį, o antrojoje remiamasi Fišerio skirstiniu, abi p reikšmės identiškos, nes viena statistika išsireiškia per kitą ($F = t^2$). Todėl kituose pavyzdžiuose antrosios lentelės nebeaptarinėsime. Greta paties įverčio $\hat{\gamma}_{00} = 499{,}47$ pateiktas jo standartinis nuokrypis 9,88 (visur atsakymus apvaliname iki dviejų ženklų po kablelio). Statistika $t = 499{,}47/9{,}88 = 50{,}55$. Kadangi p reikšmė labai maža (mažesnė už 0,01), tai darome išvadą, kad statistiškai reikšmingai $\gamma_{00} \neq 0$. Tai ne itin vertinga išvada, nes ji tereiškia,

Information Criteria[a]

-2 Rectricted Log Likelihood	6415,181
Akaike's Information Criterion (AIC)	6419,181
Hurvich and Tsai's Criterion (AICC)	6419,202
Bozdogan's Criterion (CAIC)	6429,829
Schwarz's Bayesian Criterion (BIC)	6427,829

The information criteria are displayed in smaller-is-better forms.
a. Dependent Variable: MAT.

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	25,493	2557,424	,000

a. Dependent Variable: MAT.

$F = t^2$

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Intercept	499,4717	9,8766460	25,493	50,571	,000	479,1503243	519,7931040

a. Dependent Variable: MAT.

$\hat{\gamma}_{00}$ p reikšmė Pasikliautinieji intervalai

Estimates of Covariance Parameters[a]

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
					Lower Bound	Uper Bound
Residual	5117,191	313,8830	16,303	,000	4537,533760	5770,898707
Intercept [subject Variance	2356,107	736,8559	3,198	,001	1276,402556	4349,129239

a. Dependent Variable: MAT.

$\hat{\tau}_{00}$ p reikšmės

2.1.13 pav. SPSS rezultatai taikant besąlyginį modelį

jog vidutiniai mokinių testo pasiekimai nėra lygūs nuliui. Lentelėje yra ir $\hat{\gamma}_{00}$ 95 % pasikliautinasis intervalas: [479,15; 519,79].

Atsitiktinių paklaidų dispersijų įverčiai ir tų įverčių 95 % pasikliautinieji intervalai pateikiami paskutinėje lentelėje. Mokinių individualumą apibūdinančios dispersijos įvertis $\hat{\sigma}^2 = 5117{,}19$, jos pasikliautinasis intervalas yra [4537,53; 5770,90]. Maža naudojamo kriterijaus p reikšmė ($p < 0{,}05$) leidžia teigti, kad $\hat{\sigma}^2 > 0$. Gavome, kad dėl tam tikrų individualių mokinių skirtumų jų rezultatai skiriasi. Mūsų tikslas – išsiaiškinti, nuo ko priklauso mokinio matematikos testo rezultatai. Todėl didelė $\hat{\sigma}^2$ reikšmė – nėra gerai. Konstatavome faktą, kad skiriasi, bet nežinome, kodėl. Panašiai interpretuojama ir didelė $\hat{\tau}_{00} = 2356{,}11$ reikšmė (pasikliautinasis intervalas [1276,40; 4349,13], $p < 0{,}01$)

– mokyklų įtaka statistiškai reikšmingai atsispindi mokinių rezultatuose, bet kas tai lemia, nežinome.

Atsitiktinio postūmio ir posvyrio modelio analizė. Nagrinėsime antrąjį 1.5 skyrelio modelį, t. y. tarsime, kad kintamojo *MAT* reikšmė priklauso nuo jo socialinio ir ekonominio statuso ir ta priklausomybė skirtingose mokyklose nevienoda. Taigi bandysime išsiaiškinti, ar mokinio testo rezultatas priklauso nuo jo socialinio ir ekonominio statuso.

Mokinio lygmuo:

$$MAT = \beta_0 + \beta_1 CSES + e.$$

Mokyklos lygmuo:

$$\begin{cases} \beta_0 = \gamma_{00} + u_0, \\ \beta_1 = \gamma_{10} + u_1. \end{cases}$$

Jungtinė lygtis:

$$MAT = \gamma_{00} + \gamma_{10} CSES + [u_1 CSES + u_0 + e].$$

Fiksuotasis kintamasis yra *CSES*, atsitiktiniai kintamieji – postūmis ir *CSES*. Primename, kad $\mathbf{D}(u_0) = \tau_{00}$, $\mathbf{D}(u_1) = \tau_{11}$, $\mathbb{C}\mathrm{ov}(u_0, u_1) = \mathbb{C}\mathrm{ov}(\beta_1, \beta_2) = \tau_{10} = \tau_{01}$, $\mathbf{D}e = \sigma^2$.

SPSS programos MIXED procedūros sintaksė šiuo atveju yra tokia:

```
MIXED mat WITH CSES
/CRITERIA = ...
/FIXED = CSES | SSTYPE(3)
/METHOD = REML /PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT CSES | SUBJECT(IDMOK) COVTYPE(UN) .
```

Gauti rezultatai pateikti 2.1.14 paveiksle. Palyginti su besąlyginiu modeliu, visi informaciniai indeksai sumažėjo, t. y. naujasis modelis geriau tinka turimiems duomenims. Postūmio įvertis pasikeitė labai mažai – $\hat{\gamma}_{00} = 502{,}29$. Fiksuoto parametro γ_{10} įvertis $\hat{\gamma}_{10} = 7{,}93$. Taigi norėdami šį modelį taikyti, pavyzdžiui, mokinio matematiniams rezultatams prognozuoti, turėtume naudotis lygtimi

$$MAT = 502{,}29 + 7{,}93 CSES.$$

Ši lygtis parodo, kad kiekvienas *SES* balas, viršijantis vidutinę *SES* reikšmę, vidutiniškai padidina matematikos testo rezultatus 7,93 taško.

Kaip matyti iš trečiosios lentelės, $\hat{\sigma}^2 = 4294{,}72$. Šis įvertis statistiškai reikšmingai didesnis už nulį. Todėl išlieka ankstesnio pavyzdžio išvada, kad mokinių rezultatai statistiškai reikšmingai skiriasi dėl tam tikrų jų individualių savybių. Įtraukdami į modelį socialinį ir ekonominį mokinio statusą, bandėme dalį individualių mokinių skirtumų paaiškinti būtent šiuo veiksniu.

Dispersijos įverčio $\hat{\sigma}^2$ sumažėjimas, palyginti su ankstesniu modeliu, parodo, ar labai naudinga buvo įtraukti naują kintamąjį.

Besąlyginio modelio mokinio lygmens paklaidos įvertis buvo $\hat{\sigma}^2 = 5117{,}19$, todėl įvertis sumažėjo

$$\frac{5117{,}19 - 4294{,}72}{5117{,}19} = 0{,}16.$$

Information Criteria[a]

-2 Rectricted Log Likelihood	6415,056
Akaike's Information Criterion (AIC)	6153,056
Hurvich and Tsai's Criterion (AICC)	6153,131
Bozdogan's Criterion (CAIC)	6174,230
Schwarz's Bayesian Criterion (BIC)	6170,230

The information criteria are displayed in smaller-is-better forms.
a. Dependent Variable: MAT.

Estimates of Fixed Effects[a]

	Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
							Lower Bound	Uper Bound
$\hat{\gamma}_{00}$	Intercept	502,2935	9,906121	25,427	50,705	,000	481,908848	522,678184
$\hat{\gamma}_{10}$	CSES	7,925654	1,220362	21,223	6,495	,000	5,389396	10,461912

a. Dependent Variable: MAT.

Estimates of Covariance Parameters[a]

	Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
							Lower Bound	Uper Bound
$\hat{\sigma}^2$	Residual		4294,718	274,4819	15,647	,000	3789,073455	4867,838836
$\hat{\tau}_{00}$	Intercept + CSES	UN (1,1)	2407,353	742,4251	3,243	,001	1315,315740	4406,049724
$\hat{\tau}_{01}$	[subject = IDMOK]	UN (2,1)	46,955524	73,113748	,642	,521	-96,344789	190,255837
$\hat{\tau}_{11}$		UN (2,2)	8,883918	10,975893	,809	,418	,788793	100,056666

a. Dependent Variable: MAT.

2.1.14 pav. SPSS rezultatai taikant atsitiktinio postūmio ir posvyrio modelį

Taigi galima sakyti, kad įtraukus į modelį *CSES* mokinio lygmens atsitiktinės paklaidos dispersija sumažėjo 16 %. Dažniausiai tai interpretuojama taip: palyginti su besąlyginiu modeliu, nepaaiškintų mokinių rezultatų skirtumų liko 16 % mažiau.

Analogiškai interpretuojamas ir $\hat{\tau}_{00} = 2407{,}353$ pokytis. Palyginti su besąlyginiu modeliu, šis įvertis net šiek tiek išaugo. Vis dėlto tas įverčio padidėjimas yra labai mažas (apie 2 %). Todėl darytume išvadą, kad šis modelis geresnis už besąlyginį. Matome, kad statistinė hipotezė $\tau_{00} = 0$ atmetama, o hipotezės $\tau_{11} = 0$ ir $\tau_{10} = 0$ neatmetamos. Neatmesta hipotezė $\tau_{11} = 0$ yra indikatorius, kad galbūt testo rezultatų priklausomybė nuo *CSES* visose mokyklose yra tokia pat. Neatmesta hipotezė $\tau_{10} = 0$ tiesiog parodo, kad β_0 ir β_1 nekoreliuoja, t. y. tarp pastovaus mokyklos „priedo“ ir socialinio bei ekonominio statuso koreliacijos nėra. Pavyzdžiui, negalima tikėtis, kad mokyklose, kuriose rezultatų vidurkis β_0 didesnis, rezultatų priklausomybė nuo *CSES* taip pat stipresnė (β_1 didesnis).

Kaip ir besąlyginio modelio atveju, lentelėse pateikti parametrų įverčių pasikliautinieji 95 % intervalai.

Modelio su antrojo lygmens intervaliniu kintamuoju analizė. Į modelį įtrauksime antrojo lygmens kintamąjį *MSES*, kuris paaiškins dalį mokyklų skirtumų.

Mokinio lygmuo:

$$MAT = \beta_0 + \beta_1 CSES + e.$$

Mokyklos lygmuo:

$$\begin{cases} \beta_0 = \gamma_{00} + \gamma_{01} MSES + u_0, \\ \beta_1 = \gamma_{10} + u_1. \end{cases}$$

Jungtinė lygtis:

$$MAT = \gamma_{00} + \gamma_{01} MSES + \gamma_{10} CSES + [u_1 CSES + u_0 + e].$$

Fiksuotieji kintamieji yra *CSES*, *MSES*, atsitiktiniai kintamieji – postūmis ir *CSES*. SPSS programos MIXED procedūros sintaksė šiuo atveju yra tokia:

```
MIXED mat WITH CSES MSES
/CRITERIA =. . .
/FIXED = CSES MSES | SSTYPE(3)
/METHOD = REML /PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT CSES | SUBJECT(IDMOK) COVTYPE(UN) .
```

Gauti rezultatai pateikti 2.1.15 paveiksle. Neparodyti informaciniai indeksai, kurie, palyginti su prieš tai buvusio modelio, truputį sumažėjo, pavyzdžiui, $AIC = 6117{,}83$. Tai leidžia tikėtis, kad šio pavyzdžio modelis duomenims tinka geriau. Postūmio parametro įvertis pasikeitė radikaliai ($\hat{\gamma}_{00} = 175{,}94$). Taip atsitiko todėl, kad dabar β_0 elgesį modeliuojame atsižvelgdami ir į *MSES* elgesį ($\hat{\gamma}_{01} = 16{,}56$).

Modelį papildėme mokyklos lygmens kintamuoju *MSES*. Ankstesniuose pavyzdžiuose buvo taip: mokykla daro įtaką, bet nuo ko ji priklauso – neaišku. Dabar bandome paaiškinti tą poveikį mokyklos socialinio ir ekonominio statuso vidurkiu. Tikimės, kad mokyklose, kur susirenka aukštesniojo socialinio ir ekonominio sluoksnio mokiniai, yra atitinkama atmosfera, skatinanti (ar trukdanti) siekti žinių. Gali būti, kad bandydami interpretuoti priklausomybes ir klystame. Vis dėlto, jeigu yra kokie nors mokyklos lygmens

Estimates of Fixed Effects[a]

	Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
							Lower Bound	Uper Bound
$\hat{\gamma}_{00}$	Intercept	175,9439	42,631950	24,688	4,127	,000	88,085443	263,802297
$\hat{\gamma}_{10}$	CSES	8,044992	1,227057	21,239	6,556	,000	5,494931	10,595052
$\hat{\gamma}_{01}$	MESE	16,561816	2,135069	24,060	7,757	,000	12,155834	20,967797

a. Dependent Variable: MAT.

Estimates of Covariance Parameters[a]

	Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
							Lower Bound	Uper Bound
$\hat{\sigma}^2$	Residual		4290,921	274,0261	15,659	,000	3786,092741	4863,062231
$\hat{\tau}_{00}$	Intercept + CSES	UN (1,1)	559,4769	227,0540	2,464	,014	252,543766	1239,446005
$\hat{\tau}_{01}$	[subject = IDMOK]	UN (2,1)	,371530	37,426627	,010	,992	-72,983311	73,726370
$\hat{\tau}_{11}$		UN (2,2)	9,191156	11,104040	,828	,408	,861023	98,112804

a. Dependent Variable: MAT.

2.1.15 pav. SPSS rezultatai taikant modelį su antrojo lygmens intervaliniu kintamuoju

veiksniai, turintys įtakos mokinio rezultatams ir *susiję su MSES*, jie atsispindės naujojo modelio $\hat{\tau}_{00}$ įvertyje. Logika čia tokia – visi modeliu nepaaiškinti mokinių skirtumai, atsirandantys dėl mokyklos įtakos, atsispindi atsitiktinėse paklaidose u_k. Kuo tų paklaidų dispersijos τ_{kk} didesnės, tuo daugiau lieka nepaaiškintų skirtumų.

Sėkmingai parinktas kintamasis paklaidų dispersijas turi sumažinti.

MSES įtraukėme tik į postūmio β_0 lygtį, todėl didžiausią dėmesį skiriame atitinkamai paklaidai τ_{00}. Jeigu būtume *MSES* įtraukę ir į β_1, tai turėtume pažiūrėti, kaip pasikeitė ir τ_{11}. Taigi, palyginti su prieš tai nagrinėtu modeliu, įvertis $\hat{\tau}_{00}$ sumažėjo nuo 2407,35 iki 559,48, t. y. 76,7 % ((2407,35 − 559,48)/2407,35 = 0,767...). Tai labai didelis pokytis, liudijantis, kad naujasis modelis daug geriau paaiškina mokyklos lygmens įtaką. Kaip ir anksteniame modelyje, įsitikiname, kad τ_{11} ir τ_{10} statistiškai reikšmingai nesiskiria nuo 0. Paiskliautinųjų intervalų pokyčiai buvo neesminiai.

Nors svarbiausias trečiojo modelio tikslas buvo geriau paaiškinti mokyklos įtaką mokiniams, tačiau $\hat{\sigma}^2 = 4290{,}92$ irgi truputį mažesnis nei antrojo pavyzdžio. Taigi darome vienareikšmę išvadą, kad šis modelis duomenims tinka geriau nei ankstesnis.

Modelio su antrojo lygmens kategoriniu kintamuoju analizė. Įtrauksime į modelį antrojo lygmens pseudokintamąjį *VK*, nurodantį, kad ta mokykla Vilniaus miesto.

Mokinio lygmuo:

$$MAT = \beta_0 + \beta_1 CSES + e.$$

Mokyklos lygmuo:

$$\begin{cases} \beta_0 = \gamma_{00} + \gamma_{01} MSES + \gamma_{02} VK + u_0, \\ \beta_1 = \gamma_{10} + \gamma_{12} VK + u_1. \end{cases}$$

Jungtinė lygtis:

$$\begin{aligned} MAT = {} & \gamma_{00} + \gamma_{01} MSES + \gamma_{10} CSES + \gamma_{02} VK + \gamma_{12} VK \cdot CSES \\ & + [u_1 CSES + u_0 + e]. \end{aligned} \tag{2.1.26}$$

Fiksuotieji kintamieji yra *CSES*, *MSES*, *VK* ir $VK \cdot CSES$, atsitiktiniai kintamieji – postūmis ir *CSES*. Kintamasis *VK* kategorinis, todėl SPSS programos MIXED procedūros sintaksė yra tokia:

```
MIXED mat BY VK WITH MSES CSES
/CRITERIA = ...
/FIXED = VK MSES CSES CSES*VK | SSTYPE(3)
/METHOD = REML /PRINT = CORB SOLUTION TESTCOV
/RANDOM INTERCEPT CSES | SUBJECT(IDMOK) COVTYPE(UN) .
```

Gauti rezultatai pateikti 2.1.16 paveiksle. Neparodyti informaciniai indeksai, kurie palyginti su prieš tai buvusiu modeliu, truputį sumažėjo. Pavyzdžiui, $AIC = 6099{,}13$. Fiksuoto poveikio parametrų įverčiai leidžia sudaryti testo rezultatų priklausomybės lygtis Vilniaus ($VK = 0$) ir kaimo ($VK = 1$) mokykloms. Paaiškinsime, kaip tai daroma.

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Intercept	-,696504	75,392268	21,749	-,009	,993	-157,155272	155,762264
[vk=,00]	47,929499	17,504921	20,264	2,738	,013	11,445297	84,413702
[vk=1,00]	0[a]	0	.	.	.	.	.
MSES	23,931960	3,316798	21,875	7,215	,000	17.051064	30,812855
CSES	9,341851	2,073059	19,300	4,506	,000	5,007449	13,673252
CSES([vk=,00])	-2,071545	2,593525	19,947	-,799	,434	-7,482466	3,339376
CSES([vk=1,00])	0[a]	0	.	.	.	.	.

a. This parameter is set to zero because it is redundant.
b. Dependent Variable: MAT.

Estimates of Covariance Parameters[a]

	Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
$\hat{\sigma}^2$	Residual	4293,985	274,5003	15,643	,000	3788,314166	4867,152997
$\hat{\tau}_{00}$	Intercept + CSES UN (1,1)	381,2217	184,6343	2,065	,039	147,543928	984,994491
$\hat{\tau}_{01}$	[subject = IDMOK] UN (2,1)	9,121615	36,932427	,247	,805	-63,264612	81,507842
$\hat{\tau}_{11}$	UN (2,2)	10,277841	11,706588	,878	,380	1,102485	95,814436

a. Dependent Variable: MAT.

2.1.16 pav. SPSS rezultatai taikant modelį su antrojo lygmens kategoriniu kintamuoju

Pirmojoje lentelėje nurodyti įverčiai $\hat{\gamma}_{00} = -0{,}69$, $\hat{\gamma}_{10} = 23{,}93$ ir $\hat{\gamma}_{01} = 9{,}34$. Juos įstatę į jungtinę lygtį, gauname

$$MAT = -0{,}69 + 23{,}93MSES + 9{,}34CSES + \gamma_{02}VK + \gamma_{12}VK \cdot CSES.$$

Likę lentelės koeficientai nurodo, ką reikia įstatyti vietoj $\hat{\gamma}_{02}$ ir $\hat{\gamma}_{12}$, esant atitinkamai *VK* reikšmei. Kai $VK = 1$, tai abudu koeficientai yra lygūs 0. Taigi kaimo mokyklų mokinių matematikos rezultatus aprašo tokia lygtis:

$$MAT_{\text{Kaimas}} = -0{,}69 + 23{,}93MSES + 9{,}34CSES.$$

Iš jos matyti, kad kiekvienas mokinio *SES* balas, viršijantis mokyklos vidurkį, prie mokinio rezultato prideda 9,34 balo. Tuo tarpu vienu balu padidėjus *MSES*, *MAT* išauga iki 23,93 balo. Analogiškai Vilniaus mokyklų ($VK = 0$) mokinių pasiekimus aprašo tokia lygtis:

$$\begin{aligned} MAT_{\text{Vilnius}} &= -0{,}69 + 23{,}93MSES + 9{,}34CSES + 47{,}92 - 2{,}07CSES \\ &= 47{,}23 + 23{,}93MSES + 7{,}27CSES. \end{aligned}$$

Patikrinę Voldo statistinių hipotezių p reikšmes, matome, kad ne visos statistinės hipotezės atmetamos. Neatmesta nulinė hipotezė H_0: $\gamma_{00} = 0$ leidžia įtarti, kad vidutinius mokyklų rezultatų skirtumus lemia mokyklos vieta. Matome, kad prie Vilniaus mokyklų mokinių rezultatų pridedamas mokyklos „priedas“ 47,233 balo. Neatmesta hipotezė H_0: $\gamma_{12} = 0$ rodo, kad mokyklos vietos ir socialinio bei ekonominio mokinio statuso sąveika yra statistiškai nereikšminga. Taigi nors sociologinio ir ekonominio statuso įtaka rezultatams Vilniaus mieste silpnesnė – daugiklis prie *CSES* (7,27) yra mažesnis nei kaimo mokyklose (9,34), tačiau tai dar neatskleidžia tikrosios padėties. Visiškai įmanoma, kad taip yra dėl imties atsitiktinumo – juk tiriame tik dalį mokyklų.

Įvedėme papildomą antrojo lygmens kintamąjį *VK*, t. y. bandėme tiksliau nusakyti, kas lemia mokyklos įtaką. Nepaaiškintų veiksnių turėjo likti mažiau, todėl galime tikėtis, kad sumažės atitinkami atsitiktinių koeficientų β_{01} ir β_1 dispersijų įverčiai. Iš tikrųjų iš antros lentelės matyti, kad postūmio dispersijos įvertis $\hat{\tau}_{00} = 381{,}22$, palyginti su modeliu, į kurį įtrauktas antrojo lygmens intervalinis kintamasis ($\hat{\tau}_{00} = 559{,}47$), sumažėjo beveik 32 %. Taip pat sutrumpėjo ir įverčio pasikliautinasis intervalas. Daugiklio prie *CSES* dispersijos įvertis $\hat{\tau}_{11} = 10{,}27$ truputį padidėjo (buvo 9,19), tačiau, kaip ir prieš tai nagrinėtame pavyzdyje, statistinė hipotezė dėl jo lygybės nuliui neatmetama. Neatmetama ir hipotezė dėl koeficientų nekoreliuotumo (H_0: $\tau_{10} = 0$). Mokinio lygmens atsitiktinės paklaidos įvertis $\hat{\sigma}^2 = 4293{,}98$, palyginti su ankstesniu pavyzdžiu ($\hat{\sigma}^2 = 4290{,}92$), beveik nepasikeitė (padidėjo 0,07 procento).

Apibendrindami galime konstatuoti, kad šis modelis geriau tinka duomenims, nei ankstesnieji modeliai, tačiau daug neatmestų hipotezių dėl parametrų lygybės nuliui reikalauja papildomo tyrimo.

Visų keturių modelių kai kurios charakteristikos pateikiamos 2.1.2 lentelėje. Smulkiai neaptarinėsime kitų galimų modelių. Tik norime atkreipti dėmesį, kad iš modelio su ka-

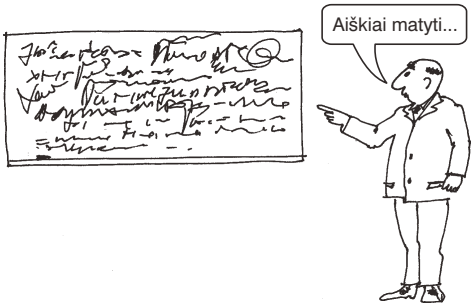
2.1.2 lentelė. Modelių palyginimas

	Modelis	Informaciniai indeksai	Fiksuoto poveikio parametrai	Atsitiktinio poveikio parametrai
1	Besąlyginis: $MAT = \beta_0 + e$ $\beta_0 = \gamma_{00} + u_0$	$-2LL = 6425{,}18$ $AIC = 6419{,}18$ $BIC = 6427{,}83$	$\hat{\gamma}_{00} = 499{,}5^{**}$	$\hat{\sigma}^2 = 5117{,}2^{**}$ $\hat{\tau}_{00} = 2356{,}1^{**}$
2	Atsitiktinio posvyrio ir postūmio: $MAT = \beta_0 + \beta_1 CSES + e$ $\beta_0 = \gamma_{00} + u_0$ $\beta_1 = \gamma_{10} + u_1$	$-2LL = 6145{,}1$ $AIC = 6153{,}1$ $BIC = 6170{,}2$	$\hat{\gamma}_{00} = 502{,}3^{**}$ $\hat{\gamma}_{10} = 7{,}9^{**}$	$\hat{\sigma}^2 = 4294{,}7^{**}$ $\hat{\tau}_{00} = 2407{,}4^{**}$ $\hat{\tau}_{01} = 46{,}9$ $\hat{\tau}_{11} = 8{,}9$
3	Su antrojo lygmens intervaliniu kintamuoju: $MAT = \beta_0 + \beta_1 CSES + e$ $\beta_0 = \gamma_{00} + \gamma_{01} MSES + u_0$ $\beta_1 = \gamma_{10} + u_1$	$-2LL = 6109{,}8$ $AIC = 6117{,}8$ $BIC = 6134{,}9$	$\hat{\gamma}_{00} = 175{,}9^{**}$ $\hat{\gamma}_{10} = 8{,}0^{**}$ $\hat{\gamma}_{01} = 16{,}6^{**}$	$\hat{\sigma}^2 = 4290{,}9^{**}$ $\hat{\tau}_{00} = 559{,}5^{*}$ $\hat{\tau}_{01} = 0{,}4$ $\hat{\tau}_{11} = 9{,}2$
4	Su antrojo lygmens kategoriniu kintamuoju: $MAT = \beta_0 + \beta_1 CSES + e$ $\beta_0 = \gamma_{00} + \gamma_{01} MSES + \gamma_{02} VK + u_0$ $\beta_1 = \gamma_{10} + \gamma_{11} VK + u_1$	$-2LL = 6091{,}1$ $AIC = 6099{,}1$ $BIC = 6116{,}3$	$\hat{\gamma}_{00} = -0.7$ $[VK = 0]$ $\hat{\gamma}_{02} = 47{,}9^{*}$ $[VK = 1]$ $\hat{\gamma}_{02} = 0$ $\hat{\gamma}_{01} = 23{,}9^{**}$ $\hat{\gamma}_{10} = 9{,}3^{**}$ $[VK = 0]$ $\hat{\gamma}_{11} = -2{,}1$ $[VK = 1]$ $\hat{\gamma}_{11} = 0$	$\hat{\sigma}^2 = 4293{,}9^{**}$ $\hat{\tau}_{00} = 381{,}2^{*}$ $\hat{\tau}_{10} = 9{,}1$ $\hat{\tau}_{11} = 10{,}3$

$^{*}p < 0{,}05$. $^{**}p < 0{,}01$.

tegoriniu antrojo lygmens kintamuoju pašalinus kintamųjų sąveiką $VK \cdot CSES$ modelio parametrai mažai keičiasi ($\hat{\tau}_{00} = 381$, $\hat{\sigma}^2 = 4356$), nors lieka neatmestos statistinės hipotezės H_0: $\gamma_{00} = 0$, H_0: $\tau_{11} = 0$ ir H_0: $\tau_{10} = 0$. Bet koks bandymas pakeisti atsitiktinius koeficientus neatsitiktiniais smarkiai pablogina parametrų dispersijų įverčius. Pavyzdžiui, tarus, kad $\tau_{11} = 0$ (t. y. β_1 neatsitiktinis), gaunama $\hat{\tau}_{00} = 555$. Todėl minėtąjį modelį galima laikyti geriausiu iš hierarchinių modelių su kintamaisiais *CSES*, *VK* ir *MSES*. Tai nereiškia, kad nėra kitų geresnių modelių, bet juose turėtų būti kitų kintamųjų.

Hierarchinių modelių tikslas – geriau paaiškinti spėjamas kintamųjų priklausomybes. Pateiksime porą pavyzdžių, iš kurių matyti, kad atsižvelgiant vien į skaitines modelio charakteristikas dar nebūtinai gaunami prasmingi modeliai.



Modelio su stipriai koreliuojančiais kintamaisiais analizė. Nagrinėsime atsitiktinio postūmio ir posvyrio modelį, tik su kitu mokinio lygmens kintamuoju – jo fizikos testo rezultatu *FIZ*.

Mokinio lygmuo:

$$MAT = \beta_0 + \beta_1 FIZ + e.$$

Mokyklos lygmuo:

$$\begin{cases} \beta_0 = \gamma_{00} + u_0, \\ \beta_1 = \gamma_{10} + u_1. \end{cases}$$

Jungtinė lygtis:

$$MAT = \gamma_{00} + \gamma_{10} FIZ + [u_1 FIZ + u_0 + e].$$

Fiksuotas kintamasis yra *FIZ*, atsitiktiniai kintamieji – posvyris ir *FIZ*. SPSS programos MIXED procedūros sintaksė yra tokia:

```
MIXED mat WITH FIZ
/CRITERIA = ...
/FIXED = FIZ | SSTYPE(3)
/METHOD = REML /PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT FIZ | SUBJECT(IDMOK) COVTYPE(UN) .
```

Gauti rezultatai pavaizduoti 2.1.17 paveiksle. Matome, kad fiksuoto poveikio parametras statistiškai reikšmingai skiriasi nuo nulio. Mokinio lygmens atsitiktinės paklaidos dispersijos įvertis $\hat{\sigma}^2 = 3552{,}42$ yra mažesnis nei ankstesniuose modeliuose (mažiausias jų buvo $\hat{\sigma}^2 = 4290{,}92$). *Hipotezė, kad atsitiktinis parametras lygus nuliui, nėkart nebuvo*

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Intercept	134,8471	25,26785	20,689	5,337	,000	82,251516	187,442632
FIZ	,699102	,047010	21,374	14,871	,000	,60144	,796760

a. Dependent Variable: MAT.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Residual		3552,420	223,4042	15,901	,000	3140,465454	4018,413079
Intercept +	UN (1,1)	3041,955	4696,271	,648	,517	147,58076	62699,77649
FIZ [subject	UN (2,1)	-4,211332	8,484440	-,496	,620	-20,840529	12,417865
= IDMOK]	UN (2,2)	,008326	,015879	,524	,600	,000198	,349806

a. Dependent Variable: MAT.

2.1.17 pav. SPSS rezultatai taikant modelį su stipriai koreliuojančiais kintamaisiais

atmesta. Tai indikatorius, kad antrojo lygmens kintamųjų nebereikia. Iš tikro, padarę prielaidą, kad visi atsitiktinių poveikių parametrai lygūs nuliui, gauname tiesinės regresijos lygtį $MAT = \gamma_{00} + \gamma_{10}FIZ + e$. Tolesni tyrimai parodė, kad tiesinė regresija šiuo atveju puikiai tinka duomenims. Taip yra dėl to, kad matematikos testo rezultatai gana stipriai koreliuoja su fizikos testo rezultatais (koreliacija lygi 0,627).

Atrodytų viskas gerai, suradome paprastą, duomenimis gerai tinkamą modelį. Tačiau prisiminkime, kad tyrimo tikslas – nustatyti, kas turi įtakos matematikos testo rezultatui. Gauta išvada – geriau mokantys fiziką geriau moka ir matematiką – ne itin naudinga. Kažin ar galima įtarti, jog mokinys, nusprendęs nebesimokyti fizikos, pradės blogiau suprasti matematiką[12].

Išvada. Stipriai su rezultatu koreliuojantis pirmojo lygmens kintamasis modelį pagerina ne visada.

Koreliuojančiųjų kintamųjų įtaka hierarchinei struktūrai. Parodysime, kaip įtrauktas kintamasis *FIZ* gali pakeisti hierarchinį modelį.

Mokinio lygmuo:

$$MAT = \beta_0 + \beta_1 CSES + \beta_2 FIZ + e.$$

Mokyklos lygmuo:

$$\begin{cases} \beta_0 = \gamma_{00} + \gamma_{01} MSES + \gamma_{02} VK + u_0, \\ \beta_1 = \gamma_{10} + u_1, \\ \beta_2 = \gamma_{20} + u_2. \end{cases}$$

Jungtinė lygtis:

$$\begin{aligned} MAT = {} & \gamma_{00} + \gamma_{01} MSES + \gamma_{10} CSES + \gamma_{20} FIZ + \gamma_{02} VK \\ & + [u_1 CSES + u_2 FIZ + u_0 + e]. \end{aligned} \tag{2.1.27}$$

[12] O štai be matematikos fiziką suprasti nelengva.

Fiksuotieji kintamieji yra *CSES*, *MSES*, *VK* ir *FIZ*, atsitiktiniai kintamieji – posvyris, *CSES* ir *FIZ*. SPSS programos MIXED procedūros sintaksė yra tokia:

```
MIXED mat BY VK WITH FIZ MSES CSES
/CRITERIA =. . .
/FIXED = FIZ VK MSES CSES | SSTYPE(3)
/METHOD = REML /PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT FIZ CSES | SUBJECT(IDMOK) COVTYPE(UN) .
```

Gauti rezultatai pavaizduoti 2.1.18 paveiksle. Iš antrosios lentelės matyti, kad mokyklos lygmens kintamųjų dispersijų įverčių nebeliko – liko tik kovariacijų įverčiai. Atrodytų, kad modelis idealiai aprašo mokyklos įtaką. Tačiau iš pirmos lentelės matyti, jog beveik 100 % galima garantuoti, kad jokie kintamieji modelyje nereikalingi. Taigi modelis tampa visiškai nelogiškas.

Įtraukdami į modelį stipriai su *MAT* koreliuojantį, bet abejotinais priežastiniais ryšiais susijusį kintamąjį, galime modelį visiškai sugadinti.
Neatmestos hipotezės apie visų antrojo lygmens dispersijų lygybę nuliui – indikatorius, kad pirmajame lygmenyje yra stipriai su priklausomu kintamuoju koreliuojantis dydis.

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Intercept	-123,778	57,227695	,000	-2,163	1,000	-8783482,88	8783235,326
FIZ	,596091	,044218	,000	13,481	1,000	-8481,148723	8482,340904
[vk=,00]	23,033509	13,361909	,000	1,737	1,000	-2234856,56	2234902,631
[vk=1,00]	0[a]	0	.	.	.	.	.
MSES	15,200273	2,577352	,000	5,898	1,000	-151137,614	151168,0149
CSES	4,46011	,984773	,001	4,533	,995	-169114,689	169123,6170

a. This parameter is set to zero because it is redundant.
b. Dependent Variable: MAT.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Residual		3271,406	204,9526	15,962	,000	2893,390014	3698,809935
Intercept + FIZ	UN (1,1)	,000000[a]	,000000	.	.	.	.
+ CSES	UN (2,1)	-2,290289	,094568	-24,218	,000	-2,475638	-2,104939
[subject =	UN (2,2)	,000665[a]	,000000	.	.	.	.
IDMOK]	UN (3,1)	26,153538	101,6642	,257	,797	-173,104695	225,411771
	UN (3,2)	-,026010	,183579	-,142	,887	-,385819	,333799
	UN (3,3)	1,185540[a]	,000000	.	.	.	.

a. This covariance parmameter is redundant. The test statistic and confidance interval cannot be computed.
b. Dependent Variable: MAT.

2.1.18 pav. SPSS rezultatai

1.10. HLM modelių analizė naudojantis SAS programa

Remdamasis statistika galiu įrodyti viską, išskyrus tiesą.
Dž. Kaningas

SAS ir SPSS programomis atliekamos HLM analizės labai panašios. Naudojamas tas pats jungtinis modelis, fiksuotieji ir atsitiktiniai kintamieji, grupavimo kintamieji, ta pati rezultatų struktūra ir kt. Tačiau naudojantis SAS programa galima atlikti gilesnę analizę. Be to, SPSS programa patogiau naudotis. Abi programos yra nepigūs komerciniai produktai.

SAS programoje HLM modelių analizė atliekama taikant procedūrą PROC MIXED.

Tie procedūros sakiniai ir parametrai, kurių reikia klasikinei hierarchinių modelių analizei, pateikiamos 2.1.3 lentelėje.

2.1.3 lentelė. PROC MIXED procedūros sintaksės paaiškinimai

PROC MIXED DATA=duomenų rinkinys COVTEST METHOD=analizės metodas;	Nurodomas duomenų rinkinys ir kaip bus vertinami parametrai. Jeigu metodas nenurodomas, taikomas REML metodas. COVTEST reiškia, kad bus įvertintos liekamųjų paklaidų kovariacijos ir dispersijos
CLASS W V	Išvardijami kategoriniai (grupavimo) kintamamieji W ir V.
MODEL Y = X Z/ SOLUTION CHISQ COVB	Nurodomas priklausomas kintamasis Y ir išvardijami visi fiksuotieji modelio kintamieji X ir Z. Parinktis CHISQ reiškia hipotezių apie fiksuotų poveikio parametrų lygybę nuliui tikrinimą. COVB naudojamas skaičiuojant fiksuoto poveikio parametrų standartines paklaidas
RANDOM INTERCEPT X / SUBJECT=W TYPE=UN;	Po RANDOM išvardijami atsitiktiniai kintamieji. Grupavimo kintamasis W nurodomas po SUBJECT. Kovariacijų matricos pavidalą nurodo parinktis TYPE. TYPE=UN reiškia, kad įvertinamos visos kintamųjų kovariacijos ir dispersijos
ESTIMATE	Įvertinama $\mathbf{C}\boldsymbol{\gamma}$ esant pasirinktam vektoriniam kontrastui C
CONTRAST	Vartotojo suformuluotų hipotezių apie kontrastus aprašymas

Mišriuose modeliuose labai svarbi kintamųjų sąveika. Tarkime, kad toliau nagrinėjame matematikos testo rezultatų HLM modelius ir norime nustatyti, kaip mokyklos aplinka sustiprina (susilpnina) socialinio ir ekonominio statuso poveikį mokinio matematikos testo rezultatams. Kitaip tariant, norime įvertinti konteksto poveikį mokiniui. Taigi reikia įvertinti koeficientus prie $CSES \cdot MSES$ ir $VK \cdot CSES$ (prisiminkime, kad VK žymi mokyklos tipą: 0 – Vilniaus mokykla, 1 – kaimo mokykla).

Analizuosime HLM modelį, kuriame yra viena kintamųjų sąveika. Modelio pirmasis lygmuo yra toks:

$$MAT = \beta_0 + \beta_1 CSES + e.$$

Antrasis lygmuo užrašomas lygtimis

$$\beta_0 = \gamma_{00} + \gamma_{01} VK + u_0,$$
$$\beta_1 = \gamma_{10} + \gamma_{11} VK + u_1.$$

Taigi jungtinis modelis yra

$$\begin{aligned} MAT &= \gamma_{00} + \gamma_{01} VK + u_0 + (\gamma_{10} + \gamma_{11} VK + u_1) CSES + e \\ &= [\gamma_{00} + \gamma_{01} VK + \gamma_{10} CSES + \gamma_{11} VK \cdot CES] \\ &\quad + [u_0 + u_1 CESE] + e. \end{aligned} \tag{2.1.28}$$

Panašiai, kaip ir tirdami modelius SPSS programa, nustatome, kad kintamieji *CSES*, *VK* ir $CSES \cdot VK$ yra fiksuoti, o *CSES* ir postūmis – atsitiktiniai. Pažvelkime į neatsitiktinę jungtinio modelio dalį. Jeigu $\gamma_{11} > 0$, tai Vilniaus mokiniams ($VK = 1$) modelyje atsiras papildoma rezultato priklausomybė nuo socialinio ir ekonominio statuso.

Šio modelio SAS programos sintaksė yra tokia:

```
PROC MIXED DATA =STAT3HLM1 COVTEST;
CLASS IDMOK;
MODEL MAT = CSES VK CSES*VK /solution covb;
RANDOM INTERCEPT CSES / SUBJECT = IDMOK TYPE = UN; RUN;
```

Skaičiavimo rezultatai yra pateikti 2.1.19 pav. Juos galima interpretuoti taip:

- mokinių, kurių *CSES* yra aukštesnis, *MAT* rezultatai geresni (fiksuoto poveikio parametras prie *CSES* teigiamas);
- vidutinį *SES* turintys Vilniaus mokiniai geriau atlieka testo užduotis nei kaimo mokiniai (fiksuoto poveikio parametras prie *VK* statistiškai reikšmingai teigiamas);
- *MAT* nuo *CSES* vienodai priklauso tiek nagrinėjant Vilniaus, tiek ir kaimo mokyklas (sąveikos $VK \cdot CSES$ koeficientas statistiškai nereikšmingas).

Galutinė išvada tokia: duomenys nepatvirtino, kad modelyje reikalinga kintamųjų sąveika $VK \cdot CSES$.

Ankstesniuose skyreliuose nagrinėtų pavyzdžių analizės sintaksė aprašyta 2.1.4 lentelėje. Čia STAT3HLM1 yra duomenų aibės pavadinimas. Nesunku pastebėti, kad SAS sintaksė labai panaši į naudojamą SPSS programoje. Visi rezultatai ir jų interpretacijos nesiskiria nuo gautųjų SPSS programa. Todėl jų neaprašinėsime.

Sakiniai CONTRAST ir ESTIMATE naudojami tikrinant hipotezes apie parametrų tiesinius darinius. SAS programos PROC MIXED procedūra papildoma sakiniu

```
ESTIMATE 'Postūmio įvertis Vilniui' intercept 1 VK 1;
```

kuris parodo, kad bus tikrinama hipotezė, ar $1 \cdot \hat{\gamma}_{00} + 1 \cdot \hat{\gamma}_{01} = 0$. Pridėtasis komentaras „Postūmio įvertis Vilniui“ padės lengviau rasti šios hipotezės tikrinimo rezultatus tarp visų SAS programa gautų rezultatų.

Įtraukus į programą sakinį CONTRAST, bus tikrinama hipotezė apie kelių fiksuoto poveikio parametrų tiesinių darinių lygybę nuliui. Jau 1.8 skyrelyje minėjome, kad taip

Model Information

Data Set	WORK.STAT3HLM1
Dependent Variable	mat
Covariance Structure	Unstructured
Subject Effect	idmok
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Laisvo pavidalo G

2 lygmens grupavimo kintamasis

$\sigma^2 I$

Fiksuotų poveikių reikšmingumo tikrinimo parinktys

Dimensions

Covariance Parameters	4
Columns in X	4
Columns in Z Per Subject	2
Subjects	27
Max Obs Per Subject	39

$\tau_{00}, \tau_{10}, \tau_{01}, \sigma^2$

Fiksuoto poveikio (kartu su postūmiu) parametrų skaičius

Antro lygmens elementų(mokyklų) skaičius

Maksimalus mokinių skaičius vienoje mokykloje

Atsitiktinio poveikio (kartu su postūmiu) parametrų skaičius

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6224.29597697	
1	2	6124.12008399	0.00003790
2	1	6124.01341848	0.00000102
3	1	6124.01075490	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	idmok	1678.07	547.52	3.06	0.0011
UN(2,1)	idmok	19.4904	60.1521	0.32	0.7459
UN(2,2)	idmok	9.7362	11.4629	0.85	0.1978
Residual		4292.04	274.21	15.65	<.0001

Laisvo pavidalo matricos G įverčiai; reikšmingumas

σ^2

Fit Statistics

-2 Res Log Likelihood	6124.0
AIC (smaller is better)	6132.0
AICC (smaller is better)	6132.1
BIC (smaller is better)	6137.2

Informacinių indeksų reikšmės. Naudingos lyginant modelius.

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	100.29	<.0001

lyginimo su nuliniu modeliu rezultatai

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	482.42	10.4058	25	46.36	<.0001
cses	7.1991	1.5465	25	4.65	<.0001
vk	58.6765	17.7514	489	3.31	0.0010
cses*vk	2.1745	2.5734	489	0.84	0.3985

Fiksuoto poveikio parametrų įverčiai; reikšmingumas

2.1.19 pav. SAS rezultato pavyzdys

galima patikrinti hipotezę apie tai, ar modelyje iš viso reikalingas konkretus kintamasis. Tarkime, norime patikrinti hipotezę, kad kintamasis *VK* (2.1.28) modelyje iš viso nereikalingas. Taigi reikėtų patikrinti, ar $\gamma_{01} = \gamma_{11} = 0$. Nesunku įsitikinti, kad ši hipotezė ekvivalenti tokiai:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \gamma_{01} \\ \gamma_{11} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

2.1.4 lentelė. Nagrinėtų modelių PROC MIXED sintaksė

	Modelis	PROC MIXED sintaksė
1	Besąlyginis: $MAT = \beta_0 + e$ $\beta_0 = \gamma_{00} + u_0$	PROC MIXED DATA =STAT3HLM1; CLASS IDMOK; MODEL MAT = /solution chisq; RANDOM INTERCEPT / SUBJECT = IDMOK TYPE = UN; RUN;
2	Atsitiktinio posvyrio ir postūmio: $MAT = \beta_0 + \beta_1 CSES + e$ $\beta_0 = \gamma_{00} + u_0$ $\beta_1 = \gamma_{10} + u_1$	PROC MIXED DATA =STAT3HLM1; CLASS IDMOK; MODEL MAT = CSES /solution chisq; RANDOM INTERCEPT CSES / SUBJECT = IDMOK TYPE = UN; RUN;
3	Su antrojo lygmens intervaliniu kintamuoju: $MAT = \beta_0 + \beta_1 CSES + e$ $\beta_0 = \gamma_{00} + \gamma_{01} MSES + u_0$ $\beta_1 = \gamma_{10} + u_1$	PROC MIXED DATA =STAT3HLM1; CLASS IDMOK; MODEL MAT = CSES MSES /solution chisq; RANDOM INTERCEPT CSES / SUBJECT = IDMOK TYPE = UN; RUN;
4	Su antrojo lygmens kategoriniu kintamuoju: $MAT = \beta_0 + \beta_1 CSES + e$ $\beta_0 = \gamma_{00} + \gamma_{01} MSES + \gamma_{02} VK + u_0$ $\beta_1 = \gamma_{10} + \gamma_{11} VK + u_1$	PROC MIXED DATA =STAT3HLM1; CLASS IDMOK; MODEL MAT = CSES MSES VK CSES*VK /solution chisq; RANDOM INTERCEPT CSES / SUBJECT = IDMOK TYPE = UN; RUN;

Sakinyje CONTRAST nurodome ne pačius fiksuotus parametrus, o kintamuosius, kurie iš jų dauginami. Todėl lygčių sistemą užrašome taip:

*CONTRAST 'Nera vietos poveikio' vk 1 cses*vk 1, vk 1 cses*vk -1;*

Nulinė hipotezė teigia, kad lygčių sistema teisinga, t. y. kad abu parametrai lygūs nuliui. Alternatyvioji hipotezė – kad taip nėra. Taigi galime daryti prielaidą, kad kintamasis *VK* modeliui nereikalingas tik tada, kai nulinė hipotezė neatmetama. Sprendimas priimamas atsižvelgus į chi kvadrato statistikos reikšmę. Žinoma, naudojantis SAS programa, lengviau rasti p reikšmę ir ją palyginti su iš anksto pasirinktu reikšmingumo lygmeniu. Tarkime, kad pasirinkome reikšmingumo lygmenį $\alpha = 0{,}05$. Tada nulinės hipotezės neatmesime (nuspręsime, kad gyvenamosios vietos įtaka matematikos rezultatams statistiškai nereikšminga), jeigu $p \geqslant \alpha$. SAS programa gauti rezultatai pateikiami 2.1.20 paveiksle. Matome, kad $p < 0{,}05$, todėl galima teigti, kad vietovės įtaka rezultatams statistiškai reikšminga.

Contrasts

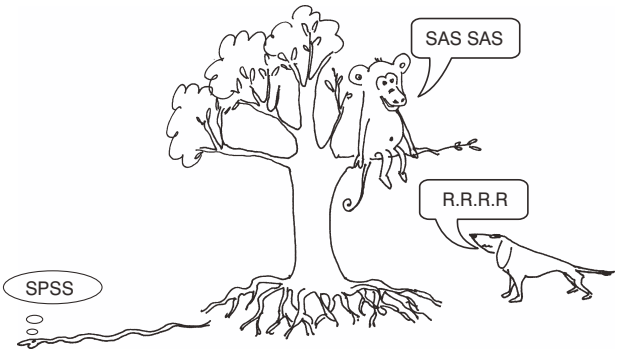
Label	Num DF	Den DF	F Value	Pr > F
Nera vietos poveikio	2	514	5.65	0.0037

2.1.20 pav. Vietovės poveikio įvertinimas naudojantis SAS programa

1.11. HLM pavyzdžių analizė naudojantis R programa

Jei iškart nepavyko, pavadink bandymą versija 1.0.

Šiame skyrelyje aprašysime kaip atliekama HLM modelių analizė naudojantis R programa. Ja galima atlikti išsamesnę analizę nei SPSS programa, tačiau ji nėra tokia „draugiška“ vartotojui.



Naudojama sintaksė. Iš pradžių aptarsime R programos sintaksę hierarchinių modelių analizei. Norint HLM analizę atlikti R programa, galima naudotis funkcijų biblioteka *nlme*[13]: *library(nlme)*. Standartinei jungtinio HLM modelio analizei naudojama *lme* funkcija, kurios sintaksė yra tokia:

lme(priklausomas kintamasis~ fiksuoto poveikio kintamieji, duomenų vardas, atsitiktinio poveikio kintamieji | grupavimo kintamasis).

Tarkime, turime tokį jungtinį modelį:

$$Y = \gamma_{00} + \gamma_{01}W + \gamma_{02}V + \gamma_{10}X + \gamma_{11}WX + [u_0 + Xu_1 + e].$$

Fiksuotieji kintamieji yra X, W, V, atsitiktiniai kintamieji – postūmis ir X. Tarkime, kad kintamasis *ID* nurodo, kokiam antrojo lygmens elementui priklauso duomenys. Duomenų vardas – *datafile*. Tada sakinys *lme* atrodo taip:

lme(Y s~1+X+W+V, data=datafile, random=1+X | ID).

Funkcijos *summary, anova, print, plot* naudojamos analizės rezultatų išvedimui. Standartinei HLM analizei užtenka funkcijos *summary*. Dabar nesunkiai galime užrašyti 1.9 skyrelyje nagrinėtus modelius.

1 *Besąlyginio modelio* jungtinė lygtis:

$$MAT = \gamma_{00} + u_0 + e.$$

R programa:

[13] Galima naudoti ir funkcijų bibliotekos *lme4* funkciją *lmer*.

```
model.0<-lme(MAT ~ 1, data=dat, random=~ 1|IDMOK)
summary(model.0)
```

2 *Atsitiktinio postūmio ir posvyrio* modelio jungtinė lygtis:

$$MAT = \gamma_{00} + \gamma_{10}CSES + [u_1 CSES + u_0 + e].$$

R programa:

```
model.2<-lme(MAT ~ 1+CSES, data=dat, random=~1+CSES|IDMOK)
summary(model.2)
```

3 *Modelio su antrojo lygmens kategoriniu kintamuoju* jungtinė lygtis:

$$MAT = \gamma_{00}+\gamma_{01}MSES+\gamma_{10}CSES+\gamma_{02}VK+\gamma_{12}VK*CSES+[u_1 CSES+u_0+e].$$

R programa:

```
model.4<-lme(MAT ~ 1+CSES+MSES+VK+VK*CSES,
data=dat, random=~1+CSES|IDMOK)
summary(model.4)
```

Rezultatai. Jie visiškai analogiški gautiems SPSS programa. Pateikiame modelio su antrojo lygmens kategoriniu kintamuoju analizės naudojantis R programa rezultatus (žr. 2.1.21 pav.). Pirmojoje lentelėje nurodytos informacinių indeksų reikšmės. Antroji lentelė skirta atsitiktinio poveikio parametrams. Skirtingai nuo SPSS programos, joje yra ne dispersijos ir kovariacijos, bet standartiniai nuokrypiai ir koreliacijos. Žinoma, vienas charakteristikas nesunku išsireikšti per kitas. Taigi suapvalinę iki dviejų ženklų po kablelio gauname:

$$\sqrt{\hat{\tau}_{00}} = 19{,}53, \qquad \sqrt{\hat{\tau}_{11}} = 3{,}21, \qquad \hat{\sigma} = 65{,}53, \qquad \frac{\hat{\tau}_{01}}{\sqrt{\hat{\tau}_{00}\hat{\tau}_{11}}} = 0{,}15.$$

Norint gauti pačius dispersijų įverčius, reikėtų naudoti parinktį *VarCorr(model.4)*. Trečiojoje lentelėje yra fiksuoto poveikio parametrų įverčiai: $\hat{\gamma}_{00} = 47{,}23$, $\hat{\gamma}_{10} = 7{,}27$ ir t. t. Ketvirtojoje lentelėje pateikta fiksuoto poveikio įverčių koreliacijų matrica. Ji įdomi tik tuo, kad didelės koreliacijos rodo, jog modelis netinka.

Dviejų modelių palyginimas. Komanda *Anova* pateikia tris charakteristikas, leidžiančias palyginti du modelius. Lyginamos *AIC*, *BIC* reikšmės ir tikrinama hipotezė apie dviejų modelių tikėtinumų santykių testo p reikšmę. Dviejų HLM modelių palyginimui reikia naudoti didžiausio tikėtinumo metodu gautus įverčius, todėl atitinkamai atnaujiname ir pačius modelius:

```
model.2B=update(model.2, method="ML")
model.4B=update(model.4,method="ML")
anova(model.2B,model.4B)
```

```
> summary(model.4)
Linear mixed-effects model fit by REML
 Data: dat
     AIC     BIC   logLik
 6109.135 6147.726 -3045.567

Random effects:
 Formula: ~1 + CSES | IDMOK
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev   Corr
(Intercept) 19.524839 (Intr)
CSES        3.205896 0.146
Residual    65.528512

Fixed effects: MAT ~ 1 + CSES + MSES + VK + VK * CSES
            Value Std.Error  DF   t-value p-value
(Intercept) 47.23311  60.83746 514  0.776382  0.4379
CSES         7.27030   1.55839 514  4.665276  0.0000
MSES        23.93195   3.31679  24  7.215393  0.0000
VK         -47.92947  17.50488  24 -2.738063  0.0115
CSES:VK      2.07155   2.59352 514  0.798739  0.4248
 Correlation:
      (Intr) CSES   MSES   VK
CSES   -0.006
MSES   -0.995  0.012
VK      0.787 -0.029 -0.825
CSES:VK  0.006 -0.601 -0.009  0.042

Standardized Within-Group Residuals:
      Min        Q1       Med        Q3       Max
-2.86523233 -0.64336811 -0.01344669  0.60060296  3.06657464

Number of Observations: 543
Number of Groups: 27
```

$$\frac{\hat{\tau}_{01}}{\sqrt{\hat{\tau}_{00}\hat{\tau}_{11}}}$$

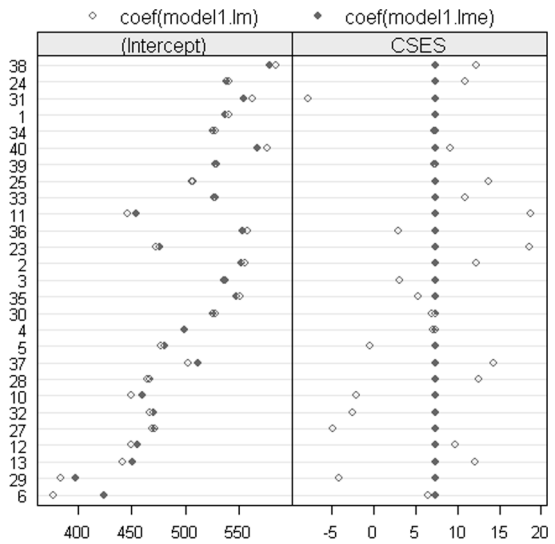
2.1.21 pav. R programa gauti rezultatai, kai modelyje yra antrojo lygmens kategorinis kintamasis

Palyginę atsitiktinio posvyrio ir postūmio modelį su modeliu, turinčiu antrojo lygmens kategorinį kintamąjį (žr. 2.1.4 lentelės antrąjį ir ketvirtąjį modelius), matome, kad p reikšmė maža ($p < 0{,}05$). Gautas rezultatas leidžia spėti, kad ketvirtas modelis geresnis už antrąjį (žr. 2.1.22 pav.).

```
> anova(model.2,model.4)
     Model df     AIC     BIC   logLik  Test  L.Ratio p-value
model.2     1  6 6165.651 6191.434 -3076.826
model.4     2  9 6130.467 6169.141 -3056.233 1 vs 2 41.18479  <.0001
>
```

2.1.22 pav. Dviejų modelių palyginimas naudojantis R programa

Palyginimas su tiesinės regresijos modeliu. Šio dalies pradžioje kalbėdami apie HLM modelių svarbą nurodėme, kad neracionalu tirti regresijos modelius, sudarytus kiekvienai grupei atskirai. Dabar paaiškinsime, kaip tai matyti iš grafiko.



2.1.23 pav. Parametrų įverčių išsibarstymas

Žemiau užrašyta R programa, kai nagrinėjamas matematikos pasiekimų HLM modelis su atsitiktiniais postūmio ir posvyrio parametrais. Kartu nurodyta, kad tiriant modelį regresijos skaičiuojamos kiekvienai mokyklai atskirai.

```
model1.lme<-lme(MAT ~ 1+CSES,random= ~ 1+CSES|IDMOK, data = dat.grp)
model1.lm<-lmList(MAT ~ 1+CSES|IDMOK,data=dat.grp)
plot(compareFits(coef(model1.lm),coef(model1.lme))) )
```

Kiekvienos mokyklos parametrų įverčių išsibarstymas pavaizduotas 2.1.23 paveiksle. Matome, kad HLM parametrų įverčių yra stabilesni, nes juos galima uždengti siauresne juostele, t. y. jie mažiau išsibarstę.

1.12. Modelių diagnostika

Kompiuteriai labai greitai daro tikslias klaidas.

Modelio diagnostika – tai modelio prielaidų tikrinimas, t. y. tikrinimas, ar reikiamos prielaidų sąlygos tenkinamos, o jei netenkinamos – tai kodėl. Šiame skyrelyje aptarsime dvi svarbias prielaidas:

1 Pirmojo lygmens kintamųjų paklaidos e turi būti normaliai pasiskirsčiusios ir turėti tą pačią dispersiją, t. y. $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. Tai vadinamoji paklaidų homogeniškumo prielaida.

2 Antrojo lygmens paklaidos u turi būti normaliai pasiskirsčiusios, t. y. $u_{kj} \sim \mathcal{N}(0, \tau_{kk})$, $k = 1, \ldots, K$, $j = 1, 2, \ldots, J$.

Jei šios prielaidos nėra tenkinamos, galima suabejoti modelio parametrų įverčių patikimumu, ar net paties modelio tinkamumu. Todėl modelio diagnostika yra svarbus HLM analizės elementas. Apsiribosime aprašymu, kaip patikrinti, ar prielaidos teisingos, nesigilindami į priežastis, dėl ko viena ar kita prielaida netenkinama. Paprasčiausias būdas – vizualiai įsitikinti, kad paklaidos normaliai pasiskirsčiusios, t. y. žvilgtelėti į paprastus, bet informatyvius kvantilių grafikus (*qq-plots*).

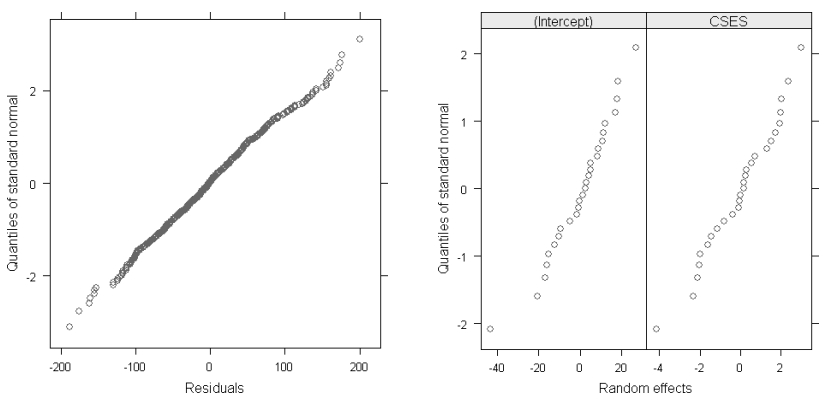
Kvantilių grafikas braižomas taip: x ašyje atidedami standartinio normaliojo skirstinio $1/n, 2/n, \ldots, 1$ kvantiliai (čia n – imties didumas), o y ašyje – atitinkami tiriamojo dydžio empiriniai kvantiliai (variacinės eilutės pirmoji, antroji, $\ldots$, n-oji reikšmės). Kuo labiau tiriamasis dydis panašus į normalųjį, tuo gauti taškai yra arčiau tiesės. Beje, kuo imtis didesnė, tuo lengviau suprasti, ar normalumo sąlyga tenkinama.

Ketvirtojo 2.1.4 lentelės modelio R programos sakinys pirmojo ir antrojo lygmenų paklaidų e ir u kvantilių grafikams braižyti yra toks:

```
qqnorm(model.4,~resid(.))
```

```
qqnorm(model.4,~ranef(.))
```

Iš gautų grafikų (žr. 2.1.24 pav.) matyti, kad abiejų lygmenų paklaidas galima laikyti normaliai pasiskirsčiusiomis.

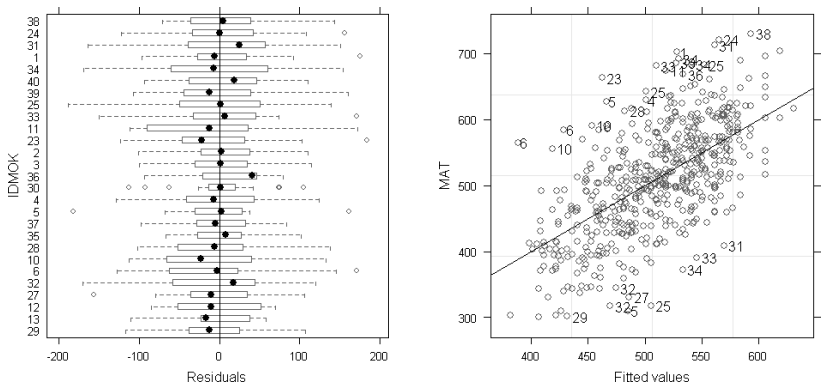


a) e paklaidų kvantilių grafikas b) u paklaidų kvantilių grafikas

2.1.24 pav. R programa gauti kvantilių grafikai homogeniškumui tikrinti

Kiekvienos mokyklos paklaidų išsibarstymo diagrama parodo, ar paklaidos yra homogeninės. Kuo vienodesnė „juosta“ uždengia taškus, tuo paklaidos labiau homogeninės. R programos sakinys, skirtas stačiakampei diagramai braižyti, yra toks:

```
plot(model.lme,IDMOK ~ resid(.),abline=0)
```



a) Paklaidų išsibarstymo diagrama b) Prognozės tikslumas

2.1.25 pav.

Gauti rezultatai pateikti 2.1.25 paveiksle, a). Matyti, kad visose mokyklose paklaidų sklaida yra panaši.

Pabaigoje pateikiame R programos sakinį spėjimo kokybei įvertinti „iš akies“:

```
plot(model.lme,MAT~ fitted(.),id=0.05,adj=-0.3,abline=c(0,1))
```

Gautas brėžinys parodytas 2.1.25 paveiksle, b). Tiesė – prognozuojamos reikšmės, o taškų „debesėlis“ – tikrosios imties duomenų reikšmės. Kuo „debesėlis“ siauresnis ir vienodesnis, tuo prognozės tikslesnės.

1.13. Daugialygiai modeliai

Socialinių mokslų atstovai jaučia slaptą pavydą biostatistikams ir fizikams, nes šie turi laimę dirbti su normaliaisiais duomenimis, o patiems tenka kankintis su ranginiais ir nominaliaisiais.

Šiame skyrelyje trumpai aptarsime, kaip sudaromi ir tiriami hierarchiniai modeliai, turintys daugiau nei du lygmenis. Praktiškai visa analizė ir modelių sudarymo logika lieka ta pati. Laikomasi tokio principo:

- Visos liekamosios paklaidos turi normaliuosius skirstinius. Jų vidurkiai yra nuliniai, o dispersijos nežinomos.
- To paties lygmens kintamųjų paklaidos gali koreliuoti, bet skirtingų lygmenų kintamųjų paklaidos nekoreliuoja.

Tarkime, kad žinome ne tik mokinio matematikos testo rezultatus *MAT*, jo socialinį ir ekonominį statusą *CSES*, mokyklos socialinį ir ekonominį statusą *MSES*, bet ir klasės socialinį ir ekonominį statusą *KSES*. Mokyklą nurodo kintamasis *IDMOK*, klasę – *IDKLA*. Praleisdami mokinį ir klases žyminčius indeksus užrašome besąlyginį modelį:

Mokinio lygmuo:

$$MAT = \beta_0 + e.$$

Klasės lygmuo:

$$\beta_0 = \gamma_{00} + u_0.$$

Mokyklos lygmuo:

$$\gamma_{00} = \alpha_{000} + r_{00}.$$

Čia $e \sim \mathcal{N}(0, \sigma^2)$, $u_0 \sim \mathcal{N}(0, \tau_{00})$, $r_{00} \sim \mathcal{N}(0, \varkappa_{0000})$.

Standartinė modelio analizė reikštų parametrų $\hat{\alpha}_{0000}$, $\hat{\sigma}^2$, $\hat{\tau_{00}}$, $\hat{\varkappa}_{0000}$ įverčių skaičiavimą ir hipotezių apie jų lygybę nuliui tikrinimą. Nauja tai, kad skaičiuojami du tarpklasiniai antrojo ir trečiojo lygmenų koreliacijos koeficientai, t. y.

$$ICC_2 = \frac{\hat{\tau}_{00}}{\hat{\sigma}^2 + \hat{\tau}_{00} + \hat{\varkappa}_{0000}}, \qquad ICC_3 = \frac{\hat{\varkappa}_{0000}}{\hat{\sigma}^2 + \hat{\tau}_{00} + \hat{\varkappa}_{0000}}.$$

Mažos koeficientų reikšmės rodo, kad hierarchinis modelis netinka. Sudėtingesni modeliai užrašomi panašiai, kaip ir dviejų lygmenų. Tarkime, kad yra tokios spėjamos priklausomybės.

Mokinio lygmuo:

$$MAT = \beta_0 + \beta_1 CSES + e.$$

Klasės lygmuo:

$$\beta_0 = \gamma_{00} + \gamma_{01} KSES + u_0,$$
$$\beta_1 = \gamma_{10} + u_1.$$

Mokyklos lygmuo:

$$\gamma_{00} = \alpha_{000} + \alpha_{001} MSES + r_{00},$$
$$\gamma_{01} = \alpha_{010} + \alpha_{011} MSES + r_{01},$$
$$\gamma_{10} = \alpha_{100} + r_{10}.$$

Čia $e \sim \mathcal{N}(0, \sigma^2)$, $u_j \sim \mathcal{N}(0, \tau_{jj})$, $u_1 \sim \mathcal{N}(0, \tau_{11})$, $\mathbb{C}\mathrm{ov}(u_0, u_1) = \tau_{01}$, $r_{ij} \sim \mathcal{N}(0, \varkappa_{ijij})$, $\mathbb{C}\mathrm{ov}(r_{ij}, r_{mk}) = \varkappa_{ijmk}$. Jungtinė lygtis:

$$\begin{aligned} MAT = {} & \alpha_{000} + \alpha_{001} MSES + \alpha_{010} KSES + \alpha_{011} KSES \cdot MSES \\ & + \alpha_{100} CSES + [e + u_0 + u_1 CSES + r_{00} + r_{01} KSES + r_{10} CSES]. \end{aligned}$$

Standartinė modelio analizė reikštų parametrų $\hat{\alpha}_{ijk}$, $\hat{\sigma}^2$, $\hat{\tau_{ij}}$, $\hat{\varkappa}_{ijmk}$ įverčių skaičiavimą ir hipotezių apie jų lygybę nuliui tikrinimą. Gautas informacinių indeksų reikšmes lygintume su besąlyginio modelio reikšmėmis. Primename, kad mažesnė informacinio indekso reikšmė yra rodiklis, kad modelis geriau tinka duomenims. Lyginant du modelius, skaičiuojami ir dispersijų įverčių procentiniai pokyčiai. Mažesnės dispersijos rodo geresnį modelio tinkamumą.

SPSS ir R programose atsižvelgiama į jungtinę lygtį. Kintamieji *MSES*, *KSES*, $KSES \cdot MSES$ ir *CSES* ir postūmis yra fiksuoti, o *CSES* ir posvyris – antrojo lygmens atsitiktiniai kintamieji. Juos nustatome, atsižvelgę į kintamuosius prie u_j. Analogiškai, atsižvelgdami į r_{ij}, nustatome, kad trečiojo lygmens atsitiktiniai kintamieji yra postūmis, *CSES* ir *KSES*. Taikant šį modelį, SPSS programa yra tokia:

```
MIXED mat WITH KSES CSES MSES
/CRITERIA = CIN(95) MXITER(100) MXSTEP(5) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE)
LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED = KSES CSES MSES KSES*MSES | SSTYPE(3)
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT CSES | SUBJECT(IDKLA) COVTYPE(UN)
/RANDOM INTERCEPT KSES CSES | SUBJECT(IDMOK) COVTYPE(UN) .
```

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Uper Bound
Residual		4209,247	273,9584	15,365	,000	3705,135595	4781,947599
Intercept + CSES	UN (1,1)	1,657245[a]	,000000	.	.	.	.
[subject=IDKLA]	UN (2,1)	-2,426233	50,897007	-,048	,962	-102,182534	97,330069
	UN (2,2)	3,579983[a]	,000000	.	.	.	.
Intercept + KSES	UN (1,1)	542,7525	224,1867	2,421	,015	241,550871	1219,537231
+ CSES [subject	UN (2,1)	-1,938949	130,0737	-,015	,988	-256,878812	253,000915
= IDMOK]	UN (2,2)	65,310960	118,8563	,549	,583	1,844694	2312,319220
	UN (3,1)	30,943269	65,329672	,474	,636	-97,100535	158,987072
	Un (3,2)	1,878219	36,891212	,051	,959	-70,427229	74,183666
	Un (3,3)	5,319241[a]	17,136108	,310	,756	,009631	2937,806512

a. This covariance parmameter is redundant. The test statistic and confidance interval cannot be computed.
b. Dependent Variable: MAT.

2.1.26 pav. SPSS programa gauti trijų lygių modelio parametrų įverčiai

Kaip matyti iš 2.1.26 paveikslo, atsitiktinio poveikio parametrų įverčiai pateikiami kiekvienam lygmeniui atskirai. Pirmasis $U(1,1)$ atitinka $\hat{\tau}_{00} = 1{,}66$. Antrasis $U(1,1)$ atitinka $\hat{\varkappa}_{0000} = 2{,}42$ ir t. t. Pavyzdžiui, $\hat{\tau}_{01} = -2{,}43$, $\hat{\tau}_{11} = 3{,}58$. Matome, kad trijų lygmenų modelis blogai suderintas su duomenimis. Taip atsitiko dėl to, kad kiekvienoje mokykloje buvo apklaustos tik dvi klasės. Taigi *MSES* ir *KSES* įgyja beveik vienodas reikšmes. Tai dar kartą rodo, kad hierarchinių modelių negalima sudarinėti aklai, neatsižvelgiant į duomenis.

atsitiktinio poveikio parametras	HLM	informaciniai indeksai
	HLM lygmenys	
fiksuoto poveikio parametras	*ICC*	

UŽDAVINIAI

1. Faile **stat3hlm1.sav** yra duomenys apie Lietuvos mokinių tiksliųjų mokslų gebėjimus. Turime tokius kintamuosius: *CSES* – centruotas mokinio socialinis ir ekonominis statusas jo mokykloje; *MSES* – mokyklos socialinio ir ekonominio statuso vidurkis;

IDMOK – mokyklos indeksas, *VK* – mokyklos tipas (0 — Vilniaus mokykla, 1 – kaimo mokykla); *FIZ* – fizikos testo rezultatai. Išnagrinėkite ir palyginkite išvardytus modelius.

a) Besąlyginis modelis:

$$\begin{aligned} FIZ &= \beta_0 + e, \\ \beta_0 &= \gamma_{00} + u_0. \end{aligned}$$

b) Atsitiktinio posvyrio ir postūmio modelis:

$$\begin{aligned} FIZ &= \beta_0 + \beta_1 CSES + e, \\ \beta_0 &= \gamma_{00} + u_0, \\ \beta_1 &= \gamma_{10} + u_1. \end{aligned}$$

c) Modelis su antrojo lygmens intervaliniu kintamuoju:

$$\begin{aligned} FIZ &= \beta_0 + \beta_1 CSES + e, \\ \beta_0 &= \gamma_{00} + \gamma_{01} MSES + u_0, \\ \beta_1 &= \gamma_{10} + u_1. \end{aligned}$$

d) Modelis su antrojo lygmens pseudokintamuoju:

$$\begin{aligned} FIZ &= \beta_0 + \beta_1 CSES + e, \\ \beta_0 &= \gamma_{00} + \gamma_{01} MSES + \gamma_{02} VK + u_0, \\ \beta_1 &= \gamma_{10} + \gamma_{11} VK + u_1. \end{aligned}$$

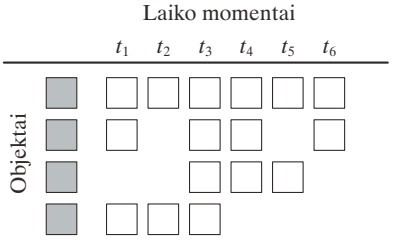
2. Kurie ankstesnio uždavinio modelių parametrai yra atsitiktinio poveikio, o kurie – fiksuoto? Kokios yra parametrų interpretacijos?

2. HLM MODELIŲ TAIKYMAS PANELINIŲ DUOMENŲ ANALIZEI

Kartą gyveno Džekas, kuris turguje pardavęs karvę įsigijo greitai augančių pupų. Iš visų pasodintų pupų sudygo vienintelė. Taip prasideda viena iš žinomiausių anglų pasakų. Tolesnį pupos augimą skirtingi pasakos variantai nusako nevienodai. Kartais pasakojama, kad, pupa per pirmą naktį išaugo iki lovos, per antrąją – iki stogo[14], o per trečią – iki dangaus. Pagal kitą variantą iki dangaus pupa išaugo iškart, per vieną naktį.

Jeigu sudarant pupos augimo modelį reikėtų atsaižvelgti į visus pasakos variantus, turėtume taikyti atitinkamai modifikuotą hierarchinį modeliavimą.

Visuose iki šiol nagrinėtuose hierarchiniuose modeliuose pirmojo lygmens kintamieji buvo nepriklausomi. Vis dėlto neretai prireikia ir sudėtingesnių modelių. Šiame skyrelyje glaustai aptariama situacija, kai kiekvienam objektui koks nors požymis matuojamas keletą kartų skirtingais laiko momentais. Priklausomybė atsiranda dėl pakartotinių kintamojo matavimų. Įprasta duomenų hierarchija čia nebūdinga. Greičiau galima sakyti, kad turime duomenų grupes (stebėjimus laiko atžvilgiu), kurias vienija antrojo lygmens elementai (individai, objektai). Taigi šiuo atveju pirmasis lygmuo – laiko momentai; antrasis lygmuo – objektai (žmonės, šalys, bandomieji triušiai). Duomenų struktūrą patogiau atvaizduoti kitaip nei ankstesniuose skyreliuose – skiriant kiekvienam respondentui atskirą eilutę (žr. 2.2.1 pav.).



2.2.1 pav. Hierarchinė panelinių duomenų struktūra

[14] Kurį Džekas – tikras žaliasis – praardė.

2.1. Paneliniai duomenys

Visi įrodymai užima vieną eilutę, jei lenta pakankamai ilga.

Neišvengiamai turime aptarti terminiją. Anglų kalboje duomenims, kai matavimai atliekami erdvėje ir laike, vartojami du terminai: *panel data* (paneliniai duomenys) ir *longitudinal data* (ilgalaikiai stebėjimai). Apskritai šie terminai laikytini sinonimais. Vis dėlto kai kurie statistikai mano, kad paneliniai duomenys gaunami tada, kai objektų nemažai, bet jie stebimi tik keletą kartų. Tuo tarpu ilgalaikiai stebėjimai atitinka situaciją, kai keletas objektų stebimi daug kartų (žr. 2.2.2 pav.). Mes vartosime tik vieną terminą, jau prigijusį lietuviškoje terminijoje, – *paneliniai duomenys*.

Panelinių duomenų pavyzdžiai:

1 Informacija apie Europos Sąjungos šalių pastarojo dešimtmečio BVP.

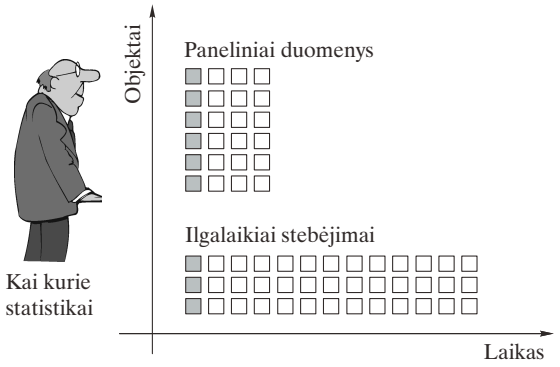
2 Atsitiktinai parinktų vaikų ūgis 8-ais, 12-ais, 14-ais ir 16-ais gyvenimo metais.

3 VU studentų išlaidos pramogoms kiekvienais studijų metais.

4 Bedarbių depresijos požymiai praėjus 2, 4, ir 6 mėnesiams po užsiregistravimo darbo biržoje.

5 Skaitymo gebėjimai po 1-ųjų, 2-ųjų ir 3-ųjų mokymosi metų.

6 Piktybinio naviko didumas po kiekvieno iš 6 chemoterapijos kursų.



2.2.2 pav. Ką mano kai kurie statistikai

Paneliniai duomenys nagrinėjami ir ekonometrijoje. Ekonometriniuose modeliuose pirmojo ir antrojo lygmens kintamieji dažnai apibrėžiami kaip atsitiktiniai dydžiai, turintys įvairius (dažniausiai normaliuosius) skirstinius. Šioje knygoje visi modeliai sudaromi esant fiksuotoms aiškinamųjų kintamųjų reikšmėms, atsitiktinės tik matavimo paklaidos.

Duomenys, gauti kintant matavimų laikui, taip pat naudojami laiko eilutėse ir blokuotųjų duomenų dispersinėje analizėje. Skirtumas tarp panelinių duomenų ir laiko eilučių beveik toks pat kaip tarp klasikinių HLM modelių ir tiesinės regresijos – naudojama daugiau informacijos nei agreguojant duomenis. Be to, paneliniai duomenys gaunami atliekant palyginti nedaug vieno objekto stebėjimų. Skirtingai nuo blokuotųjų duomenų dispersinės analizės, tiriant panelinius duomenis, modeliai sudaromi individualiems objektams, o ne vidurkiams. Dažnai panelinių duomenų analizėje reikia nustatyti ne tik grupių ar laiko įtaką – jiems būdinga lygmenų tarpusavio sąveika[15]. Pavyzdžiui, labai tikėtina, kad mergaičių vertinimų tendencijos skiriasi nuo berniukų vertinimų tendencijų (sąveika *laikas* ↔ *lytis*) ir pan.

Nėra labai griežtų normų, nusakančių, kiek turi būti pakartotinių matavimų. Ištisos knygos skiriamos modeliams, kuriuose kiekvienas objektas matuotas tik 3–4 kartus. To pakanka, kad būtų galima sudaryti gerą bendrąjį modelį. Panelinių duomenų analizėje duomenys gali būti nesubalansuoti, t. y. nebūtinai kiekvienu laiko momentu turi būti duomenų apie visus objektus. Tai matyti iš 2.2.1 paveikslo.

Sudarant pirmojo lygmens modelius, atsižvelgiama į laiko faktorių – pakartotinius matavimus ir dėl to atsiradusią matavimo paklaidų priklausomybę. Galimi du skirtingi modelių sudarymo būdai. Pirmuoju atveju laiko komponentė į modelį įtraukiama tiesiogiai, kaip papildomas regresorius, o paklaidos jau laikomos nepriklausomomis. Taip sukonstruoti modeliai vadinami *augimo modeliais*. Antruoju atveju parenkamas paklaidų kovariacijų matricos pavidalas, t. y. nusakoma, kaip paklaidos susijusios. Žinoma, sudaromi ir abiejų modelių hibridai.

Nagrinėjant panelinius duomenis, skiriamos dvi sąvokos: matavimo momentas[16] t ir prabėgęs tarp matavimų laikas $Laik$. Skirtingų objektų laikas tarp matavimų gali skirtis. Pavyzdžiui, tiriant vieną objektą, tarp pirmojo ($t = 1$) ir antrojo ($t = 2$) matavimų galėjo praeiti 15 ($Laik = 15$) dienų, o tiriant antrąjį objektą – 20 ($Laik = 20$) dienų. Vis dėlto labai dažnai visi matavimai atliekami tiksliai tuo pat metu. Dažnai būna, kad tikslesnės informacijos apie laiką tarp matavimų neturime. Tokiais atvejais laiko kintamojo reikšmė sutampa su matavimo momentu ($Laik = t$), kuris dažniausiai pradedamas koduoti nuo nulio. Pavydžiui, jeigu pacientų svorį matavome tik dietos kurso pradžioje, po mėnesio ir po dviejų mėnesių, tai natūralu kintamųjų $Laik$ ir t reikšmes koduoti $0, 1, 2$. Taip koduojama, kad lengviau būtų interpretuoti – kai modelyje imama $Laik = 0$, tai gaunamas vidutinis svoris pradiniame tyrimo etape.

[15] Angl. *Cross-Level Interactions*.

[16] Angl. *wave*.

2.2. Augimo modeliai

Ar girdėjote apie statistiką, patekusį į kalėjimą? Dabar jis turi nulį laisvės laipsnių.

Tradiciškai jais vadinami HLM modeliai, kuriuose priklausomybė nuo laiko užrašoma išreikštu pavidalu – tam tikra laiko funkcija, įtraukiama į pirmojo lymens lygtį. Paprasčiausiu atveju įtraukiama tiesiog konstanta, padauginta iš laiko, ekonometrų kalba – tiriamas modelis su laiko trendu. Paprasčiausias augimo modelis yra besąlyginis HLM modelis su tiesiniu trendu. Jis dar vadinamas *besąlyginiu* (*nuliniu*) *augimo modeliu*. Jis yra šiek tiek sudėtingesnis už tradicinį besąlyginį HLM modelį. Aprašysime jį tiksliau.

Tarkime, turime $j = 1, 2, \ldots, J$ objektų, dėl kurių laiko momentais $t = 1, 2, \ldots, T$ buvo matuojama normaliojo kintamojo reikšmė Y_{tj}.

Pirmojo lygmens lygtis:

$$Y_{tj} = \beta_{0j} + \beta_{1j} Laik + e_{tj}. \tag{2.2.1}$$

Antrojo lygmens lygtys:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{2.2.2}$$

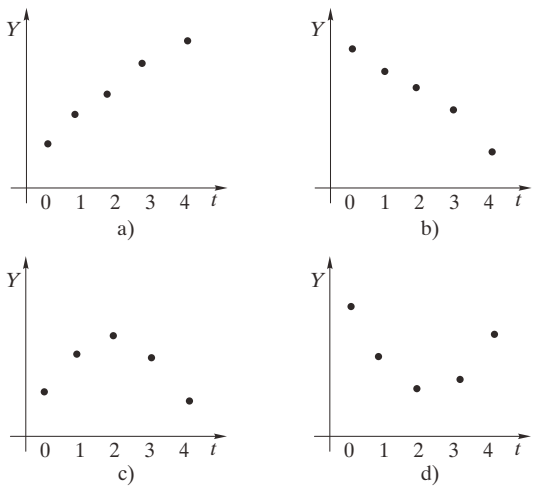
$$\beta_{1j} = \gamma_{10} + u_{1j}. \tag{2.2.3}$$

Pastebėsime, kad laiko indeksas t (pirmo lygmens indeksas) panelinių duomenų modeliuose dažnai rašomas antruoju. Mes jį parašėme pirmuoju, norėdami išlaikyti jau naudotą indeksavimo sistemą, kai pirmasis indeksas yra skiriamas pirmajam lygmeniui. Kaip ir įprastuose HLM modeliuose, tariame, kad pirmojo lygmens liekamosios paklaidos yra nepriklausomos.

Modelio prielaidos:

1) $e_{tj} \sim \mathcal{N}(0, \sigma^2)$ nepriklausomi visiems $t = 1, \ldots, T$, $j = 1, \ldots, J$;
2) $u_{0j} \sim \mathcal{N}(0, \tau_{00})$ nepriklausomi visiems $j = 1, \ldots, J$;
3) $u_{1j} \sim \mathcal{N}(0, \tau_{11})$ nepriklausomi visiems $j = 1, \ldots, J$;
4) $\mathbb{C}\mathrm{ov}(u_{0j}, u_{1j}) = \tau_{01} = \tau_{10}$;
5) e_{ij} nekoreliuoja su u_{0j} ir u_{1j} visiems $j = 1, \ldots, J$.

Atkreipiame dėmesį, kad pirmojo lygmens paklaidos e_{tj} jau yra nepriklausomos, t. y. tariame, kad priklausomybę nuo laiko modelyje aprašėme išsamiai. Panašiai, kaip ir įprastuose HLM modeliuose, γ_{00}, γ_{10} vadinami fiksuoto poveikio parametrais, o σ^2, τ_{00}, τ_{10}, τ_{11} – atsitiktinio poveikio parametrais. Modelio tyrimas niekuo nesisikria nuo jau nagrinėtųjų. Beje, labai dažnai besąlyginis augimo modelis lyginamas ir su besąlyginiu HLM modeliu, aprašytu 1.3 skyrelyje.



2.2.3 pav. Tiesiniai ir kvadratiniai trendai

Visi kiti augimo modeliai konstruojami įtraukiant naujus antrojo (neretai ir pirmojo) lygmens kintamuosius. Laikas į modelį įtraukiamas kaip tolydusis kintamasis (net ir tuo atveju, kai jis sutampa su matavimo momentais). Aprašytasis modelis laiko atžvilgiu yra tiesinis. Jeigu manome, kad laiko įtaka yra netiesinė, tai galima konstruoti kvadratinį augimo modelį ir t. t.

Kvadratinio augimo modelio pirmojo lygmens lygtis yra tokia:

$$Y_{tj} = \beta_0 + \beta_1 Laik + \beta_2 Laik^2 + e_{tj}.$$

Kokį augimo modelį – tiesinį ar kvadratinį – taikyti, gali padėti peržiūrėtos Y sklaidos matavimų atžvilgiu diagramos. Tarkime, kad keturių objektų Y reikšmės, matuotos 0, 1, 2, 3, 4 laiko momentais, pavaizduotos 2.2.3 paveiksle. Gauti trendai a) ir b) atvejais yra tiesiniai, o c) ir d) atvejais – kvadratiniai. Sudarant bet kokį augimo modelį, vertėtų iš pradžių kiekvienam objektui nubraižyti tiriamojo kintamojo kitimo laike grafikus.

2.3. Koreliuojančiųjų paklaidų modeliai

Žmogus, kuris nuolat daro kvailas klaidas, matyt, toks ir yra.

Nuo bendrųjų HLM modelių, aprašytų 1.4 skyrelyje (2.1.5), (2.1.6) lygtimis, koreliuojančiųjų paklaidų modeliai skiriasi tuo, kad pirmojo lygmens paklaidos laikui bėgant koreliuoja:

$$Y_{tj} = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{ktj} + r_{tj};$$

čia $\mathbb{C}\mathrm{ov}(r_{tj}, r_{sj}) \neq 0$. Priešingai augimo modeliams, laikas į modelį tiesiogiai neįtraukimas. Visos prielaidos apie antrojo lygmens kintamuosius yra tos pačios, kaip ir sudarant bendrąjį HLM modelį. Skirtingi koreliuojančiųjų paklaidų modeliai atsiranda dėl skirtingų prielaidų apie r_{tj} kovariacijų matricą $\mathbf{R}$ (žr. 1.6 skyrelį). Kokią matricos struktūrą pasirinkti specifikuojant modelį, priklauso nuo išankstinės informacijos, duomenų grafikų.

Pagrindinės ir dažniausiai naudojamos paklaidų matricų struktūros nurodytos 2.2.1 lentelėje. Pirmajame stulpelyje yra struktūrų pavadinimai. Antrajame stulpelyje nurodyta, kiek nežinomų parametrų reikia įvertinti norint aprašyti matricą, trečiajame – kaip atrodo kovariacija $\mathbb{C}\mathrm{ov}(r_{tj}, r_{sj})$, t. y. paklaidų kovariacija laiko momentais t ir s (elementas (t, s)). Čia $t, s = 1, \ldots, T$ yra kiekvieno objekto stebėjimų skaičius. Kronekerio simbolis $\delta_{ts} = 0$, kai $t \neq s$, ir $\delta_{ts} = 1$, kai $t = s$. Paprastumo dėlei tariame, kad kiekvieno objekto matavimų skaičius yra vienodas.

2.2.1 lentelė. Matricų struktūros

Pavadinimas	Parametrų skaičius	Elementas (t, s)
Sudėtinės simetrijos (CS)	2	$\sigma_1^2 + \sigma_2^2 \delta_{ts}$
Sudėtinės simetrijos heterogeninė	$T + 1$	$\sigma_t^2 \delta_{ts} + \sigma_t \sigma_s \rho (1 - \delta_{ts})$
Autoregresinė AR(1)	2	$\sigma^2 \rho^{\lvert t-s \rvert}$
Nestruktūrizuota (UN)	$T(T+1)/2$	σ_{ts}^2

Pateiksime visų išvardytų matricų pavyzdžių, kai $T = 4$. Šiuo atveju sudėtinės simetrijos kovariacijų matrica yra

$$\mathbf{R} = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_2^2 & \sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_2^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}.$$

Matome, kad iš tikro reikia įvertinti tik du parametrus: σ_1^2 ir σ_2^2. Toks matricos pavidalas gali būti tinkamas, kai modelyje daugikliai β_{kj}, $k = 1, \ldots, K$, yra neatsitiktiniai.

Sudėtinės simetrijos heterogeninė kovariacijų matrica

$$\mathbf{R} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$$

turi 5 nežinomus parametrus. Autoregresinė kovariacijų matrica atsiranda tada, kai pirmojo lygmens paklaidos sudaro pirmosios eilės autoregresinį procesą:

$$\mathbf{R} = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{pmatrix}.$$

Galų gale nestruktūrizuota matrica reiškia, kad reikia įvertinti visas dispersijas ir visas kovariacijas:

$$\mathbf{R} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 & \sigma_{24}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 & \sigma_{34}^2 \\ \sigma_{41}^2 & \sigma_{42}^2 & \sigma_{43}^2 & \sigma_{44}^2 \end{pmatrix}.$$

Visas liekanų priklausomybes norima įvertinti kiek galima tiksliau, todėl atrodytų, kad nestruktūrizuotoji kovariacijų matrica tam geriausiai tinka. Vis dėlto būtent šiuo atveju reikia įvertinti daugiausia nežinomų parametrų. Tai atsispindės informaciniuose indeksuose *AIC* ir *BIC*, kurie didėja, kai tenka vertinti daugiau parametrų (primename, kad modelis tuo geriau tinka duomenims, kuo informacinių indeksų reikšmės mažesnės). Todėl praktiškai rekomenduotina patikrinti keletą pasirinktų liekanų priklausomybių, atsižvelgti ir į *AIC*, *BIC* reikšmes ir hipotezių apie parametrų skirtumus nuo nulio išvadas. Visa kita nesiskiria nuo standartinių HLM modelių analizės.

Mišrieji modeliai. Galimi augimo modeliai su AR(1) kovariacijų matricomis pirmojo lygmens liekanoms ir pan.

2.4. Augimo modelių analizė naudojantis SPSS programa

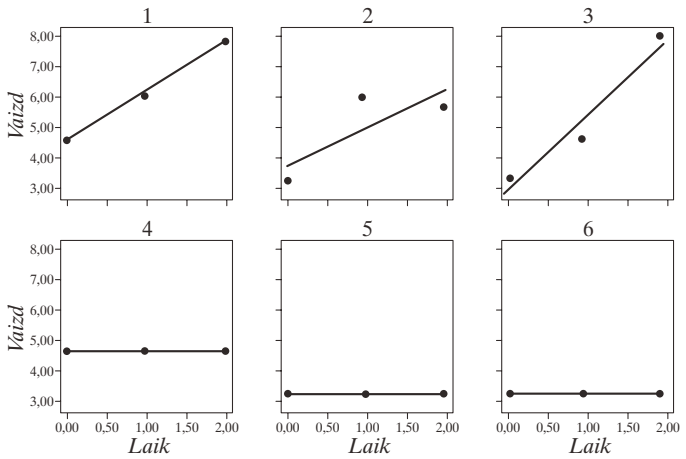
Šiuo metu kompiuterizacija primena lenktynes tarp programuotojų, kuriančių vis lengviau vartotojui suprantamas programas, ir gamtos, kuriančios vis žioplesnius vartotojus. Kol kas laimi gamta.

Šešiasdešimt paauglių iš vaikų namų dalyvavo vaizduotės ugdymo programoje. Jų vaizduotės lygis (kintamasis *Vaizd*) buvo matuojamas programos pradžioje ($Laik = 0$), po pusmečio ($Laik = 1$), ir po metų ($Laik = 2$). Be to, tyrimo pradžioje buvo įvertinta neigiama vaiką supusios aplinkos įtaka vaizduotės lavinimui (kintamasis *Apl*). Norima ištirti vaizduotės raidą bėgant laikui.

Visų pirma reikia, kad duomenys turėtų HLM analizei tinkamą formą, t. y. kiekvienam matavimui būtų skiriama atskira eilutė (žr. 2.2.4 pav.). SPSS programoje yra

	id	Lytis	Laikas	vaizd	aplinka
1	1	1,00	,00	4,47	8,22
2	1	1,00	1,00	5,94	8,22
3	1	1,00	2,00	7,88	8,22
4	2	1,00	,00	3,00	6,42
5	2	1,00	1,00	5,94	6,42
6	2	1,00	2,00	5,55	6,42
7	3	1,00	,00	3,00	10,00
8	3	1,00	1,00	4,47	10,00
9	3	1,00	2,00	8,09	10,00

2.2.4 pav. Duomenų struktūra



2.2.5 pav. Sklaidos diagramos ir regresijos tiesės

galimybė labiau įprastos struktūros (vienam respondentui – viena eilutė) duomenis atitinkamai transformuoti.

Norint geriau įsivaizduoti, kaip laikui bėgant kintamasis *Vaizd* kinta, rekomenduojama nusibraižyti atitinkamas sklaidos diagramas. Pirmųjų šešių mokinių sklaidos diagramas matome 2.2.5 paveiksle. Kartu pavaizduotos ir regresijos tiesės. Žinoma, tokias diagramas reikėtų nubraižyti visiems tirtiems mokiniams.

Matome, kad laikui bėgant galima tikėtis tiesinės vaizduotės plėtros. Todėl į modelį laikas bus įtraukiamas tiesiškai. Kaip ir ankstesniuose skyreliuose, skirtuose SPSS, SAS ir R programoms, rašydami modelį dalį indeksų praleisime. Iš pradžių suformuluosime tiriamus modelius.

Besąlyginis HLM modelis. Tradiciškai tyrimas pradedamas ignoruojant priklausomybę nuo laiko, t. y. nuo tradicinio besąlyginio modelio:

$$\begin{aligned} Vaizd &= \beta_0 + e, \\ \beta_0 &= \gamma_{00} + u_0. \end{aligned}$$

Jungtinis modelis:

$$Vaizd = \gamma_{00} + [e + u_0].$$

Jo sintaksė niekuo nesiskiria nuo aprašytos 1.9 skyrelyje.

Besąlyginis augimo modelis. Į modelį įtraukiame tiesinę priklausomybę nuo laiko:

$$\begin{aligned} Vaizd &= \beta_0 + \beta_1 Laik + e, \\ \beta_0 &= \gamma_{00} + u_0, \\ \beta_1 &= \gamma_{10} + u_1. \end{aligned}$$

Jungtinis modelis:

$$Vaizd = \gamma_{00} + \gamma_{10} Laik + [u_0 + u_1 Laik].$$

Šiuo atveju programos sintaksė yra tokia:

```
MIXED Vaizd WITH Laik
/CRITERIA = ...
/FIXED = Laik | SSTYPE(3)
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/RANDOM INTERCEPT Laik | SUBJECT(id) COVTYPE(UN) .
```

Atkreiptinas dėmesys į tai, kad *Laik* į modelį įtraukiamas kaip tolydusis kintamasis.

Sudėtingesnis augimo modelis. Į modelį įtraukiame kintamąjį *Apl*:

$$\begin{aligned} Vaizd &= \beta_0 + \beta_1 Laik + e, \\ \beta_0 &= \gamma_{00} + \gamma_{01} Apl + u_0, \\ \beta_1 &= \gamma_{10} + u_1. \end{aligned}$$

Jungtinis modelis:

$$Vaizd = \gamma_{00} + \gamma_{01} Apl + \gamma_{10} Laik + [u_0 + u_1 Laik].$$

2.2.2 lentelė. Lyginamoji augimo modelių lentelė

Modelis	Informaciniai indeksai	Fiksuoto poveikio parametrai	Atsitiktinio poveikio parametrai
Besąlyginis: $Vaizd = \beta_0 + e$ $\beta_0 = \gamma_{00} + u_0$	$-2LL = 615{,}7$ $AIC = 619{,}7$ $BIC = 626{,}1$	$\hat{\gamma}_{00} = 4{,}2^{**}$	$\hat{\sigma}^2 = 1{,}2^{**}$ $\hat{\tau}_{00} = 0{,}8^{**}$
Besąlyginio augimo: $Vaizd = \beta_0 + \beta_1 Laik + e$ $\beta_0 = \gamma_{00} + u_0$ $\beta_1 = \gamma_{10} + u_1$	$-2LL = 554{,}1$ $AIC = 562{,}1$ $BIC = 574{,}8$	$\hat{\gamma}_{00} = 3{,}6^{**}$ $\hat{\gamma}_{10} = 0{,}6^{**}$	$\hat{\sigma}^2 = 0{,}6^{**}$ $\hat{\tau}_{00} = 0{,}5^{*}$ $\hat{\tau}_{01} = 0{,}1$ $\hat{\tau}_{11} = 04^{*}$
Sudėtingesnis: $Vaizd = \beta_0 + \beta_1 Laik + e$ $\beta_0 = \gamma_{00} + \gamma_{01} Apl + u_0$ $\beta_1 = \gamma_{10} + u_1$	$-2LL = 531{,}6$ $AIC = 539{,}6$ $BIC = 552{,}3$	$\hat{\gamma}_{00} = 7{,}3^{**}$ $\hat{\gamma}_{10} = 0{,}6^{**}$ $\hat{\gamma}_{01} = -0{,}4^{**}$	$\hat{\sigma}^2 = 0{,}6^{**}$ $\hat{\tau}_{00} = 0{,}2$ $\hat{\tau}_{01} = 0{,}1$ $\hat{\tau}_{11} = 0{,}4^{**}$
Su lygmenų sąveika: $Vaizd = \beta_0 + \beta_1 Laik + e$ $\beta_0 = \gamma_{00} + \gamma_{01} Apl + u_0$ $\beta_1 = \gamma_{10} + \gamma_{11} Apl + u_1$	$-2LL = 534{,}7$ $AIC = 542{,}9$ $BIC = 555{,}3$	$\hat{\gamma}_{00} = 7{,}3$ $\hat{\gamma}_{10} = 0{,}3$ $\hat{\gamma}_{01} = -0{,}4^{**}$ $\hat{\gamma}_{11} = 0{,}02$	$\hat{\sigma}^2 = 0{,}6^{**}$ $\hat{\tau}_{00} = 0{,}2$ $\hat{\tau}_{10} = 0{,}1$ $\hat{\tau}_{11} = 0{,}4^{**}$

$^{*}p < 0{,}05$. $^{**}p < 0{,}01$.

Augimo modelis su lygmenų sąveika. Kintamąjį *Apl* įtraukiame į abi antrojo lygmens lygtis:

$$\begin{aligned} Vaizd &= \beta_0 + \beta_1 Laik + e, \\ \beta_0 &= \gamma_{00} + \gamma_{01} Apl + u_0, \\ \beta_1 &= \gamma_{10} + \gamma_{11} Apl + u_1. \end{aligned}$$

Jungtinis modelis:

$$\begin{aligned} Vaizd = {} & \gamma_{00} + \gamma_{01} Apl + \gamma_{10} Laik + \gamma_{11} Apl \cdot Laik \\ & + [u_0 + u_1 Laik + e]. \end{aligned}$$

Gauti visų pavyzdžių rezultatai pateikiami 2.2.2 lentelėje.

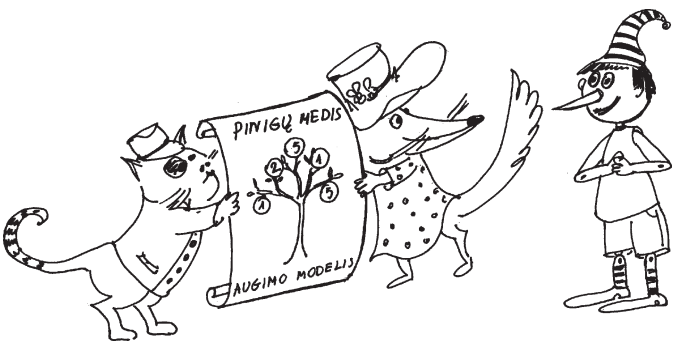
Matome, kad įtraukus į modelį laiką labai sumažėjo skirtumus tarp respondentų aprašanti dispersija $\hat{\sigma}^2$:

$$\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{1{,}2 - 0{,}6}{1{,}2} = 0{,}50.$$

Taigi galima teigti, kad 50 % respondentų skirtumų lėmė laiko faktorius. Toliau lygindami modelius ir atkreipdami dėmesį į statistiškai nereikšmingus koeficientų įverčius, nusprendžiame, kad geriausiai tinka trečiasis modelis.

2.5. Augimo modelių analizė naudojantis R programa

Dirbant R programa, augimo modelių analizei naudojama jau anksčiau minėta funkcijų bibliotekos *nlme* funkcija *lme*. Vienas iš iššūkių yra paruošti panelinius duomenis taip, kad jie būtų „suprantami“ R programai. Čia mes pateiksime aprašymą, kaip, naudojantis R programos funkcijomis, įprastos formos duomenys konvertuojami į priimtiną formatą. Čia įprasta vadinama tokia duomenų forma, kai duomenų eilutė yra vieno individo (objekto) stebėjimų aibė. Pavyzdžiui, duomenų eilutė yra lytis, pasirinkto požymio matavimas laiko momentu 1, pasirinkto požymio matavimas laiko momentu 2, pasirinkto požymio matavimas laiko momentu 3 (tarkime, laiko tarpas tarp matavimų yra mėnuo). Toks duomenų formatas neturi laiko stulpelio ir į modelį jo neįmanoma įtraukti. Ką daryti?



Funkcijų bibliotekos *multilevel* funkcija *make.univ* atlieka reikiamą duomenų sistemos transformavimą. Pirmasis privalomas funkcijos parametras – transformuojama duomenų

aibė; antrasis privalomas parametras – duomenų sistemos stulpeliai su pakartotiniais stebėjimais. Galima dar nurodyti dviejų kintamųjų (TIME ir MULTDV), kuriuos sukuria funkcija, vardus.

Pelių navikų augimo duomenų ir transformuotųjų duomenų fragmentas pateikiamas 2.2.6 paveiksle.

Kaip taikomi augimo modeliai, iliustruosime remdamiesi tokiais tyrimo duomenimis. Su 20-čia pelių buvo atliktas toks eksperimentas. Kiekvienai pelei buvo įskiepytas navikas: pirmasis įskiepytas 0-ę eksperimento dieną, antrasis – 2-ą dieną. Abiejų navikų matavimai

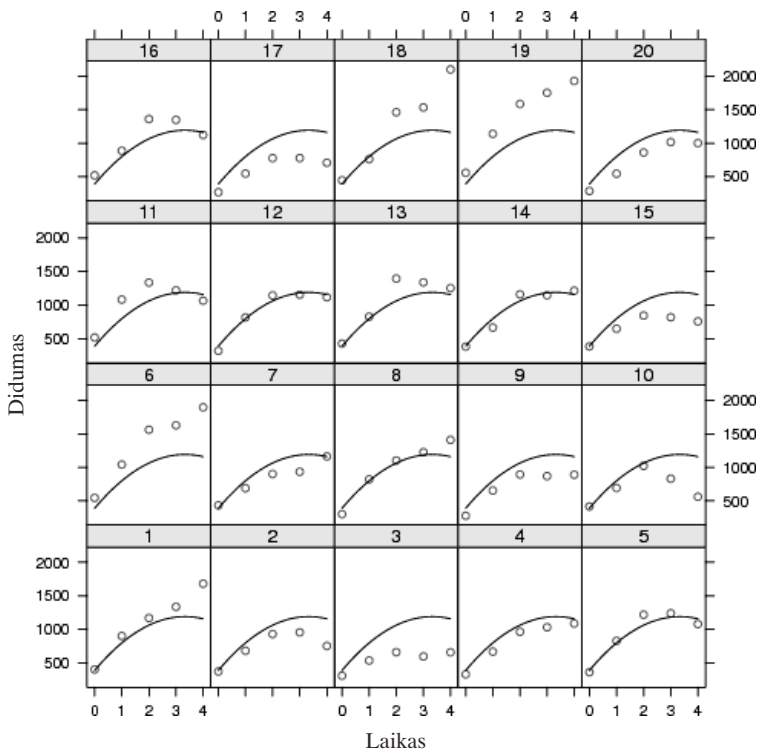
```
> y[1:3,]
  eksp nr  p1  p2   p3   p4   p5
1    0  1 404 903 1170 1337 1682
2    0  2 374 682  931  957  754
3    0  3 310 537  661  598  659
```

a) Navikų augimo duomenys

```
> univ.y=make.univ(y,y[,3:7])
> univ.y[1:9,]
   eksp nr  p1  p2   p3   p4   p5 TIME MULTDV
1     0  1 404 903 1170 1337 1682    0    404
1.1   0  1 404 903 1170 1337 1682    1    903
1.2   0  1 404 903 1170 1337 1682    2   1170
1.3   0  1 404 903 1170 1337 1682    3   1337
1.4   0  1 404 903 1170 1337 1682    4   1682
2     0  2 374 682  931  957  754    0    374
2.1   0  2 374 682  931  957  754    1    682
2.2   0  2 374 682  931  957  754    2    931
2.3   0  2 374 682  931  957  754    3    957
```

a) Transformuoti duomenys

2.2.6 pav. R programa gauta duomenų transformacija



2.2.7 pav. Pelių navikų augimo duomenys

```
> model2b<-lme(MULTDV~TIME+I(TIME^2),random=~1|nr,
+ data=y,na.action=na.omit)
> summary(model2b)
Linear mixed-effects model fit by REML
 Data: z
       AIC     BIC   logLik
  1341.827 1354.701 -665.9136

Random effects:
 Formula: ~1 | nr
         (Intercept) Residual
StdDev:    221.4399 172.0603

Fixed effects: MULTDV ~ TIME + I(TIME^2)
                 Value Std.Error DF   t-value p-value
(Intercept)   389.1129  61.34206 78  6.343329       0
TIME          481.9943  42.89199 78 11.237396       0
I(TIME^2)     -72.3786  10.28257 78 -7.038959       0
 Correlation:
              (Intr) TIME
TIME          -0.434
I(TIME^2)      0.335 -0.959

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3        Max
-2.36373720 -0.48388246 -0.04326762  0.44133989  3.69108861

Number of Observations: 100
Number of Groups: 20
```

2.2.8 pav. Netiesinio augimo modelio analizės, atliktos R programa, rezultatai

pradėti nuo 5-os dienos ir kartoti kas antrą dieną. Nuo 7-os eksperimento dienos 5 dienas iš eilės pelėms leistas vaistas.

Iš 2.2.7 paveiksle pateiktų pirmo naviko augimo grafikų matyti, kad dalies pelių naviko augimas sulėtėjo (ar net pradėjo mažėti). Tačiau ar galime tvirtinti, kad tokia tendencija statistiškai reikšminga?

Toliau pateikiamas augimo modelis duomenims tinka geriau nei nulinis ar tiesinis augimo modelis:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \varepsilon_{ij}.$$

Paklaidos yra autokoreliuotos:

$$\varepsilon_{ij} = \rho \varepsilon_{(i-1)j} + \delta_{ij}.$$

Iš pateiktų 2.2.8 paveiksle rezultatų matome, kad vaistų poveikis reikšmingas (koeficientas prie t^2 neigiamas ir statistiškai reikšmingas).

augimo modeliai	koreliuojančiųjų paklaidų modeliai	lygmenų sąveika paneliniai duomenys

UŽDAVINIAI

1. Faile **stabdymas.sav** yra automobilių stabdžių testavimo duomenys. Turime tokius kintamuosius: *autoid* – automobilio kodas; *ssustojo* – laikas, per kurį automobilis sustojo; *SPIDOMTR* – greitis pradedant stabdyti; *slapias* – kelio būklė (1 – šlapias asfaltas, 0 – sausas). Išnagrinėkite ir palyginkite du modelius.

   Pirmasis modelis:

$$\begin{aligned} ssustojo &= \beta_0 + \beta_1 SPIDOMTR + e, \\ \beta_0 &= \gamma_{00} + u_0, \\ \beta_1 &= \gamma_{10} + u_1. \end{aligned}$$

   Antrasis modelis:

$$\begin{aligned} ssustojo &= \beta_0 + \beta_1 SPIDOMTR + e, \\ \beta_0 &= \gamma_{00} + \gamma_{01} slapias + u_0, \\ \beta_1 &= \gamma_{10} + u_1. \end{aligned}$$

   Atlikite pradinę duomenų analizę ir atsakykite, kodėl kelio būklę įtraukti į modelį netikslinga.

2. Faile **st3panel.sav** pateikiami hipotetiniai duomenys apie dvidešimties respondentų alergiškumo statistikai laipsnį (kintamasis y), paskaičius vadovėlį *Statistika ir jos taikymai* 5, 6, 7, 8 ir 9 valandas. Sudarykite ir ištirkite augimo modelius su tiesiniu ir kvadratiniu trendais.

3. APIBENDRINTIEJI HIERARCHINIAI MODELIAI

Problemos formuluotė dažnai yra svarbesnė nei jos sprendimas, kuris gali priklausyti vien nuo matematinių ar eksperimentinių gebėjimų.

A. Einšteinas

Turėdami kelis vykusius modelius, mokslininkai retai kada atsispiria pagundai sudaryti vieną bendrą jungtinį modelį. Istorijoje apstu tokių modelių pavyzdžių. Garsūs jungtinių teorinių biomodelių kūrėjai buvo senovės graikai. Įžiūrėjo panašumų tarp vyro ir žirgo, sujungė ir štai – kentauras. Sujungė jautį su vyru – Minotauras. Jungtinis moters ir liūtės modelis – tai graikiškasis sfinksas, moters ir paukštės – harpija. Iš pirmo žvilgsnio tokie modeliai gal kiek keistoki, tačiau paaiškėjo, kad jie pavydėtinai atsparūs laikui.

Tokių modelių („kentaurų“ ir „harpijų“) esama ir statistikoje. Šiame skyrelyje nagrinėsime jungtinį modelį, sudarytą iš apibendrintojo tiesinio ir hierarchinio modelių. Tokie modeliai reikalingi įvairiems tyrimams:

1 Ar dalyvavimas rinkimuose priklauso nuo rinkėjo amžiaus, išsilavinimo, socialinio ir ekonominio statuso, lyties? Atsižvelgiama į rinkėjų gyvenamosios vietovės ypatybes.

2 Ar operacijos sėkmė priklauso nuo ligonio amžiaus, svorio, kraujo spaudimo, lyties? Atsižvelgiama į operaciją dariusio chirurgo patirtį.

3 Ar butų vagys, rinkdamiesi „taikinį“, atsižvelgia į vietos, aplinkos, apsaugos faktorius?

4 Ar noras pakeisti atsarginių detalių (kompiuterinės technikos ir pan.) tiekėją priklauso nuo prekės kainos? Atsižvelgiama į detales tiekiančios firmos didumą ir jos prekybos apimtis.

5 Ar draudimo išmokų skaičius priklauso nuo apdraustų automobilių eksploatavimo trukmės, apdraudusiųjų lyties ir amžiaus? Atsižvelgiama į draudimo firmos bendrąjį klientų skaičių.

Priešingai nei ankstesniuose skyreliuose, priklausomas kinamasis nėra normalusis (binominis: dalyvaus – nedalyvaus, operacija pavyks – operacija nepavyks; Puasono: išmokų skaičių skirstinys ir pan.).

Apibendrintieji HLM dar vadinami apibendrintaisiais tiesiniais mišriaisiais modeliais (GLMM[17]).

3.1. GLMM struktūra

Užduotis: *reikia išskleisti* $(a+b)^n$. *Sprendimas*: $(\ a\ +\ b\ )^{\ n}$.

Pats pavadinimas GLMM sufleruoja, kad kalbama apie apibendrintųjų tiesinių modelių (prisiminkime, kad jų trumpinys GLM) ir HLM „hibridą“. Priminsime šių dviejų modelių klasių bendrąsias matricines išraiškas.

GLM atveju tariama, kad priklausomų kintamųjų aibės $Y_1, Y_2, \ldots, Y_n$ kiekvieno nario Y_i skirstinys priklauso eksponentinių skirstinių šeimai, o priklausomybę nuo regresorių galima užrašyti taip:

$$g(\mathbf{EY}) = \mathbf{X}\boldsymbol{\gamma}.$$

Primename, kad $\mathbf{X}$ yra struktūros matrica, nusakanti, kokių kintamųjų (regresorių) įtaką priklausomiems kintamiesiems Y tiriame, $\boldsymbol{\gamma}$ yra nežinomi koeficientai[18], o užrašas $g(\mathbf{EY})$ žymi vektorių, kurio koordinatės yra $g(\mathbf{E}Y_1), g(\mathbf{E}Y_2), \ldots, g(\mathbf{E}Y_n)$.

Prisiminkime matricinę HLM išraišką:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Zu} + \mathbf{e}.$$

Matricinis GLMM užrašas yra akivaizdus abiejų modelių junginys:

$$g(\mathbf{E}[\mathbf{Y}|\mathbf{u}]) = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Zu}; \tag{2.3.1}$$

čia $\mathbf{Y}|\mathbf{u}$ žymi sąlyginį atsitiktinį dydį $\mathbf{Y}$, kai atsižvelgiama į kintamąjį $\mathbf{u}$. Jei priklausomas kintamasis $\mathbf{Y}$ turi normalųjį skirstinį, tai jungties funkcija yra tiesiog $\mathbf{g}(\mathbf{E}[\mathbf{Y}|\mathbf{u}]) = \mathbf{E}[\mathbf{Y}|\mathbf{u}]$. Tuo atveju GLMM virsta įprastu HLM. Palyginimui modeliai nurodyti 2.3.1 lentelėje.

2.3.1 lentelė. Modelių palyginimas

Modelio tipas	Vidurkio modelis	Skirstinys
GLMM	$g(\mathbf{EY}\|\mathbf{u}) = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Zu}$	$\mathbf{Y}\|\mathbf{u}$ turi eksponentinį skirstinį; $\mathbf{u}$ – normalusis
HLM	$\mathbf{EY} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Zu}$	$\mathbf{u}$; $\mathbf{Y}\|\mathbf{u}$ – normalieji kintamieji
GLM	$g(\mathbf{EY}) = \mathbf{X}\boldsymbol{\gamma}$	$\mathbf{Y}$ turi eksponentinį skirstinį
LNM	$\mathbf{EY} = \mathbf{X}\boldsymbol{\gamma}$	$\mathbf{Y}$ – normalusis kintamasis

GLMM galima užrašyti ir lygčių sistemomis. Tarkime, kad yra $j = 1, 2, \ldots, J$ grupių ir Y_{ij}, $i = 1, 2, \ldots, n_j$ stebėjimų. Kintamieji Y_{ij} priklauso nuo X_k, $k = 1, 2, \ldots, K$, ir turi vidurkius $\mathbf{E}Y_{ij} = \mu_{ij}$. Pirmojo lygmens lygtys yra tokios:

$$g(\mu_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{Kj}X_{Kij}; \tag{2.3.2}$$

[17] GLMM – angl. *Generalized Linear Mixed Models*.

[18] Taip nežinomus koeficientus $\boldsymbol{\beta}$ pažymėjome todėl, kad būtų galima lengviau GLMM palyginti su HLM.

čia X_{kij} žymi j-osios grupės i-ojo elemento k-ojo kintamojo reikšmę. Kiekvienas iš $K+1$ pirmojo lygmens modelio parametrų (β) antrajame lygmenyje traktuojamas kaip priklausomas kintamasis. Antrojo lygmens lygtys:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \cdots + \gamma_{0S} W_{Sj} + u_{0j}, \\ \beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + \cdots + \gamma_{1S} W_{Sj} + u_{1j}, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \beta_{Kj} = \gamma_{K0} + \gamma_{K1} W_{1j} + \cdots + \gamma_{KS} W_{Sj} + u_{Kj}. \end{cases} \tag{2.3.3}$$

Tos pačios grupės antrojo lygmens paklaidos gali koreliuoti: $\mathbb{C}\mathrm{ov}(u_{kj}, u_{mj}) = \tau_{km}$. Kaip ir HLM atveju, aprašėme šiek tiek suprastintą bendrojo modelio variantą, laikydami visus β priklausančiais nuo tų pačių kintamųjų W_{sj}. Sudarant modelius, nebūtina į kiekvieną lygtį įtraukti *visus* W_{sj}. Pavyzdžiui, galima sudaryti modelį, kuriame β_{0j} priklausys tik nuo W_{1j}, o β_{1j} – tik nuo W_{2j}.

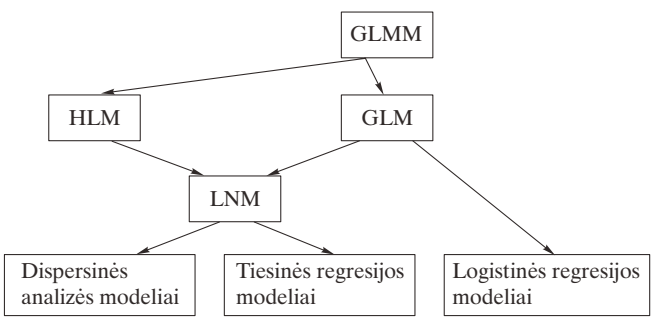
Pastaba. Apibendrintieji HLM ir apibendrintieji tiesiniai mišrieji modeliai statistinėje literatūroje ne visada yra tapatinami. GLMM – tai apibendrintasis HLM, kai **u** turi normalųjį skirstinį. Tik apie tokius modelius ir kalbama šiame skyrelyje.

Apibendrintųjų tiesinių modelių hierarchija pateikta 2.3.1 paveiksle. Matome, kad GLMM yra tiek GLM, tiek ir HLM apibendrinimas.

Pastebėsime, kad įmanomas ir kitoks, paprastesnis GLM apibendrinimas, kai į hierarchinę struktūrą neatsižvelgiama. Tiesiog nurodoma, kad Y_{ij} koreliuoja su Y_{ik}, kai $j \neq k$ (tiksliau – pateikiama šių kintamųjų kovariacija). Tokie modeliai vadinami apibendrintosiomis vertinimų lygtimis (GEE[19]). Mūsų nagrinėjamo uždavinio GEE modelis yra toks:

$$\ln\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \gamma_{00} + \gamma_{10} x_{ij}, \qquad \mathbb{C}\mathrm{ov}(Y_{ij}, Y_{ik}) = \alpha, \quad j \neq k.$$

Matome, kad GEE yra logistinės regresijos modelis, kuriame priklausomo kintamojo stebėjimai koreliuoja grupių viduje. Modelyje pasirinkta pati paprasčiausia situacija, kai visų kintamųjų porų kovariacija vienoda.



2.3.1 pav. Modelių hierarchija

[19] GEE – angl. *Generalized Estimating Equations*.

3.2. Hierarchinė logistinė regresija

Tie, kurie žino, – neprognozuoja. Tie, kurie prognozuoja, – nežino.
Lao Dzė

Apsiribosime tik atskiru tokių modelių atveju – hierarchine logistine regresija su atsitiktiniais koeficientais. Aprašysime paprastą modelį, kai yra tik vienas antrojo lygmens kintamasis Z, tik vienas pirmojo lygmens kintamasis X, o atsitiktinis – tik postūmis.

Tarsime, kad stebimas priklausomas kintamasis turi sąlyginį binominį skirstinį:

$$Y_{ij}|u_{0j} \sim Bi(n_{ij}, \pi_{ij});$$

čia $u_{0j} \sim \mathcal{N}(0, \tau_{00})$ – atsitiktinis poveikis, $j = 1, 2, \ldots, N$. Užrašysime dviejų lygmenų GLMM. Pirmojo lygmens lygtis sutampa su logistinės regresijos modelio lygtimi:

$$\text{logit}(\pi_{ij}) = \ln\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \beta_{0j} + \beta_{10}X_{ij}.$$

Antrojo lygmens lygtys nesiskiria nuo HLM lygčių:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}, \\ \beta_{1j} = \gamma_{10}. \end{cases}$$

Jungtinis (mišrusis) modelis:

$$\text{logit}(\pi_{ij}) = \ln\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{10}X_{ij} + u_{0j}. \tag{2.3.4}$$

Nesunku suprasti, kad analogiškai užrašomi ir sudėtingesni (paprastesni) modeliai.

2.3.1[20] **pavyzdys.** Tyrimo tikslas – palyginti dviejų medicininių procedūrų šalutinių poveikių skaičius. Tyrimui duomenys pateikti iš 15 atsitiktinai parinktų medicinos centrų. Natūrali prielaida, kad šalutinių poveikių tikimybės priklauso nuo procedūros tipo (A arba B). Be to, suprantama, kad medicinos centrai nėra vienodi (skirtinga personalo kvalifikacija, nevienodai laikomasi higienos normų ir t. t.). Logiška medicinos centrų poveikį modelyje laikyti atsitiktiniu, nes: a) parinkta *tik* 15 centrų iš visos populiacijos (atranką pakartojus poveikio įverčiai bus kiti), b) personalų kvalifikacija nevienoda. Apie ligoninės lygmenyje veikiančius veiksnius nieko konkretaus nežinoma, todėl antrajame lygmenyje aiškinamųjų kintamųjų nėra.

Centras	1	1	2	2	3	3	4	4	5	5
Procedūra	A	B	A	B	A	B	A	B	A	B
Procedūrų skaičius	32	23	30	28	23	24	22	22	20	21
Šalutinių poveikių skaičius	14	18	4	8	14	9	7	10	6	12
Centras	6	6	7	7	8	8	9	9	10	10
Procedūra	A	B	A	B	A	B	A	B	A	B
Procedūrų skaičius	19	20	17	17	16	15	13	14	13	13
Šalutinių poveikių skaičius	1	3	2	6	7	9	1	5	3	1
Centras	11	11	12	12	13	13	14	14	15	15
Procedūra	A	B	A	B	A	B	A	B	A	B
Procdūrų skaičius	11	12	10	9	9	9	8	8	7	8
Šalutinių poveikių skaičius	1	2	1	0	2	6	1	1	1	0

[20] Pavyzdys paimtas iš: Oliver Schabenberger, *Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models*. SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. *SUGI 30 Proceedings*. Philadelphia, Pennsylvania. April 10–13, 2005.

Pažymėkime dydžiu Y_{Aj} skaičių tų j-ojo centro pacientų, kuriems, atlikus procedūrą A, buvo užfiksuotas šalutinis poveikis. Analogiškai, tegul Y_{Bj} – skaičius tų j-ojo centro pacientų, kuriems šalutinis poveikis užfiksuotas po procedūros B. Kintamieji $Y_{Aj} \sim Bi(n_{Aj}, \pi_{Aj})$ ir $Y_{Bj} \sim Bi(n_{Bj}, \pi_{Bj})$, $j = 1, 2, \ldots, 15$, turi binominius skirstinius. Aišku, kad π_{Aj} yra pašalinių poveikių po A procedūros tikimybė j-ojoje ligoninėje. Analogiškai interpretuojama π_{Bj}.

Norėdami modelį užrašyti kompaktiškai, apibrėšime pseudokintamąjį x_{ij} – atliktos procedūros indikatorių, nurodantį, kuri procedūra (A ar B), buvo atlikta:

$$x_{ij} = \begin{cases} 1, & i = A, \\ 0, & i = B. \end{cases}$$

Dabar modelį galima užrašyti taip:

$$\ln\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \beta_{0j} + \beta_{10}x_{ij}, \qquad \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j}, \\ \beta_{1j} = \gamma_{10}. \end{cases}$$

Jungtinis (mišrusis) modelis:

$$\ln\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j}.$$

Sudarius modelį, reikia įvertinti jo nežinomus parametrus ir išsiaiškinti, ar jis tinka duomenims.

3.3. GLMM parametrų įverčiai

Šviesiausias matematiko gyvenimo momentas būna ką tik įrodžius svarbią teoremą ir dar neradus įrodyme klaidos.

Kaip ir įprastų modelių, logistinės regresijos mišriojo modelio (hierarchinės logistinės regresijos modelio su atsitiktiniais koeficientais) tyrimo etapai yra tokie: a) įvertinami parametrai, b) tikrinamos statistinės hipotezės apie parametrų reikšmes, c) įvertinamas bendrasis modelio tinkamumas duomenims. Neturime tikslo išsamiai visų etapų aprašyti. Iš esmės tai būtų ankstesnių HLM, skirtų statistinėms išvadoms formuluoti, pakartojimas. Apsiribosime tik keliomis pastabomis.

Įvertinti GLMM modelių parametrus yra daug sudėtingiau nei GLM ar HLM modelių. Trumpai supažindinsime tik su pagrindiniais įverčių radimo principais. Visi GLMM parametrų įverčių skaičiavimo metodai pagrįsti didžiausiojo tikėtinumo funkcijos logaritmo (logtikėtinumo funkcijos) optimizavimu. Tačiau galima būtų išskirti du pagrindinius optimizavimo metodų tipus:

1 Metodai, kuriuose naudojamas modelio linearizavimas.

2 Metodai, grįsti aproksimuotos (pseudo) tikėtinumo funkcijos optimizavimu.

Modelio linearizavimas reiškia, kad $g(\mathbf{EY})$ pakeičiama tiesine Teiloro formulės dalimi, o toliau tiriamas tiesinis pseudomodelis. Parametrai įvertinami pažingsniui – gaunami pradiniai įverčiai, sudaromas naujas pseudomodelis, vėl skaičiuojami parametrų įverčiai ir procedūra tęsiama tol, kol parametrų įverčių pokyčiai tampa pakankamai maži. Paminėsime du dažniausiai naudojamus modelių linearizavimo ir parametrų įverčių skaičiavimo metodus: a) svorinį kvazitikėtinumo (PQL[21]) metodą ir b) ribinį kvazitikėtinumo (MQL[22]) metodą. PQL užtikrina tikslesnę pradinio modelio aproksimaciją, tačiau ne visada pavyksta įverčius apskaičiuoti (procedūra gali nekonverguoti).

Alternatyvūs parametrų įverčių skaičiavimo metodai grindžiami tuo, kad optimizuojama ne pati logtikėtinumo funkcija, o paprastesnės struktūros funkcija. GLMM modelių tikėtinumo funkcijos išraiška gana sudėtinga. Iš esmės tai daugialypis integralas:

$$L(y) = \prod_{j=1}^{N} \int \prod_{i=1}^{n_j} f_{y_{ij}|\mathbf{u}}(Y_{ij}|\mathbf{u})\, f_u(\mathbf{u})\mathrm{d}\mathbf{u};$$

čia $f_{y_{ij}|\mathbf{u}}(Y_{ij}|\mathbf{u})$ – ribinis priklausomo kintamojo Y_{ij} skirstinio tankis (kurio išraiška yra sudėtinga), $f_u(\mathbf{u})$ – atsitiktinių poveikių skirstinio tankis.

2.3.2 pavyzdys. Nagrinėto 2.3.1 pavyzdyje modelio didžiausiojo tikėtinumo funkcija yra tokia:

$$L = \int \cdots \int \prod_{j=1}^{N} \prod_{i=1}^{n_j} \frac{\exp\{\gamma_{00} + \gamma_{10}x_{ij} + u_{0j}\}}{1 + \exp\{\gamma_{00} + \gamma_{10}x_{ij} + u_{0j}\}} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{u_{0j}^2}{2\sigma^2}} \mathrm{d}u_{01} \cdots \mathrm{d}u_{0n}.$$

Tikėtinumo funkcijos integralo negalima apskaičiuoti tiesiogiai (išskyrus išskirtinius atvejus). Todėl tikėtinumo funkcijos skaičiavimuose naudojamas integralo aproksimavimas (pseudotikėtinumo funkcija). Integralo aproksimavimui taikomi skaitinio integravimo metodai. Vienas populiaresnių yra Gauso adaptyvusis kvadratūrinis (AGQ[23]) metodas. Taip pat galima naudoti Laplaso aproksimaciją. Pastebėsime, kad skaičiavimai šiais metodais užima daug laiko. Didėjant integralo matmenų skaičiui, aproksimavimas praktiškai tampa neįmanomas.

Kokie GLMM įverčiai skaičiuojami? Takime, kad nagrinėjame modelį, aprašytą (2.3.2) ir (2.3.3) lygtimis. Primename, kad $\tau_{km} = \mathbb{C}\mathrm{ov}(u_{kj}, u_{mj})$.

GLMM parametrų įverčiai:

1) fiksuoto poveikio parametrai $\hat{\gamma}_{kj}$;
2) atsitiktinio poveikio parametrai $\hat{\tau}_{kj}$;
3) pirmojo lygmens atsitiktiniai koeficientai $\hat{\beta}_{mj}$;
4) antrojo lygmens liekamosios paklaidos $\hat{u}_{kj}$.

[21] PQL – angl. *Penalized Quasi-Likelihood*.

[22] MQL – angl. *Marginal Quasi-Likelihood*.

[23] AGQ – angl. *Adaptive Gaussian Quadrature*.

Pažymėtina, kad tuo atveju, kai GLMM virsta GLM, apskaičiuojamas ir liekamosios paklaidos e dispersijos įvertis $\hat{\sigma}^2$. Visų parametrų įverčių interpretacija nesiskiria nuo naudotos HLM modeliuose.

2.3.3 pavyzdys. Nagrinėto 2.3.1 pavyzdyje modelio parametrų įverčiai, gauti naudojant SAS programos procedūrą PROC GLIMMIX (analizės programėlės aprašytos kitame skyrelyje), pateikiami 2.3.2 paveiksle. Ligoninės lygmens paklaidos dispersijos įvertis $\hat{\tau}_{00} = 0{,}6176$ yra naudingas tik lyginant du modelius arba „iš akies“ vertinant ligoninės įtakos didumą. Vizualiai vertindami $\hat{\tau}_{00}$, įsitikiname, kad dispersijos įverčio reikšmė yra beveik to paties didumo, kaip ir parametrų $\hat{\gamma}_{00}$ ir $\hat{\gamma}_{10}$ įverčiai. Tokį paklaidos dispersijos didumą interpretuojame kaip įrodymą, kad ligoninių rezultatai skiriasi dėl tam tikrų modelyje neatspindėtų dalykų. Darome išvadą, kad modelyje trūksta ligonines apibūdinančių kintamųjų (antrojo lygmens kintamųjų).

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
$\hat{\tau}_{00}$ Intercept	center	0.6176	0.3181

Solutions for Fixed Effects

Effect	group	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	$\hat{\gamma}_{00}$	-0.8071	0.2514	14	-3.21	0.0063
group	$\hat{\gamma}_{10}$ A	-0.4896	0.2034	14	-2.41	0.0305
group	B	0	.	.	.	.

2.3.2 pav. Parametrų įverčiai

Žinodami parametrų įverčius, galime užrašyti tikimybės, kad po procedūros A bus pašalinių reiškinių (pažymėkime šią tikimybę π_A), įvertį:

$$\hat{\pi}_A = \frac{\exp\{\hat{\gamma}_{00} + \hat{\gamma}_{10}\}}{1 + \exp\{\hat{\gamma}_{00} + \hat{\gamma}_{10}\}} = \frac{\exp\{-0{,}8071 - 0{,}4896\}}{1 + \exp\{-0{,}8071 - 0{,}489\}} = 0{,}2147.$$

Analogiškai, pažymėję simboliu π_B tikimybę, kad pašalinių reiškinių bus po procedūros B, gauname

$$\hat{\pi}_B = \frac{\exp\{\hat{\gamma}_{00}\}}{1 + \exp\{\hat{\gamma}_{00}\}} = \frac{\exp\{-0{,}8071\}}{1 + \exp\{-0{,}8071\}} = 0{,}3085.$$

Įvertinome pašalinių poveikių tikimybes turėdami omenyje *visą* ligoninių populiaciją (beje, gavome, kad ši tikimybė didesnė po procedūros B).

Norėdami įvertinti pašalinių poveikių tikimybes konkrečioje ligoninėje, turėtume apskaičiuoti įverčius $\hat{u}_{Aj}$ ir $\hat{u}_{Bj}$. Tada galima būtų pasinaudoti formulėmis:

$$\hat{\pi}_{Aj} = \frac{\exp\{\hat{\gamma}_{00} + \hat{\gamma}_{10} + \hat{u}_{Aj}\}}{1 + \exp\{\hat{\gamma}_{00} + \hat{\gamma}_{10} + \hat{u}_{Aj}\}}, \qquad \hat{\pi}_{Bj} = \frac{\exp\{\hat{\gamma}_{00} + \hat{u}_{Bj}\}}{1 + \exp\{\hat{\gamma}_{00} + \hat{u}_{Bj}\}}.$$

2.3.4 pavyzdys. Statistine SAS programa apskaičiuotos šalutinių poveikių tikimybės kiekvienoje ligoninėje pateikiamos 2.3.3 paveiksle. Matome, kad tikimybė sulaukti pašalinių reiškinių po procedūros A pirmoje ligoninėje įvertinta taip:

$$\pi_{A1} = 0{,}40773.$$

Analogiškai, $\pi_{B1} = 0{,}52902$ ir t. t.

Liekamųjų paklaidų įverčiai $\hat{u}_{Aj}$ ir $\hat{u}_{Bj}$ kiekvienai ligoninei pateikiami 2.3.4 paveiksle (pateikiamas tik SAS programa gauto rezultato fragmentas). Matome, kad $\hat{u}_{A1} = 0{,}9233$, $\hat{u}_{B1} = -0{,}2573$ ir t. t.

Obs	center	group	n	SideEffect	PredProb	PredProb_PA
1	1	A	32	14	0.40773	0.21474
2	1	B	33	18	0.52902	0.30852
3	2	A	30	4	0.17452	0.21474
4	2	B	28	8	0.25647	0.30852
.	.	.	. . .		. . .	. . .
27	14	A	8	1	0.14126	0.21474
28	14	B	8	1	0.21160	0.30852
29	15	A	7	1	0.11501	0.21474
30	15	B	8	0	0.17494	0.30852

2.3.3 pav. Tikimybių prognozės

Solution for Random Effects

Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	center 1	0.9233	0.3198	14	2.89	0.0119
Intercept	center 2	-0.2573	0.3591	14	-0.72	0.4854
. . .	. . .	. . .	. . .		. . .	. . .
Intercept	center 14	-0.5082	0.5233	14	-0.97	0.3479
Intercept	center 15	-0.7440	0.5501	14	-1.35	0.1977

2.3.4 pav. Atsitiktiniai poveikiai

3.4. GLMM hipotezių tikrinimas ir tinkamumas duomenims

Japonai maistui vartoja nedaug riebalų, geria nedaug raudonojo vyno ir rečiau nei britai ar amerikiečiai patiria infarktą. Kita vertus, prancūzai vartoja daug riebalų, geria daug raudonojo vyno ir taip pat rečiau nei britai ar amerikiečiai patiria infarktą. Išvada: infarkto riziką didina ne maistas, o anglų kalba.

GLMM yra GLM ir HLM hibridas, todėl ir tikrinamų hipotezių sąrašas sutampa su sąrašu, sudarytu iš GLM ir HLM atitinkamų skyrelių. Paminėsime tik keletą dažniausiai tikrinamų hipotezių.

Statistinės hipotezės formuluojamos apie:

1) fiksuoto poveikio parametrus γ_{kj};
2) atsitiktinio poveikio parametrus ir τ_{kj};
3) atsitiktinius koeficientus β_{kj}.

Hipotezių formuluotės (o ir rezultatų interpretavimas) nesiskiria nuo aprašytų I dalies 2.5 skyrelyje ir šios dalies 2.8 skyrelyje.

2.3.5 pavyzdys. Tarkime, turime modelį, aprašytą 2.3.1 pavyzdyje. Norime sužinoti: a) ar šalutinių poveikių galimybė priklauso nuo taikytos medicininės procedūros tipo; b) ar galima sakyti, kad skirtinguose centruose šalutinių poveikių galimybės varijuoja. Atsakymus į šiuos klausimus gausime patikrinę tokias statistines hipotezes:

$$A\colon \begin{cases} H_0\colon\ \gamma_{10} = 0, \\ H_1\colon\ \gamma_{10} \neq 0, \end{cases} \qquad B\colon \begin{cases} H_0\colon\ \tau_{00} = 0, \\ H_1\colon\ \tau_{00} > 0. \end{cases}$$

Hipotezės A tikrinimo rezultatą matome 2.3.2. paveiksle. Kadangi $p = 0{,}0305 < 0{,}05$, tai darome išvadą, jog γ_{10} statistiškai reikšmingai skiriasi nuo 0. Vadinasi, šalutinių poveikių galimybė statistiškai reikšmingai priklauso nuo taikytos medicininės procedūros tipo.

Tests of Covariance Parameters
Based on the Residual Pseudo-Likelihood

Label	DF	-2 Res Log P-Like	ChiSq	Pr > ChiSq	Note
Atstikt. poveik. dispersija	1	108.26	26.81	<.0001	MI

2.3.5 pav. Atsitiktinių poveikių dispersijos svarba

Hipotezės apie medicinos centrų skirtumus rezultatai pateikti 2.3.5 paveiksle. Kadangi $p < 0{,}0001$, tai atsitiktinių poveikių įtraukimas į modelį turi prasmę. Šį rezultatą galima interpretuoti ir kaip rodiklį, kad ligoninių įtaka šalutinių poveikių atsiradimui po procedūrų yra statistiškai reikšminga.

Norint įvertinti bendrąjį modelio tinkamumą duomenims, galima naudotis informacinių indeksų (Akaikės indekso *AIC*, *BIC* ir pan.) reikšmėmis. Tačiau šios reikšmės prasmingos tik lyginant du modelius (mažesnė reikšmė rodo geresnį modelio tinkamumą duomenims). Be to, reikėtų prisiminti, kad lyginti galima tik tokius modelius, kurių vienas turi visus kito modelio regresorius. Kita alternatyva modelio tinkamumui įvertinti tokia: reikia atsižvelgti į liekamųjų paklaidų kvadratų sumą, t. y. į chi kvadrato statistikos reikšmę. Modelis gerai tinka duomenims, jeigu ši reikšmė statistiškai nereikšminga. Vis dėlto, kai duomenų daug, beveik visos statistikos būna statistiškai reikšmingos. Todėl dažniau taikomas daug švelnesnis kriterijus – chi kvadrato statistikos, padalytos iš laisvės laipsnių, reikšmė lyginama su vienetu.

GLMM pakankamai gerai tinka duomenims, jeigu chi kvadrato ir laisvės laipsnių santykis nedaug viršija arba yra mažesnis už vienetą.

Žinoma toks apibrėžimas gana netikslus, nes „nedaug“ – miglota sąvoka. Bet kuriuo atveju svarbu, kad minėtas santykis būtų mažesnis už du.

Fit Statistics

-2 Res Log Pseudo-Likelihood	81.44
Generalized Chi-Square	30.69
Gener. Chi-Square / DF	1.10

2.3.6 pav. Suderinamumo matai

2.3.6 pavyzdys. Nagrinėtame 2.3.1 pavyzdyje chi kvadrato statistika statistiškai reikšmingai skiriasi nuo nulio, tačiau statistikos ir laisvės laipsnių santykis lygus 1,10 (žr. 2.3.6 pav.). Šis santykis tikrai nedaug viršija vienetą, todėl darome išvadą, kad pavyzdžio modelis pakankamai gerai tinka duomenims.

3.5. Modelių analizė naudojantis statistinėmis programomis

Pasaulio darbininkai ir profesiniai darbuotojai greitai pasidalys į dvi skirtingas grupes. Tuos, kurie kontroliuos kompiuterius, ir tuos, kurie bus kompiuterių kontroliuojami. Jums geriau būtų patekti į pirmąją grupę.

L. D. Eigenas

Kaip matome iš GLMM aprašymo, jei $\mathbf{u} = \mathbf{0}$, tai modelis virsta apibendrintuoju tiesiniu modeliu (GLM). Taigi ir šiame skyrelyje aprašomos procedūros atskiru atveju gali būti naudojamos GLM analizei. Kurį iš modelių rinktis, turėtų paaiškėti iš analizės rezultatų. Jei atsitiktinio poveikio parametrai nėra statistiškai reišmingi, tai vargu ar galima kalbėti apie GLMM. Geriau tada apsiriboti GLM. Autorių žiniomis, SPSS programoje (iki 17 versijos) GLMM analizės procedūrų nėra.

SAS programoje (9.2 versijoje) GLMM analizei yra naudojama procedūra PROC GLIMMIX. Jos sintaksė yra labai panaši į procedūros PROC MIXED sintaksę. Procedūros sakiniai ir jų paskirtys žinomos iš ankstesnių skyrių. Pavyzdžiui, RANDOM sakinys „atkeliavo“ iš PROC MIXED – jame išvardijami atsitiktiniai poveikiai; MODEL sakiniu aprašome modelį, CONTRAST sakinyje nurodome, kokias hipotezes apie parametrų tiesinius darinius tikrinsime ir pan. Manome, kad čia pateikiama 2.3.1 pavyzdžio analizei skirtos procedūros PROC GLIMMIX sintaksė leis skaitytojui geriau suprasti PROC GLIMMIX ypatumus ir galimybes.

```
PROC GLIMMIX DATA=MULTICENTER;
CLASS CENTER GROUP;
MODEL SIDEEFFECT/N = GROUP / SOLUTION;
RANDOM INTERCEPT / SUBJECT=CENTER SOLUTION;
LSMEANS GROUP / CL ILINK;
COVTEST 'ATSTIKT. POVEIK. DISPERSIJA' ZEROG;
OUTPUT OUT=GLIMMIXOUT PRED( BLUP ILINK)=PREDPROB
PRED(NOBLUP ILINK)=PREDPROB_PA;
RUN;
PROC PRINT DATA=GLIMMIXOUT;
RUN;
```

Dalis skaičiavimo rezultatų buvo pateikta 2.3.2 ir 2.3.6 paveiksluose. Pastebėsime, kad hipotezėms apie atsitiktinių poveikių kovariacijų matricos pavidalą tikrinti naudojamas sakinys COVTEST.

Sakinys LSMEANS naudojamas grupių vidurkiams apskaičiuoti (tiksliau, $\hat{\gamma}_{00}$ ir $\hat{\gamma}_{00} + \hat{\gamma}_{10}$). Nagrinėtame pavyzdyje šie vidurkiai yra $-1{,}2966$ ir $-0{,}8071$ (žr. 2.3.7 pav.). Šiame paveiksle taip pat pateikti pirmojo lygmens grupių vidurkiai. Šiuo, hierarchinės logistinės regresijos, atveju jie sutampa su šalutinių poveikių po procedūrų A ir B tikimybėmis. Matome, kad $\hat{\pi}_A = 0{,}2147$, $\hat{\pi}_B = 0{,}3085$. Primename, kad šias tikimybes jau apskaičiavome skyrelyje, skirtame parametrų įverčiams.

The GLIMMIX Procedure

group Least Squares Means

group	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
A	-1.2966	0.2601	14	-4.99	0.0002	0.05	-1.8544	-0.7388
B	-0.8071	0.2514	14	-3.21	0.0063	0.05	-1.3462	-0.2679

Mean	Standard Error Mean	Lower Mean	Upper Mean
0.2147	0.04385	0.1354	0.3233
0.3085	0.05363	0.2065	0.4334

2.3.7 pav. Grupių vidurkiai

Sakinys OUTPUT panaudotas, norint sužinoti šalutinių poveikių tikimybių prognozes kievienam stebėjimui. Rezultatus jau aptarėme 2.3.4 pavyzdyje (žr. 2.3.3 pav.).

R programoje GLMM analizei yra naudojama funkcijų bibliotekos *lme4* funkcija *lmer*. Jos sintaksė tokia:

```
lmer(formula, data, family, method, control, start,
subset, weights, na.action, offset, contrasts,
model, ...)
```

Matome, kad šios funkcijos argumentų sąrašas yra įprastas. Vienintelis privalomas argumentas yra *formula*. Žinomi du metodai, kuriuos galima pasirinkti GLMM parametrų įverčių skaičiavimui: PQL ir Laplaso. Žemiau pateikiama programėlė, pagal kurią gaunami tokie pat poveikių įverčiai, kaip ir naudojantis procedūra SAS GLIMMIX.

```
library(lme4)
mdl = lmer(cbind(SideEffect,n) ~ group + (1 | center), data = multicenter,
family="binomial"(link="logit",method="PQL"))
ranef(mdl) # atsitiktinių poveikių reikšmės
```

Įdomu, kokius rezultatus gautume pasirinkę mūsų pavyzdžiui GEE modelį. Tam, kad lengviau būtų galima palyginti rezultatus, modelius užrašysime pasirinkę skirtingus koeficientų žymėjimus. Tarkime, kad GLMM modelis yra

$$\text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j},$$

o GEE užrašomas taip:

$$\text{logit}(\pi_{ij}) = \alpha_{00} + \alpha_{10}x_{ij}.$$

Žinoma, kad:

- $|\hat{\alpha}_{10}| < |\hat{\gamma}_{10}|$;
- jei $\tau_{00} = 0$, tai $\hat{\alpha}_{10} = \hat{\gamma}_{10}$;
- jei $\tau_{00} >> 0$, tai $\hat{\alpha}_{10}$ ir $\hat{\gamma}_{10}$ skirtumai bus labai dideli.

Pirmiausia užrašykime, kaip atrodo SPSS, SAS ir R programose GEE analizei naudojamų procedūrų sintaksė. SPSS programoje ji tokia:

```
* SPSS.
* Generalized Estimating Equations.
GENLIN SideEffect OF n BY group (ORDER=ASCENDING)
/MODEL group INTERCEPT=YES      DISTRIBUTION=BINOMIAL LINK=LOGIT

/CRITERIA ANALYSISTYPE=3(WALD)
/EMMEANS TABLES=group SCALE=ORIGINAL
/REPEATED SUBJECT=center SORT=YES CORRTYPE=UNSTRUCTURED
ADJUSTCORR=YES COVB=ROBUST
/MISSING CLASSMISSING=EXCLUDE
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION workingcorr.
```

SAS programoje naudotini tokie sakiniai:

```
/*SAS GEE*/
PROC GENMOD DATA=EXCL.CENTER;
CLASS CENTER GROUP;
MODEL SIDEEFFECT/N = GROUP / D=BIN LINK=LOGIT;
REPEATED SUBJECT=CENTER / TYPE=CS;
RUN;
```

R programa, skirta GEE, yra tokia:

```
# R GEE
library(gee)
a.ml <- gee(cbind(SideEffect,n) ~ group ,id= center, data = multicenter,
family="binomial",corstr="unstructured")
```

Kokias išvadas galima padaryti pažiūrėjus į SAS programa gautų rezultatų lenteles (žr. 2.3.8 pav.)?

GEE fiksuotų poveikių įverčių reikšmės absoliuočiuoju didumu yra mažesnės nei

Exchangeable Working Correlation

$\hat{\alpha}$ **Correlation** 0.6292499521

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-0.5958	0.2092	-1.0058	-0.1859	-2.85	0.0044
group	A	-0.4520	0.2102	-0.8640	-0.0400	-2.15	0.0315
group	B	0.0000	0.0000	0.0000	0.0000	.	.

2.3.8 pav. GEE analizės rezultatai

analogiškos logistinės regresijos mišraus modelio. Be to, įverčių skirtumai yra nedideli (nes $\hat{\tau}_{00}$ reikšmė nedaug skiriasi nuo 0).

Jeigu tyrimo tikslas yra įvertinti šalutinių poveikių skirtumus, tuomet galima rinktis tiek GLMM, tiek ir GEE modelius bei atitinkamas procedūras. Jeigu norima išsiaiškinti skirtumus tarp medicinos centrų (galbūt norint pratęsti tyrimą ir išsiaiškinti, kodėl šie skirtumai egzistuoja), tai verčiau rinktis GLMM.

GEE	hierarchinis apibendrintas	hierarchinė logistinė
GLMM	tiesinis modelis	regresija

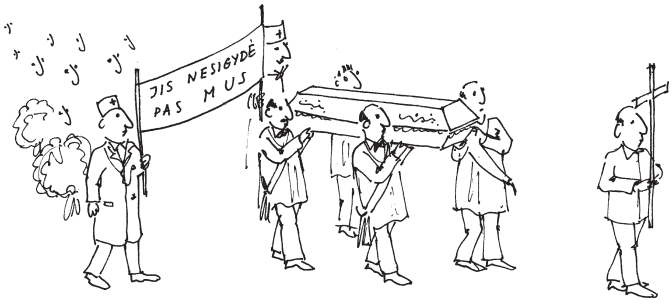
UŽDAVINIAI

1. Pritaikykite 2.3.1 pavyzdžio duomenims logistinės regresijos modelį. Paanalizuokite, ar gauti rezultatai labai skiriasi nuo GLMM ir GEE rezultatų? Kokias išvadas galima padaryti?

2. Tiriame VU Matematikos ir informatikos fakulteto studentus. Faile **hlmlogistine.sav** yra tokie kintamieji: *ATESTVID* – atestato vidurkis, *PRGR* – studijų programos kodas (ekonometrija, finansų ir draudimo matematika, informatika, matematika ir matematikos taikymai, matematikos ir informatikos mokymas, programų sistemos, statistika), *GRUPD* – grupės numeris, *BUSENID* – studijų rezultatas (1 – baigė, 0 – nebaigė), *LYTIS* – studento lytis (1 – V, 2 – M), *IVERTINIM* – pirmo kurso matematinės analizės egzamino pažymys, *nr* – dvireikšmis kintamasis (1 – studijuojamoji programa pateko tarp 5 pirmųjų pasirinkimų, 0 – kitais atvejais). Sudarykite hierarchinį logistinės regresijos modelį kintamajam *BUSENID*. Palyginkite su nehierarchinės logistinės regresijos modeliu.

STATISTINIAI PRIEŽASTINĖS ANALIZĖS MODELIAI

3



Situacija, kai priežastis eina po pasekmės, galima tik tada, kai gydytojas eina paskui paciento karstą.

Ši knygos dalis skiriama kintamųjų priklausomybių analizei. Aptarsime modelius, kurie dažniausiai taikomi *spėjamiems* priežastiniams ryšiams tirti.

Remdamasis teoriniais sociologijos (psichologijos, medicinos, ekonomikos) samprotavimais, tyrėjas sudaro stebimų kintamųjų priklausomybes atspindintį modelį. Po to tikrina, ar jis tinka turimiems duomenimis. Dažniausiai tokie modeliai sudaromi bandant suprasti tiriamo reiškinio priežastis – nuo ko priklauso teigiamas rinkėjų požiūris į politinę partiją, kas lemia dėstytojo gebėjimą bendrauti su studentais, kaip pasireiškia šešėlinė ekonomika. Norime pabrėžti, kad statistinis tyrimas tik padeda atsakyti į klausimą, ar kintamieji susiję, bet ne į klausimą, ar ryšys priežastinis. Priežastinio ryšio ir priklausomybės tyrimas nėra vienas ir tas pats. Pavyzdžiui, tarp vyrų ir moterų vidutinių gyvenimo trukmių yra stipri koreliacija. Bet ar ši koreliacija rodo priežastinį ryšį? Labai abejotina. Greičiau jau gyvenimo kokybės augimas (kurį savo ruožtu lemia daugybė faktorių) sąlygoja, kad tiek vyrai, tiek moterys gyvena ilgiau.

Statistika sprendžia klausimą, ar kintamieji yra priklausomi.
Ar priklausomybė priežastinė – sprendžia pats tyrėjas.

Norėdamas pagrįsti priklausomybės priežastingumą, tyrėjas remiasi savo tiriamos srities teorija ir (ar) sveiku protu. Jei jau spėjamas kintamųjų priežastinis ryšys, tai jie turi koreliuoti ir priežastis eiti prieš pasekmę (išgertas alkoholio kiekis daro įtaką vairuotojų reakcijos greičiui, bet ne atvirkščiai). Tikimės, kad skaitytojas šių pastabų nepamirš. Toliau to fakto, kad visi priežastiniai ryšiai yra tik spėjami, nebepabrėšime.

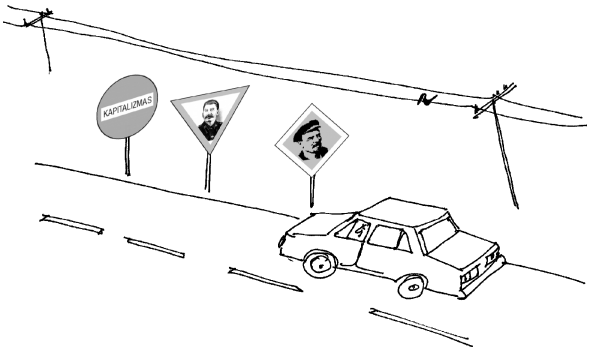
Sudarant spėjamų priežastinių ryšių modelius nepavyksta išvengti subjektyvumo. Todėl norėtume skaitytojui priminti jau ne kartą autorių išsakytą mintį. Kai yra keletas modelių, statistika padeda atsakyti į klausimą, kuris jų geriau tinka duomenims. Nereikia tikėtis, kad statistika atsakys į klausimą, kuris modelis teisingesnis.

Tinkamesnis duomenims modelis nebūtinai yra geresnis ir logiškesnis.

Modelį pasirenka ir, reikalui esant, jį nuo įkyrių oponentų apgina pats tyrėjas, kuris turėtų būti geras savo srities specialistas. Pakartosime jau ne sykį minėtą tiesą – statistika tėra tyrimo įrankis. Panašiai kaip geras teptukas tapytojui – labai svarbus, bet pats paveikslų netapo ir už tai, kad vietoj portreto atsiranda abstraktus peizažas, neatsako. Mūsų tikslas yra išmokti tuo įrankiu (statistika) naudotis – patikrinti modelio savybes ir jį priimti arba atmesti.

1. KELIŲ ANALIZĖ

Sovietiniai laikai. Naktis. Ukrainos stepe keliu važiuoja pasiklydęs automobilis. Keleiviai niūriai žvelgia pro langus. Tuščia, nyku – nei žmonių, nei nuorodų. Galų gale tolumoje pasirodo kelio ženklas. Visuotinis pagyvėjimas. Automobilis privažiuoja, o ženklas rodo, kad visi keliai veda į komunizmą.



Kelių analizė, pasiūlyta praėjusio šimtmečio trečiąjį dešimtmetį, buvo pirmasis bandymas priežastinius ryšius nagrinėti taikant regresiją. Tai toks regresijos išplėtojimas, kai dalis nepriklausomų kintamųjų (indikatorių, regresorių, aiškinamųjų kintamųjų) kartu yra ir priklausomi kintamieji. Kelių analizės tikslas – įvertinti tiriamų kintamųjų sąryšius. Šiuo metu kelių analizė tėra tik viena iš sudėtinių struktūrinių lygčių analizės dalių. Šioje knygoje ji nagrinėjama atskirai, nes paprastesniais modeliais galima lengviau iliustruoti naujas sąvokas, paaiškinti skaičiuojamų rodiklių prasmę ir statistinio modelio tyrimo etapus.

Iki šiol lietuviškos struktūrinių lygčių modelių terminijos faktiškai nebuvo. Iš pirmo žvilgsnio terminas *kelių analizė* turėtų priklausyti susisiekimo ministerijos sričiai. Tačiau jis daug trumpesnis už pagal prasmę tikslesnį terminą *spėjamų priklausomybių tarp matuojamų intervalinių kintamųjų analizė*. Šioje knygoje kelių analizei žymėti vartojame ir visuotinai priimtą anglišką trumpinį PA[1]. Manome, kad tai leis skaitytojui lengviau orientuotis skaitant papildomą literatūrą ar naudojant statistines programas.

Kituose skyreliuose aptarsime PA modelio sudarymo, modelio identifikavimo, parametrų vertinimo, tinkamumo duomenims ir interpretavimo problemas.

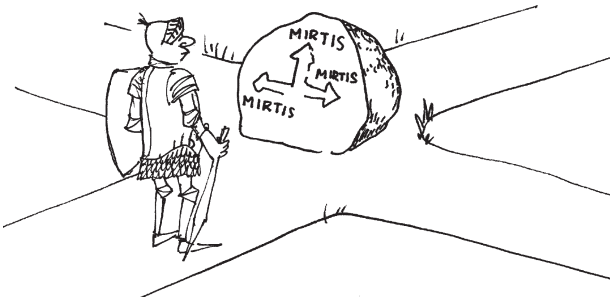
1.1. Kelių analizės modelio aprašymas

Kartais rečiau pasirenkamas kelias yra kelias, rečiau pasirenkamas dėl rimtų priežasčių.
Dž. Seinfeldas

Norint sudaryti PA modelį, reikia jį aprašyti lygtimis arba pavaizduoti diagrama. Būtent šiame etape, remiantis nagrinėjamo dalyko teorija, nusprendžiama, kokie kintamieji

[1] PA – angl. *Path Analysis.*

bus priklausomi, ir užrašomos jų spėjamos priklausomybės. Kokius ryšius pasirinkti – sprendžia pats tyrėjas, atsižvelgdamas į išankstinę informaciją.



Jau minėjome, kad kelių analizės modelis yra išplėtotas regresijos modelis. Norėdami geriau suprasti, kaip tokie modeliai sudaromi, panagrinėkime paprastą pavyzdį.

3.1.1 pavyzdys. Specialūs klausimynai padėjo įvertinti 229 atsitiktinai pasirinktų respondentų pasitenkinimą gyvenimu (kintamasis *Gyvenimas*), santykius darbovietėje (*Darbas*), santykius šeimoje (*Šeima*) ir vidutinę respondento algą (*Alga*). Visi kintamieji yra intervaliniai. Kintamųjų kovariacijų matrica pateikta 3.1.1 lentelėje.

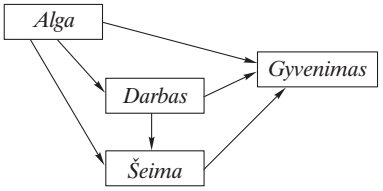
3.1.1 lentelė. Empirinė kintamųjų kovariacijų matrica

	Darbas	Gyvenimas	Šeima	Alga
Darbas	2,84			
Gyvenimas	1,10	2,72		
Šeima	2,10	1,98	5,05	
Alga	1,61	0,85	1,64	2,48

Spėjame, kad:

1) Pasitenkinimas gyvenimu priklauso nuo visų likusių kintamųjų.
2) Alga veikia santykius darbe ir šeimoje.
3) Santykiai darbe veikia santykius šeimoje.

Grafiškai spėjamas priklausomybes galime pavaizduoti tokia principine schema (žr. 3.1.1 pav.).



3.1.1 pav. Spėjamų priklausomybių schema

Užrašykime spėjamas priklausomybes regresinėmis lygtimis. Matome, kad santykiams darbe įtaką daro gaunama alga. Vis dėlto naivu tikėtis, kad santykiai darbe pri-

klauso *tik* nuo gaunamos algos. Todėl į atitinkamą lygtį įtraukiama atsitiktinė liekamoji paklaida e_1, žyminti visa tai, nuo ko dar galėtų tie santykiai priklausyti. Analogiškai kitose lygtyse įtraukiamos paklaidos e_2 ir e_3. Tradiciškai tariama, kad visos paklaidos turi normalųjį skirstinį su nuliniais vidurkiais. Gaunama tokia vadinamųjų struktūrinių lygčių sistema:

$$\begin{cases} Darbas = b_1 Alga + e_1, \\ Šeima = b_2 Alga + b_3 Darbas + e_2, \\ Gyvenimas = b_4 Alga + b_5 Darbas + b_6 Šeima + e_3. \end{cases} \tag{3.1.1}$$

Šia lygčių sistema aprašėme (specifikavome) nagrinėjamą modelį.

Pastaba. Lygtyse nėra laisvųjų narių. Kad būtų paprasčiau, kelių analizės aprašymuose tariama, kad kintamieji yra centruoti (t. y. kintamojo vidurkis atimtas iš kiekvieno stebėjimo).

Regresinėje analizėje įprasta nagrinėti nepriklausomus ir priklausomus kintamuosius. Struktūrinių lygčių analizėje (tap pat ir PA) vartojamos kitos sąvokos: *stebimieji*, *endogeniniai*, *egzogeniniai* ir *tarpiniai kintamieji*.

1 *Stebimieji kintamieji.* Tai kintamieji, kurie tiesiogiai stebimi (matuojami). Jie dar vadinami matuojamaisiais kintamaisiais, arba *indikatoriais*. Nagrinėtame 3.1.1 pavyzdyje visi kintamieji (*Gyvenimas, Darbas, Šeima, Alga*) yra tiesiogiai stebimi. Nestebimos tik liekamosios paklaidos e_1, e_2, e_3. Dažnai sakoma, kad PA modeliuose visi kintamieji yra stebimi. Taip yra todėl, kad istoriškai kelių analizėje liekamosios paklaidos nebuvo interpretuojamos kaip kintamieji.

2 *Endogeniniai kintamieji.* Tai vidiniai (kilę iš vidaus) kintamieji, t. y. kurių kintamumą lemia išoriniai ar vidiniai modelio kintamieji, kurių elgesį bandome paaiškinti modeliu. 3.1.1 pavyzdyje yra trys endogeniniai kintamieji – *Gyvenimas, Darbas, Šeima*.

3 *Egzogeniniai kintamieji.* Tai išoriniai (kilę iš išorės) kintamieji, t. y. kurių kintamumą lemia ne modelio kintamieji. Modeliu nebandome jo elgesio paaiškinti. 3.1.1 pavyzdyje yra vienintelis stebimas egzogeninis kintamasis – *Alga*. Beje, liekamosios paklaidos dažniausiai taip pat laikomos egzogeniniais kintamaisiais. Terminai *egzogeninis*, *endogeninis* yra sąlyginiai ir siejami tik su konkrečiu nagrinėjamu modeliu. Bet koks tiriamas reiškinys yra platesnio, sudėtingesnio reiškinio dalis. Kita vertus, sudaromas modelis niekada neapima visų stebimų ar nestebimų tiriamą reiškinį sąlygojamų požymių. Egzogeninis kintamasis dažniausiai taip pat priklauso nuo kitų kintamųjų, kurie tiesiog nėra įtraukti į nagrinėjamą modelį.

4 *Tarpiniai kintamieji* (mediatoriai). Tai endogeniniai kintamieji, kurie kartu yra kitų endogeninių kintamųjų paaiškinamieji kintamieji. 3.1.1 pavyzdyje tarpiniai kintamieji yra *Darbas* ir *Šeima*.

Taigi klasikiniame PA modelyje visi kintamieji yra stebimi, dalis jų yra egzogeniniai, likusieji – endogeniniai, o tarp endogeninių dar gali būti tarpinių kintamųjų.

Struktūrinių lygčių koeficientai dar vadinami *kelių koeficientais*. Jie parodo, kaip stipriai vienas kintamasis veikia kitą. 3.1.1 pavyzdyje kelių koeficientai yra b_1, b_2, b_3, b_4, b_5 ir b_6. Jeigu modelyje naudojami standartizuoti duomenys (duomenų z reikšmės), tai gauname standartizuotus kelių koeficientus.

Modelio parametrus sudaro kelių koeficientai, liekamųjų paklaidų dispersijos, egzogeninių kintamųjų dispersijos ir kintamųjų kovariacijos (koreliacijos). Šio skyrelio pavyzdyje yra šie parametrai: $b_1, b_2, \ldots, b_6$, liekamųjų paklaidų e_1, e_2, e_3 dispersijos ir kintamojo *Alga* dispersija. Sudarant modelį, dalį parametrų galima iš anksto fiksuoti. Endogeninių kintamųjų dispersijos nėra modelio parametrai. Jas galima vienareikšmiškai rasti žinant kitų parametrų įverčius. Išsamiau modelio parametrai aptariami 1.3 skyrelyje.

Dabar panagrinėsime prielaidas, kuriomis remiamasi sudarant PA modelio struktūrines lygtis:

1 Kintamųjų sąryšiai yra tik tiesiniai. Paprastai kalbant, priklausomybės yra $Y = a + bX + cZ$ pavidalo, nėra kintamųjų laipsnių (pvz., X^2) arba jų sandaugų (pvz., XZ).

2 Klasikiniame PA modelyje liekamosios paklaidos (nematuojami kintamieji) nekoreliuoja nei tarpusavyje, nei su kitais stebimais kintamaisiais. Pažymėtina, kad sudėtingesniuose struktūrinių lygčių modeliuose liekamųjų paklaidų koreliacija leidžiama.

3 Stebimi kintamieji yra intervaliniai ir matuojami be paklaidų. Ši sąlyga iš tikrųjų tėra 2 prielaidos patikslinimas. Struktūrinėse lygtyse liekamosios paklaidos nėra interpretuojamos kaip matavimo paklaidos.

4 Mažas stebimų kintamųjų multikolinearumas (priešingu atveju parametrų įverčiai turi dideles paklaidas). Faktiškai reikalaujama, kad kintamieji stipriai nekoreliuotų.

Pastaba. Iš tiesų, kelių analizės modelio prielaidos yra kiek griežtesnės už išvardytas. Pavyzdžiui, tariama, kad visi stebimi kintamieji pasiskirstę pagal normalųjį dėsnį. Sprendžiant praktinius uždavinius, ši prielaida dažnai pažeidžiama – stebimi skintamieji būna ir kategoriniai, ir net dichotominiai. Plačiau šią problemą aptarsime kitame skyriuje.

Be aukščiau išvardytų teorinių reikalavimų kintamiesiems, yra ir praktinių reikalavimų duomenims. Vienas jų – stebėjimų turi būti pakankamai daug, kad galima būtų užtikrinti modelio parametrų įverčių stabilumą. Kartais siūloma tokia praktinė taisyklė: rekomenduojama struktūrinių lygčių analizę taikyti tik tada, kai stebėjimų yra bent 10 kartų daugiau nei vertinamų parametrų. Analizė netaikoma, jeigu stebėjimų skaičius viršija parametrų skaičių ne daugiau kaip 5 kartus.

1.2. Modelio diagrama

„Akivaizdu“ yra pavojingiausias žodis matematikoje.
E. T. Belas

Viena iš PA ypatybių yra ta, kad modelio aprašymas palydimas diagrama. Ji yra struktūrinių lygčių grafinis atvaizdavimas, palengvinantis modelio suvokimą ir rezultatų interpretavimą. Kelių diagramos yra tik atskiras struktūrinių lygčių modelių diagramų atvejis, todėl čia išvardijamos bendrosios braižymo taisyklės. Diagramose naudojami elementai nurodyti 3.1.2 lentelėje.

3.1.2 lentelė. Kelių diagramos elementai

Žymėjimas	Reikšmė
	Stebimasis (matuojamasis) kintamasis
	Liekamoji paklaida (nematuojamas kintamasis)
	Vienpusis priežastinis ryšys
	Koreliacija (kovariacija)

Kartais norint pažymėti, kad egzogeninis kintamasis turi dispersiją, prie jo pripiešiama lenkta dviguba rodyklė. Kitaip tariant, lenkta dviguba rodyklė tarp dviejų kintamųjų žymi kovariaciją, o lenkta dvipusė rodyklė, prasidedanti ir pasibaigianti tame pačiame kintamajame, žymi jo dispersiją (kovariaciją su savimi). Toks žymėjimas yra visiškai logiškas, tačiau labai užgriozdina diagramą, joje sunku įžiūrėti kintamųjų priklausomybes. Todėl šioje knygoje egzogeninių kintamųjų dispersijų diagramose nežymėsime.

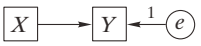
Egzogeninis kintamasis diagramoje yra tas, į kurį nėra nukreiptos jokios tiesios rodyklės (bet lenkta rodyklė gali būti). Endogeninis kintamasis diagramoje tas, į kurį nukreipta bent viena tiesi rodyklė. Be to, kiekvienas endogeninis kintamasis sujungtas su atitinkama liekamąja paklaida. Endogeninių kintamųjų dispersijos išreiškiamos per egzogeninių kintamųjų ir liekamųjų paklaidų dispersijas.

Visi egzogeniniai kintamieji turi dispersijas, kurios diagramoje nežymimos.
Visi endogeniniai kintamieji turi liekamąsias paklaidas.

Stebimieji kintamieji vaizduojami kvadratėliais, į kuriuos įrašomas kintamojo vardas, o liekamosios paklaidos (nematuojami kintamieji) vaizduojamos apskritimais su įrašytais žymėjimais. Tarkime, kad turime lygtį

$$Y = bX + e.$$

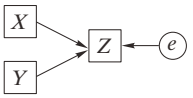
Tada kelio diagrama atrodo taip:



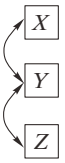
Lygtyje daugiklis prie e buvo 1, jį diagramoje ir užrašėme.

Yra nusistovėjusios tokios kelių diagramų sudarymo taisyklės:

1 Diagramos kairėje esantys kintamieji priežastingumo prasme yra ankstesni nei kintamieji, esantys diagramos dešinėje:



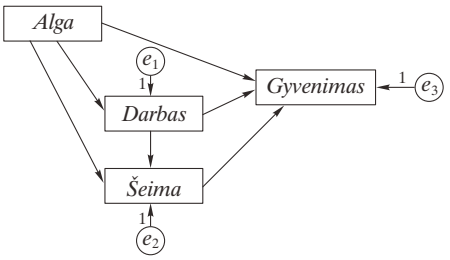
2 Kintamieji, kurie spėjamai susiję, tačiau ne priežastiniu ryšiu, vaizduojami toje pačioje horizontaliosios ašies atžvilgiu padėtyje:



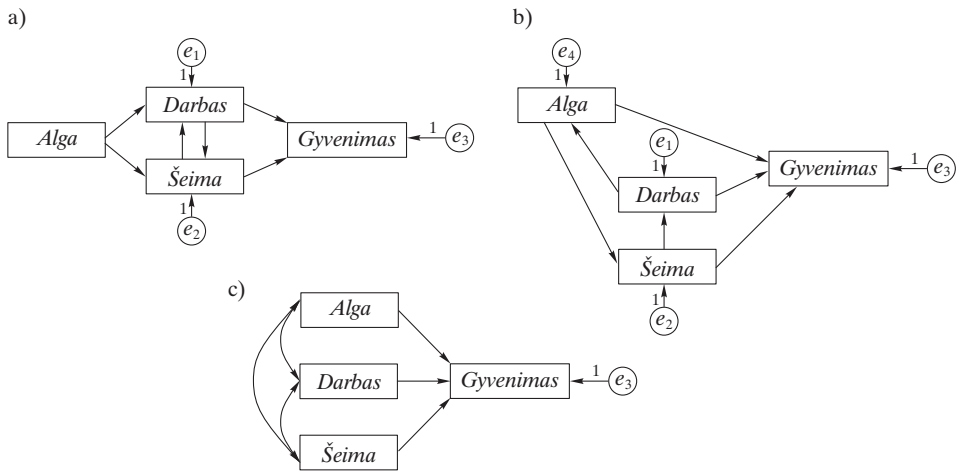
3 Spėjamas priežastinis poveikis gali būti žymimas pliuso arba minuso ženklu prie rodyklių. Tuo parodoma, koks yra vieno kintamojo pokyčio poveikis kitam kintamajam.

Žinoma, tai tik rekomendacijos. Svarbiausias reikalavimas – nubraižytos kelių diagramos turi būti suprantamos. Dabar jau galime nusibraižyti 3.1.1 pavyzdžio kelių diagramą (žr. 3.1.2 pav.). Atkreipkite dėmesį, kad diagramoje prie liekamųjų paklaidų nurodyti jau žinomi kelių koeficientai. Visi jie buvo lygūs vienetui (žr. struktūrinių lygčių (3.1.1) sistemą).

Kelių analizės modeliai skirstomi į *rekursyviuosius* ir *nerekursyviuosius*. Rekursyvaus modelio priežastiniai poveikiai yra asimetriniai (vienos krypties). Žvelgiant į tokio modelio kelių diagramą, galima pasakyti, kad kartą iškeliavęs atgal negrįši. Rekursyvaus modelio liekamosios paklaidos nekoreliuoja. Nerekursyvaus modelio bent du kintamieji veikia vienas kitą tiesiogiai ar netiesiogiai (per tarpinius kintamuosius) arba yra bent dvi koreliuojančios liekamosios paklaidos. Ar modelis rekursyvus, lengviausiai galima suprasti iš jo kelių diagramos (žr. 3.1.3 pav.). Modelis a) nerekursyvus, nes *Darbas* tiesiogiai veikia kintamąjį *Šeima* ir atvirkščiai; modelis b) nerekursyvus, nes *Šeima* tiesiogiai veikia kintamąjį *Darbas* ir yra netiesioginis atvirkštinis poveikis: *Darbas*$\longrightarrow$*Alga*$\longrightarrow$*Šeima*; modelis c) rekursyvus, nes koreliuoja tik egzogeniniai kintamieji. Beje, visų modelių



3.1.2 pav. 3.1.1 pavyzdžio kelių diagrama



3.1.3 pav. Modelių pavyzdžiai: a) nerekursyvusis, b) nerekursyvusis, c) rekursyvusis

interpretacija yra pakankamai prasminga. Pavyzdžiui, modelis a) teigia, kad santykiai šeimoje gali stipriau veikti santykius darbe nei darbo santykiai veikia santykius šeimoje.

Modelio rekursyvumas svarbus tik parametrų vertinimui (modelio tinkamumui duomenims). Vertinant parametrus, reikia išspręsti kelių lygčių sistemą. Turint nerekursyvus modelius, tai padaryti sunkiau, o kartais net neįmanoma. Klasikinėje kelių analizėje nagrinėjami tik rekursyvieji modeliai. Beje, modelis yra tikrai rekursyvus, jeigu kelių diagramoje:

1 Joks kelias neina per tą patį kintamąjį daugiau kaip vieną kartą.

2 Tarp dviejų kintamųjų nėra skirtingų krypčių kelių.

3 Per lenktą dvipusę rodyklę, vaizduojančią spėjamą koreliaciją tarp egzogeninių kintamųjų, joks kelias neina daugiau kaip kartą.

Žinoma, šie reikalavimai tėra rekomendaciniai. Sudarant modelį, tyrėjui mažiausiai reikėtų sukti galvą, ar modelis rekursyvus, ar ne. Kita vertus, nereikėtų nustebti, paaiškėjus, kad modelis duomenims iš viso netinka.

1.3. Kelių analizės modelio identifikavimas

Ką vienas matematikos vadovėlis „sako“ kitam? „Neįkyrėk, aš ir pats turiu krūvą problemų!“

Sudarius modelį, reikia jį identifikuoti – įvertinti, ar duomenų aibė teoriškai pakankama modelio parametrams įvertinti. Galima prikurti tokių modelių, kad net teoriškai to padaryti nepavyks. Toliau šiame skyrelyje nagrinėsime, kaip identifikuojami centruoti modeliai (modeliai be laisvųjų narių regresinėse lygtyse, žr., pvz., (3.1.1) sistemą).

Modelio identifikuojamumas – teorinė *modelio* savybė. Norint modelį identifikuoti, reikia patikrinti, ar yra pakankamai informacijos parametrams įvertinti. Tai primena lygčių sistemos sprendimą su daugeliu nežinomųjų. Jei nežinomųjų bus daugiau nei lygčių, tai sprendinio nerasime. Primename, kad centruoto modelio vertinamus parametrus sudaro kelių koeficientai, liekamųjų paklaidų dispersijos, egzogeninių kintamųjų dispersijos ir kintamųjų kovariacijos (koreliacijos). Naudojantis šiais parametrais, galima išreikšti endogeninių kintamųjų dispersijas. Todėl pastarosios vertinamais modelio parametrais nėra laikomos.

Vertinant parametrus, atsižvelgiama į jų kovariacijas. Kodėl ne į koreliacijas? Ir kovariacija, ir koreliacija matuoja kintamųjų tiesinės priklausomybės stiprumą.[2] Koreliacija yra standartizuota kovariacija (kovariacija tarp *standartizuotų* kintamųjų, turinčių nulinius vidurkius ir vienetines dispersijas). Ji geriau tinka, kai reikia nustatyti, kuri priklausomybė stipresnė. Kovariacijoje geriau atsispindi informacija apie kintamųjų vidurkius ir dispersijas. Beje, kintamojo kovariacija su savimi lygi jo dispersijai.

Tarkime, kad modelyje yra k stebimų kintamųjų. Tada galima rasti $k(k+1)/2$ skirtingų kovariacijų (literatūroje jos dar vadinamos duomenų taškais). Pagal vertinamų parametrų skaičių modeliai skirstomi į nepakankamai identifikuojamus, tiksliai identifikuojamus ir perteklinės identifikacijos.

1 *Nepakankamai identifikuojamas* (*neidentifikuojamasis*) *modelis*. Vertinamų parametrų skaičius didesnis už $k(k+1)/2$. Neįvedus papildomų apribojimų, tokio modelio parametrų vienareikšmiškai įvertinti negalima.

2 *Tiksliai identifikuojamas* (*pilnasis*) *modelis*. Vertinamų parametrų skaičius lygus $k(k+1)/2$. Būtent tokie yra klasikinės kelių analizės modeliai. Šiuo atveju teoriškai informacijos pakanka parametrų įverčiams apskaičiuoti, tačiau nėra prasmės tirti modelio tinkamumo duomenims laipsnio, nes įvertintoji modelio kovariacijų matrica ir empirinė kovariacijų matrica sutampa. Plačiau apie modelio tinkamumą duomenims nagrinėsime 2.6 ir 3.6 skyreliuose.

3 *Perteklinės identifikacijos modelis*. Vertinamų parametrų yra mažiau nei $k(k+1)/2$. Šiuo atveju galima įvertinti ne tik nežinomus parametrus, bet ir modelio kokybę, palyginti modelio kovariacijos matricos įvertį su empirine kovariacijų matrica. Dažniausiai perteklinės identifikacijos modeliai yra sudėtingesni struktūrinių lygčių modeliai.

Pakankamos ir perteklinės identifikacijos modeliai yra identifikuojami, todėl galima užrašyti tokią taisyklę:

Centruotas PA modelis su k kintamųjų identifikuojamas, jei vertinamų parametrų skaičius yra ne didesnis kaip $k(k+1)/2$.

Skirtumas tarp $k(k+1)/2$ ir vertinamų parametrų skaičiaus vadinamas modelio laisvės laipsnių skaičiumi ir žymimas df. Taigi modelis identifikuojamas, jeigu jo laisvės laipsnių skaičius yra neneigiamas.

[2] Plačiau žr. *Statistika ir jos taikymai*. I, p. 97–98.

3.1.3 lentelė. Modelių parametrai

Modelis	Kelio koeficientai	Kovariacijos	Dispersijos	Iš viso
3.1.3 (a)	$A \to D$, $A \to Š$, $D \to Š$, $D \to G$, $Š \to D$, $Š \to G$		A, e_1, e_2, e_3	10
3.1.3 (b)	$A \to G$, $A \to Š$, $D \to A$, $D \to G$, $Š \to D$, $Š \to G$		e_1, e_2, e_3, e_4	10
3.1.3 (c)	$A \to G$, $D \to G$, $Š \to G$	$A \leftrightarrow D$, $A \leftrightarrow Š$, $D \leftrightarrow Š$	A, D, $Š$, e_1	10

Pavaizduotų 3.1.3 paveiksle modelių parametrai pateikiami 3.1.3 lentelėje. Kad būtų trumpiau, kintamieji *Šeima*, *Darbas*, *Gyvenimas* ir *Alga* lentelėje žymimi didžiosiomis raidėmis $Š$, D, G ir A. Yra keturi kintamieji ($k = 4$, $k(k+1)/2 = 10$), todėl matome, kad visi modeliai pakankamai identifikuojami. Norint gauti perteklinės identifikacijos modelį, užtektų modelio diagramoje pašalinti vieną kelią arba koreliaciją.

Toliau šiame skyriuje nagrinėjami tik tiksliai identifikuojami modeliai.

Identifikuojamumas tėra teorinė modelio savybė, todėl nėra jokių garantijų, kad tiriant realius duomenis pavyks įvertinti identifikuojamo modelio parametrus. Pavyzdžiui, du kintamieji gali labai stipriai koreliuoti, o tada skirtingų kovariacijų jau bus mažiau nei $k(k+1)/2$.

1.4. Modelio parametrų įvertinimas

Investicinei bendrovei reikia matematiko. Po pradinės atrankos lieka du kandidatai: finansų matematikas ir ekonomistas. Į klausimą, kokio atlyginimo norėtų, matematikas atsako, kad 4000 Lt per mėnesį. „Trisdešimt tūkstančių“, – pareiškia ekonomistas. „Kodėl tiek daug?“ – stebisi personalo direktorius. „Trylika tūkstančių jums, trylika – man, o už likusius pasamdysime matematiką!“

Primename, kad nagrinėjant klasikinį (tiksliai identifikuojamą, rekursyvų, centruotų kintamųjų) PA modelį, reikia įvertinti šiuos parametrus:

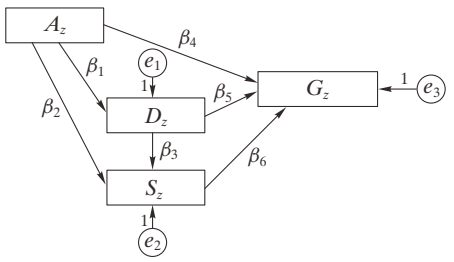
1) kelių koeficientus,
2) egzogeninių kintamųjų koreliacijas,
3) egzogeninių kintamųjų dispersijas,
4) endogeninių kintamųjų liekamųjų paklaidų dispersijas.

Jeigu kintamieji būtų necentruoti, t. y. jų vidurkiai nebūtų lygūs nuliui, dar reikėtų įvertinti kintamųjų vidurkius.

Norėdami geriau suprasti, kaip randami parametrų įverčiai, išnagrinėsime 3.1.1 pavyzdžio modelio (žr. (3.1.1) sistemą ir 3.1.2 pav.) analogą, sudarytą standartizuotiems kintamiesiems, kurių vidurkiai lygūs nuliui, o dispersijos – vienetui. Vietoje kintamųjų *Darbas*, *Alga*, *Šeima*, *Gyvenimas* imame jų z reikšmių analogus – D_z, A_z, S_z ir G_z. Šiuo atveju parametrų įvertinimui užteks kintamųjų koreliacijų matricos.

Tiriame tokią struktūrinių lygčių sistemą:

$$\begin{cases} D_z = \beta_1 A_z + e_1, \\ S_z = \beta_2 A_z + \beta_3 D_z + e_2, \\ G_z = \beta_4 A_z + \beta_5 D_z + \beta_6 S_z + e_3; \end{cases} \tag{3.1.2}$$



3.1.4 pav. Standartizuotų kintamųjų kelio diagrama

čia $e_j \sim \mathrm{N}(0, \sigma_j^2)$, $j = 1, 2, 3$. Kelių diagrama su nurodytais koeficientais pavaizduota 3.1.4 paveiksle. Reikia įvertinti šešis nežinomus kelio parametrus β_j ir tris liekamųjų paklaidų dispersijas σ_j^2, $j = 1, 2, 3$. Tarkime, kad koreliacijos tarp kintamųjų yra $\varrho(A, G) = \varrho_{AG}$, $\varrho(A, D) = \varrho_{AD}$ ir t. t. Išreikšime šias koreliacijas per nežinomus parametrus. Kintamieji turi nulinius vidurkius ir vienetines dispersijas, todėl jų koreliacijos sutampa su kintamųjų sandaugos vidurkiais. Klasikiniuose PA modeliuose reikalaujama, kad kintamiųjų liekamosios paklaidos nekoreliuotų nei tarpusavyje, nei su kitais egzogeniniais kintamaisiais. Todėl galime užrašyti:

$$\begin{aligned}\varrho_{AG} &= \mathbf{E}A_zG_z = \mathbf{E}A_z(\beta_4A_z + \beta_5D_z + \beta_6S_z + e_3)\\ &= \beta_4\mathbf{E}A_z^2 + \beta_5\mathbf{E}A_zD_z + \beta_6\mathbf{E}A_zS_z + \mathbf{E}A_z\mathbf{E}e_3\\ &= \beta_4\mathbf{D}A_z + \beta_5\varrho_{AD} + \beta_6\varrho_{AS} = \beta_4 + \beta_5\varrho_{AD} + \beta_6\varrho_{AS}.\end{aligned}$$

Analogiškai užrašę visas likusias koreliacijas, gauname šešių lygčių sistemą. Be to, žinome, kad visos dispersijos lygios vienetui. Todėl

$$1 = \mathbf{D}D_z = \mathbf{D}\beta_1A_z + \mathbf{D}e_1 = \beta_1^2\mathbf{D}A_z + \sigma_1^2 = \beta_1^2 + \sigma_1^2.$$

Analogiškai užrašome $\mathbf{D}S_z$ ir $\mathbf{D}G_z$. Iš visų lygčių, gauname tokią devynių lygčių su devyniais nežinomaisiais sistemą:

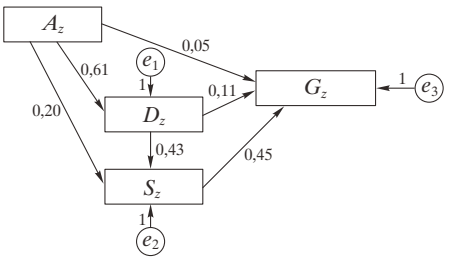
$$\begin{cases}\beta_1 = \varrho_{AD},\\ \beta_2 + \beta_3\varrho_{AD} = \varrho_{AS},\\ \beta_4 + \beta_5\varrho_{AD} + \beta_6\varrho_{AS} = \varrho_{AG},\\ \beta_1\beta_2 + \beta_1\beta_3\varrho_{AD} = \varrho_{DS},\\ \beta_1\beta_4 + \beta_1\beta_5\varrho_{AD} + \beta_1\beta_6\varrho_{AS} = \varrho_{DG},\\ (\beta_2\beta_4 + \beta_3\beta_5) + (\beta_2\beta_5 + \beta_3\beta_4)\varrho_{AD} + \beta_2\beta_6\varrho_{AS} + \beta_3\beta_6\varrho_{DS} = \varrho_{GS},\\ \beta_1^2 + \sigma_1^2 = 1,\\ \beta_2^2 + \beta_3^2 + \sigma_2^2 = 1,\\ \beta_4^2 + \beta_5^2 + \beta_6^2 + \sigma_3^2 = 1.\end{cases} \tag{3.1.3}$$

Norėdami gauti nežinomų parametrų įverčius, į (3.1.3) lygčių sistemą įstatome empirines koreliacijas (pažymėkime jas r_{AD} ir t. t.). Liks tik išspręsti gautąją sistemą.

3.1.2 pavyzdys. Pasinaudoję 3.1.1 lentele, gauname tokią empirinių koreliacijų matricą (matrica simetrinė, todėl pateikiame tik jos reikšmes po pagrindine įstrižaine):

$$\Sigma = \begin{pmatrix} r_{DD} & & & \\ r_{GD} & r_{GG} & & \\ r_{SD} & r_{SG} & r_{SS} & \\ r_{AD} & r_{AG} & r_{AS} & r_{AA} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 0{,}396 & 1 & & \\ 0{,}555 & 0{,}534 & 1 & \\ 0{,}607 & 0{,}327 & 0{,}463 & 1 \end{pmatrix}. \tag{3.1.4}$$

Įstatę šias empirines koreliacijas į (3.1.3) lygčių sistemą ir ją išsprendę, gauname tokius parametrų įverčius: $\hat{\beta}_1 = 0{,}61$, $\hat{\beta}_2 = 0{,}20$, $\hat{\beta}_3 = 0{,}43$, $\hat{\beta}_4 = 0{,}05$, $\hat{\beta}_5 = 0{,}11$, $\hat{\beta}_6 = 0{,}45$, $\hat{\sigma}_1^2 = 0{,}63$, $\hat{\sigma}_2^2 = 0{,}66$, $\hat{\sigma}_3^2 = 0{,}70$. Gautieji kelių koeficientų įverčiai pavaizduoti 3.1.5 paveiksle.



3.1.5 pav. Standartizuotų kintamųjų kelio diagrama su nurodytais kelių koeficientais

Pastaba. Kartais nagrinėjami modeliai, kuriuose visos liekamosios paklaidos turi vienetines dispersijas, t. y. $e_j \sim \mathrm{N}(0,1)$. Tada įvedami papildomi kelio koeficientai ir truputį keičiasi modelis. Pareikalavę, kad liekamosios paklaidos turėtų vienetines dispersijas, modelio parametrų priklausomybes aprašytume taip:

$$\begin{cases} D_z = \beta_1 A_z + \omega_1 e_1, \\ S_z = \beta_2 A_z + \beta_3 D_z + \omega_2 e_2, \\ G_z = \beta_4 A_z + \beta_5 D_z + \beta_6 S_z + \omega_3 e_3. \end{cases} \tag{3.1.5}$$

Vertinamų parametrų skaičius (devyni) išlieka tas pats, o ir (3.1.3) lygčių sistema nedaug pasikeičia. Ši liekamųjų paklaidų taisyklė taikoma visuose struktūrinių lygčių modeliuose:

Modeliuose fiksuojame:
arba liekamosios paklaidos kelio koeficientą,
arba liekamosios paklaidos dispersiją.

Visiškai analogiškai skaičiuojami parametrų įverčiai ir tada, kai nagrinėjami nestandartizuoti kintamieji. Tiesiog vietoje koreliacijų imamos kovariacijos. Naudojantis 3.1.1 pavyzdžio duomenimis gauti kelių koeficientų įverčiai pavaizduoti 3.1.6 paveiksle. Lygiai tokius pat kelio koeficientų įverčius gautume, jeigu taikytume mažiausiųjų kvadratų metodą kiekvienai (3.1.1) sistemos struktūrinei lygčiai.[3]

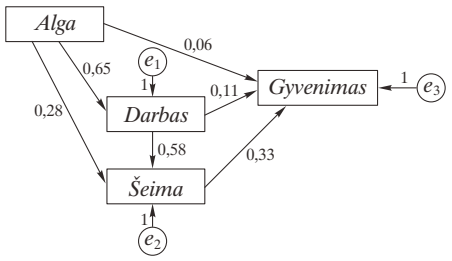
[3] Tai galima gauti naudojantis bet kokia statistine programa ir taikant teisinę regresiją kiekvienai lygčiai.

Analizuojant tiksliai identifikuojamus kelių analizės modelius, susiduriama su problemomis. Pavyzdžiui, neaišku, kuris modelis geresnis – pavaizduotas 3.1.2 paveiksle ar 3.1.3 (c) paveiksle, nes nėra kriterijų įvertinti, ar tokie modeliai tinka duomenims. Tačiau galima įvertinti kelių koeficientų statistinį reikšmingumą.

Jeigu be parametrų įverčių žinomos ir jų standartinės paklaidos (SE[4]), tai galima patikrinti hipotezę, kad parametrai nelygūs nuliui. Tiesa, imtis turi būti pakankamai didelė. Praktikoje taikomas toks, nelabai tikslus metodas. Pirmiausia parametro įvertis padalijamas iš to įverčio standartinės paklaidos. Po to tikrinama, ar gautas skaičius absoliučiuoju didumu viršija 1,96. Jeigu viršija – darome išvadą, kad parametras statistiškai reikšmingai skiriasi nuo 0. Nors iš pirmo žvilgsnio tokia procedūra atrodo itin nemokslišką, iš tikro – tai tiesiog normaliosios aproksimacijos taikymas. Parametro įvertinys, padalytas iš jo standartinės paklaidos, turi apytiksliai normalųjį skirstinį. Skaičius 1,96 tėra standartinio normaliojo skirstinio 0,975 lygmens kvantilis (0,025 lygmens kritinė reikšmė).

Tegul reikšmingumo lygmuo $\alpha = 0{,}05$.
Parametras θ statistiškai reikšmingai skiriasi nuo nulio, jei $|\hat{\theta}|/SE(\hat{\theta}) > 1{,}96$.

3.1.3 pavyzdys. Didžiausio tikėtinumo metodu įvertinti 3.1.1 pavyzdžio (žr. 3.1.6 pav.) parametrai ir jų standartinės paklaidos pateikiamos 3.1.4 lentelėje.



3.1.6 pav. 3.1.1 pavyzdžio diagrama su nurodytais kelių koeficientais

3.1.4 lentelė. Parametrų įverčiai

Parametras	Įvertis	SE	Įvertis/SE
σ_1^2	1,89	0,18	10,5
σ_2^2	3,35	0,31	10,8
σ_3^2	1,78	0,17	10,5
D*Alga*	2,46	0,23	10,7
Alga → *Šeima*	0,28	0,09	3,1
Alga → *Darbas*	0,65	0,06	10,8
Darbas → *Šeima*	0,58	0,09	6,4
Alga → *Gyvenimas*	0,06	0,08	0,8
Šeima → *Gyvenimas*	0,33	0,05	6,6
Darbas → *Gyvenimas*	0,11	0,07	1,6

[4] SE – angl. *Standard Error.*

Beje, nagrinėjant tiksliai identifikuojamus kelių analizės modelius, parametrų įverčiai, gauti didžiausiojo tikėtinumo metodu, sutampa su mažiausiųjų kvadratų metodu gautaisiais įverčiais. Nesunku įsitikinti, kad statistiškai reikšmingai nuo nulio nesiskiria tik kelio koeficientai $Alga \to Gyvenimas$ ir $Darbas \to Gyvenimas$. Tai dar kartą patvirtina išvadą, kad modelis yra tobulintinas.

1.5. Kelių koeficientų išskaidymas

Begalinė matematikų eilė stovi prie baro. Pirmasis užsisako bokalą alaus, antrasis – pusę bokalo, trečiasis – ketvirtį... „Supratau“, – sako barmenas ir pripila du bokalus.

Prisiminkime, kad PA tikslas – įvertinti spėjamus kintamųjų sąryšius. Šiame skyrelyje aprašysime, kaip naudojantis parametrų įverčiais galima apskaičiuoti vienų kintamųjų poveikį kitiems kintamiesiems.

Visą vieno kintamojo poveikį kitam kintamajam sudaro suma dviejų dydžių:

1) ***tiesioginio poveikio***, kurio didumas lygus kelio tarp tų kintamųjų koeficientui;
2) ***netiesioginio poveikio***, kurio didumas lygus kelių, einančių per vieną ar keletą tarpinių kintamųjų, sekos koeficientų sandaugai.

Visas vieno kintamojo priežastinio poveikio kitam kintamajam didumas lygus tiesioginio ir netiesioginio poveikių didumų sumai.
Tiesioginis poveikis gali būti teigiamas, o netiesioginis – neigiamas, ir atvirkščiai.
Netiesioginis poveikis gali būti didesnis už tiesioginį.

Grafiškai tiesioginio ir netiesioginio poveikio sąryšių idėja pateikta 3.1.7 paveiksle.

Panagrinėsime 3.1.1 pavyzdžio kintamojo *Alga* poveikį kintamajam *Gyvenimas* (žr. 3.1.8 pav.).

1 Tiesioginis poveikis yra $Alga \to Gyvenimas$. Jo didumas lygus 0,06.

2 Netiesioginis poveikis susideda iš kelių: $Alga \to Darbas \to Gyvenimas$, $Alga \to Šeima \to Gyvenimas$ ir $Alga \to Darbas \to Šeima \to Gyvenimas$. Netiesioginio poveikio didumas lygus $0{,}65 \cdot 0{,}11 + 0{,}28 \cdot 0{,}33 + 0{,}65 \cdot 0{,}58 \cdot 0{,}33 = 0{,}29$.

3 Visas kintamojo *Alga* poveikis kintamajam *Gyvenimas* lygus $0{,}06 + 0{,}29 = 0{,}35$.

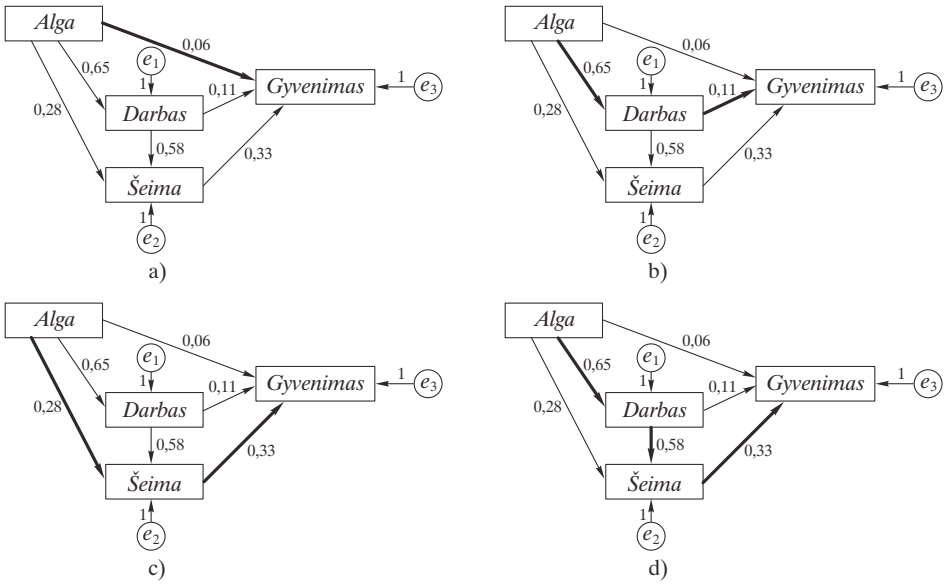
$$\boxed{X} \xrightarrow{a} \boxed{Y}$$

Tiesioginis X poveikio Y didumas $= a$

$$\boxed{X} \xrightarrow{b} \boxed{Z} \xrightarrow{c} \boxed{Y}$$

Netiesioginis X poveikio Y didumas $= bc$

3.1.7 pav. Kintamųjų priežastinio ryšio komponentai



3.1.8 pav. Kintamojo *Alga* poveikio kintamajam *Gyvenimas* išskaidymas: a) tiesioginis poveikis, b)–d) netiesioginis poveikis

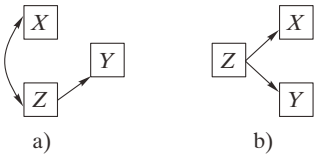
Žinoma, atitinkamas kintamųjų poveikio dedamąsias galima skaičiuoti ir turint standartizuotus kintamuosius. Pavyzdžiui, netiesioginio D_z poveikio kintamajam G_z didumas lygus $0{,}43 \cdot 0{,}45 = 0{,}19$ (žr. 3.1.5 pav.).

Kintamieji gali būti susieti ir nepriežastiniais ryšiais, kurie atsiranda dėl:

1 Kintamųjų koreliacijos. Tuo atveju sakoma, kad yra *neanalizuojamas kintamojo poveikis*.

2 Trečiojo kintamojo poveikio. Tuo atveju sakoma, kad yra *iškreiptas kintamųjų poveikis*.

Nepriežastinių ryšių poveikį iliustruoja 3.1.9 paveikslas. Pažvelgę į kelių diagramas, matome, kad neanalizuojamo poveikio dedamoji būtinai „susijusi“ su dvipuse rodykle, o iškreipto poveikio kelias pažymėtas skirtingų krypčių rodyklėmis.



3.1.9 pav. Nepriežastiniai kintamųjų ryšiai:
a) neanalizuojamas poveikis,
b) iškreiptas poveikis

Atkreipsime dėmesį į kelis kelių analizės momentus, palengvinančius modelio interpretaciją. Visų pirma reikia patikrinti, ar kelių koeficientai yra „teisingų“ ženklų.

Reikia patikrinti kelių koeficientų ženklus ir juos palyginti su kintamųjų koreliacijomis.

Teigiamas kelio tarp kintamųjų *Šeima* ir *Gyvenimas* koeficientas 0,33 (žr. 3.1.6 pav.) rodo, kad kuo geresni santykiai šeimoje, tuo didesnis ir pasitenkinimas gyvenimu. Tai visiškai natūralu. Vis dėlto, kaip ir tiesinės regresijos atveju, kartais didelė kintamųjų koreliacija iškreipia PA modelio priklausomybes (multikolinearumo problema). Todėl, gavus modelio kintamųjų priklausomybę, nesuderintą su tiesiogine kintamųjų koreliacija, vertėtų paieškoti kito modelio.

1.6. Modelio tobulinimas

Studentams liepta įsiminti telefonų knygą.
Matematikai: „O kam?“
Sociologai: „Per kiek laiko?“
Teisininkai: „O mes jau anksčiau įsiminėm.“

Jeigu modelis dėl kokių nors priežasčių tyrėjo netenkina, paprastai sudaromas naujas modelis (modelis modifikuojamas). Klasikinėje kelių analizėje apsiribojama tik tiksliai identifikuojamais rekursiniais modeliais, todėl nėra didelio modelių pasirinkimo. Modelio keitimui didesnį dėmesį skirsime kituose skyreliuose. Norime tik atkreipti dėmesį į kelis gajus mitus, susijusius su kelių analize.

Kelių diagramos su gražiomis rodyklėlėmis užburia vaizduotę. Nemažai žmonių manė (o gal ir dabar mano), kad kelių analizę galima taikyti priežastiniams ryšiams įrodyti. Kelių analizė net buvo vadinama „priežastingumo modeliavimu“. Atrodytų, jei kintamasis X sujungtas rodykle su kintamuoju Y, o Y su Z, tai, akivaizdu, kad X daro įtaką Y, o Y daro įtaką Z. Deja, taip yra toli gražu ne visada. Yra daug modelių, kuriuose pakeitus kelių rodyklių kryptis, gaunami *tiksliai* tie patys rezultatai. Priežastį ir poveikį galima nustatyti tik kruopščiai paruoštu eksperimentu, joks statistikas, kad ir kaip jis išradingai manipuliuotų duomenimis, nepavers koreliacijų priežastingumo išvadomis.

Kitas mitas yra apie tai, kad kelių analizę galima taikyti teoriniams modeliams sudaryti. PA yra modelių testavimo, o ne sudarymo (plėtojimo) procedūra. Modeliai visada yra pagrįsti išankstiniu žinojimu, patirtimi, teorija.

Kelių analizei skirtame skyrelyje yra daugybė įspėjimų ir apribojimų – tik rekursyvūs modeliai, tik matuojami kintamieji, tik tiksliai identifikuojami. Ar galima tokių apribojimų išvengti? Be abejo, taip! Tolesniuose skyreliuose tai ir panagrinėsime.

endogeninis kintamasis	liekamoji paklaida	rekursyvusis modelis
egzogeninis kintamasis	modelio identifikavimas	standartizuotasis modelis
kelių analizė	modelio laisvės laipsniai	tiesioginis poveikis
kelio koeficientas	netiesioginis poveikis	

UŽDAVINIAI

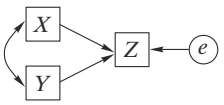
1. Žinoma trijų kintamųjų sąryšių struktūra: $X_1 \to Y$, $X_2 \to Y$, $X_1 \longleftrightarrow X_2$. Naudodamiesi kintamųjų kovariacijomis, užrašykite struktūrines lygtis ir koeficientų išraiškas.

2. Turime lygčių sistemą

$$\begin{cases} X_2 = \alpha_1 X_1 + \alpha_2 X_4 + e_1, \\ X_3 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + e_2. \end{cases}$$

   Nubraižykite kelių diagramą. Kurie kintamieji yra endogeniniai, o kurie – egzogeniniai?

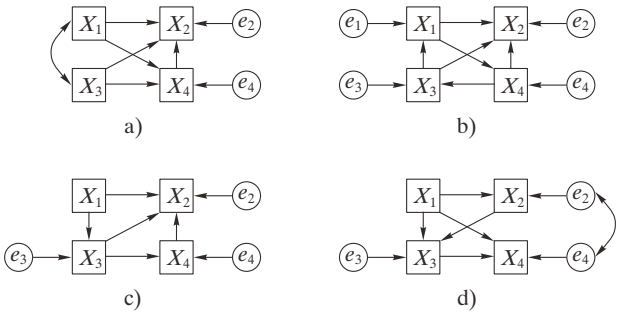
3. Tarkime, kad turime tokią kelių diagramą:



   Apskaičiuokite kelių koeficientų įverčius, kai kintamieji X, Y, Z yra standartizuoti ir jų koreliacijų matrica

$$S = \begin{pmatrix} 1 & 0,4 & 0,6 \\ 0,4 & 1 & 0,2 \\ 0,6 & 0,2 & 1 \end{pmatrix}.$$

4. Kurie iš 3.1.10 paveiksle pavaizduotų modelių yra rekursyvieji?



3.1.10 pav. Kelių analizės modeliai

5. Užrašykite modelio, pavaizduoto 3.1.10 paveiksle, c), kintamųjų X_1, X_3 ir X_4 tiesioginius ir netiesioginius poveikius kintamajam X_2.

2. PATVIRTINANČIOJI FAKTORINĖ ANALIZĖ

Kai keli pensininkai sueina į būrelį, jie dažnai sprendžia amžiną klausimą – „kas kaltas“. Iškeliamos pačios įvairiausios hipotezės: komunistai, masonai, kitataučiai, valdininkai ar šėtonas mumyse. Be to, detaliai aptariama, kaip įvairios blogybės pasireiškia. Statistiniu požiūriu pensininkai sudarinėja faktorinius modelius – įvardija spėjamą latentinį faktorių ir aptaria jo apraiškas. Tik savo modeliams patvirtinti jie netaiko statistinių metodų.

Ankstesniame skyriuje visi kintamieji buvo matuojami. Šiame skyriuje aptarsime modelius, kuriuose yra ir tiesiogiai nematuojamų kintamųjų (latentinių faktorių). Kaip matysime vėliau, čia aprašomi faktorinės analizės modeliai yra svarbi sudėtingesnių struktūrinių modelių dalis.

2.1. Dviejų rūšių faktorinės analizės

Žemės trauka yra faktorius, prisidedantis prie 73 procentų visų nelaimingų atsitikimų, susijusių su krentančiais objektais.

D. Bary

Yra dalykų, kurių tiesiogiai negalima išmatuoti – tolerancija, depresijos stiprumas, gyvenimo kokybė, šešėlinės ekonomikos apimtis, verslumas, ligos stadija. Tokie kintamieji vadinami latentiniais. Dažniausiai latentiniai kintamieji matuojami netiesiogiai, pavyzdžiui, stebint tiriamojo elgesį įvairiose situacijose, užduodant papildomus klausimus ir kt. Faktorinė analizė yra viena iš statistikos dalių, kurioje tiriamos latentinių ir stebimųjų kintamųjų tarpusavio priklausomybės. Yra dvi faktorinės analizės rūšys: tiriančioji faktorinė analizė (EFA[5]) ir patvirtinančioji faktorinė analizė (CFA[6]). Išsamiai tiriančiąją faktorinę analizę nagrinėjome [3] knygoje, todėl čia tik trumpai paaiškinsime, kuo viena faktorinė analizė skiriasi nuo kitos.

Tipinė tiriančioji faktorinė analizė atrodytų taip: pagal įvairias metodikas (tikrines reikšmes, faktoriuose išsaugotą bendrosios kintamųjų dispersijos procentą, savo slaptą troškimą) pasirenkame faktorius, kurie susiję su *visais* stebimaisiais kintamaisiais (indikatoriais). Po to ištiriame, su kuriais kintamaisiais faktoriai yra susiję stipriai. Dažniausiai tai būna vizualus vertinimas („iš akies“) – žiūrima, kurie koeficientai prie indikatorių dideli. Likusias priklausomybes ignoruojame. Tada prasideda paskutinis EFA etapas –

[5] EFA – angl. *Exploratory Factor Analysis.*

[6] CFA – angl. *Confirmatory Factor Analysis.*

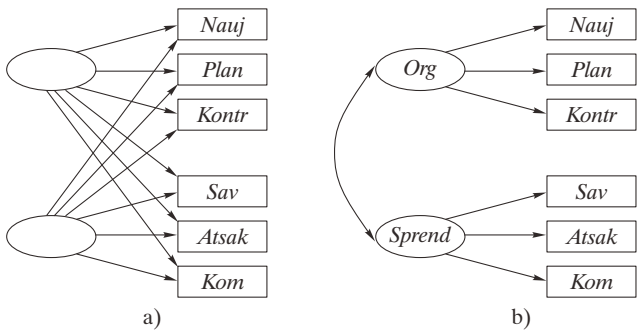
ieškome, kaip faktorius apibūdinti (kuriame jiems pavadinimus). Tuo tiriančioji faktorinė analizė ir pasibaigia. Taigi faktorių pavadinimai sugalvojami (faktoriai identifikuojami), *jau atlikus* faktorinę analizę.

3.2.1 pavyzdys. Tarkime, kad specialiais testais buvo įvertintas tam tikros pramonės šakos vadybininkų komunikabilumas (kintamasis *Kom*) bendraujant su kitais vadybininkais, sprendimų savarankiškumas (*Sav*), atsakomybės už priimtus sprendimus jutimas (*Atsak*), naujų technologijų diegimas (*Nauj*), darbo proceso planavimas (*Plan*) ir kontroliavimas (*Kontr*). Visi kintamieji intervaliniai. Jų kovariacijų matrica pateikta 3.2.1 lentelėje.

3.2.1 lentelė. Kintamųjų imties kovariacijų matrica

	Atsak	*Kom*	*Kontr*	*Nauj*	*Plan*	*Sav*
Atsak	1,85					
Kom	0,07	4,90				
Kontr	0,74	0,80	4,87			
Nauj	0,19	0,25	1,19	1,20		
Plan	0,73	0,43	3,10	1,49	5,10	
Sav	0,29	0,44	1,04	0,50	1,27	1,14

Įtariame, kad visiems išvardytiems kintamiesiems įtaką daro tik du latentiniai faktoriai. Užrašome tokią kintamųjų priklausomybių schemą, kaip pavaizduota 3.2.1 pav., a). Įvertinę faktorių įtaką tiriamiems kintamiesiems, mąstytume apie faktorių vardus.



3.2.1 pav. Vadybininko veiklos EFA (a) ir CFA (b) modelių principinės schemos

Patvirtinančioji faktorinė analizė pradedama nuo spėjamų priklausomybių. Latentiniai faktoriai „žinomi“ (jų pavadinimus sugalvojame iškart). Nurodome, koks indikatorius nuo kokio faktoriaus priklauso. Galima tirti ir faktorių koreliaciją.

3.2.2 pavyzdys. Tarkime, nusprendėme, kad vadybininką apibūdina du koreliuojantys latentiniai faktoriai: 1) organizaciniai gabumai (kintamasis *Org*), kurie atsiskleidžia planuojant darbą (*Plan*), kontroliuojant (*Kontr*) ir diegiant naujoves (*Nauj*); 2) gebėjimas priimti sprendimus (*Sprend*), kuris susijęs su vadybininko savarankiškumu (*Sav*), atsakomybės jausmu (*Atsak*) ir gebėjimu bendrauti su kitais vadybininkais (*Kom*). Spėjama CFA modelio schema pateikta 3.2.1 paveiksle, b).

Sudarę CFA modelį, tikriname, ar duomenys neprieštarauja mūsų spėjamoms priklausomybėms. Tikrinant taikomi statistiniai kriterijai ir atsižvelgiama į įvairius aprašomuosius rodiklius.

Dažniausiai tiriančiosios faktorinės analizės metodai skiriasi nuo patvirtinančiosios faktorinės analizės metodų. To pasekmė – galima situacija, kai CFA nepatvirtina, kad EFA metodu sudarytas modelis tinka duomenims.

EFA ir CFA tyrimų rezultatai esant tiems patiems duomenims gali skirtis.

Patvirtinančioji faktorinė analizė dažnai taikoma, kai norima pasitikrinti, ar gerai sudarytas klausimynas. Dažniausiai klausimai grupuojami. Pavyzdžiui, klausimai apie santykius šeimoje, klausimai, padedantys atskleisti respondento impulsyvumą, ir pan. Tos spėjamos klausimų grupės yra latentiniai CFA modelio faktoriai.

Toliau šiame skyriuje nagrinėsime tik patvirtinančiąją faktorinę analizę.

2.2. CFA modelio prielaidos ir tyrimo etapai

Teorija, brolau, – sausa šaka,
Užtai gyvenimo vaisingas medis žydi.
Mefistofelis[7]

Statistikoje visur susiduriama su tam tikru dvilypumu. Viena vertus, kriterijų savybių įrodymai reikalauja labai griežtų matematinių prielaidų (daugiamačio normaliojo skirstinio ir pan.) Kita vertus, realiai gyvenime kintamieji tokių griežtų reikalavimų dažniausiai netenkina. Laimei, dalis kriterijų gan atsparūs saikingiems prielaidų pažeidimams (yra robastiški). Todėl visuotinai priimta – taikyti statistinius metodus, net jei kintamieji nevisiškai atitinka teorinius reikalavimus. Suformuluosime *praktinius* reikalavimus, keliamus faktorinės analizės duomenims. Jie daug kuo primena reikalavimus kelių analizei.

1 *Stebimieji kintamieji* (*indikatoriai*) *turi būti intervaliniai.* Idealus troškimas, kad kintamieji būtų normaliai pasiskirstę, tačiau kartais modeliuose naudojami ir ranginiai kintamieji. Kada? Mes siūlytume tokią paprastą praktinę taisyklę: jeigu matavimų vidurkis visiškai aiškiai suvokiamas (pvz., pažymių vidurkis), o skirtingų įgyjamų reikšmių yra pakankamai daug (bent penkios–septynios reikšmės)[8], tai toks kintamasis turėtų tikti.

Primename, kad yra ir kategorinių kintamųjų (lytis, religija, rasė ir pan.), kurie CFA visai netinka.

2 *Neturi būti išskirčių.* Jeigu duomenyse yra išskirčių, labai abejotina ar verta sudaryti CFA modelį. Labai išsiskiriančios reikšmės gali iškreipti koreliacijas, kurios naudojamos tikrinant modelio tinkamumą. Kai išskirtis nenatūrali (pvz., duomenų įvedimo klaida arba tarp respondentų lietuvių pakliuvęs tipiškas įstatymams paklusnus vokietis), ją iš duomenų pašaliname.

[7] V. Gėtė, *Faustas*.

[8] Surinktuose duomenyse, o ne tik klausimyne! Galima taip kvailai paklausti, kad iš 7 balų ranginės skalės respondentai rinksis tik du atsakymus. Tada kintamasis bus dvireikšmis, o jo septyniabališkumas liks neišsipildžiusia tyrėjo svajone.

3 *Duomenų kiekis.* Reikalavimas, kad būtų pakankamai stebėjimų, – grynai praktinis. Kai stebėjimų yra per mažai, negalima tikėti sudaryto modelio teisingumu. Jeigu stebėjimų yra per daug, pagrindiniai statistiniai kriterijai rodo, kad modelis netinka duomenims. Išsamiau šią problemą aptarsime 2.6 skyrelyje. Dabar tik pažymėsime, kad dažniausiai sakoma, jog stebėjimų turi būti nuo 100 iki 300. Neretai stebėjimų skaičius siejamas su vertinamų parametrų skaičiumi. Pavyzdžiui, sakoma, kad stebėjimų turi būti nuo 10 iki 20 kartų daugiau nei parametrų. Pats kukliausias CFA modeliams skirtoje literatūroje matytas reikalavimas (dar neužtikrinantis, kad faktorinės analizės kriterijai bus pritaikomi), kad stebėjimų būtų bent 60 vienetų daugiau nei kintamųjų.

Norime atkreipti dėmesį, kad suformuluoti praktiniai reikalavimai duomenims dar negarantuoja, kad bus sukurtas geras CFA modelis. Kaip ir kelių analizėje, išskiriami tokie CFA modelų sudarymo ir tyrimo etapai (žr. 3.2.2 pav.):

1 Modelio aprašymas.

2 Modelio identifikavimas.

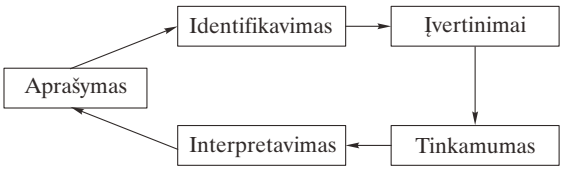
3 Parametrų įvertinimas.

4 Modelio tinkamumas duomenims.

5 Modelio interpretavimas.

6 Modelio modifikavimas.

Beje, tokie pat tyrimo etapai būdingi ir sudėtingesniems struktūrinių lygčių modeliams. Toliau kiekvieną etapą aptarsime išsamiau.



3.2.2 pav. CFA modelių sudarymo ir tyrimo schema

2.3. CFA modelio aprašymas ir identifikavimas

Svarbu ne planas, o planavimas.
G. Edvardsas

CFA modelio aprašymas – tai modelio pateikimas struktūrinių lygčių ar kelių diagramos pavidalu. Būtent šiame etape, remdamiesi teorija, sugalvojame, kokie bus latentiniai kintamieji ir su kuriais stebimais kintamaisiais (indikatoriais) jie bus susiję. CFA modeliuose stebimi kintamieji vadinami indikatoriais, nes jie atspindi (indikuoja) latentinius faktorius. Beje, CFA modeliuose kelių, jungiančių indikatorius su latentiniu faktoriumi, koeficientus įprasta vadinti *faktoriniais svoriais*. Toliau šiame skyrelyje nagrinėsime centruotus CFA modelius, t. y. tokius, kai visi indikatoriai turi nulinius vidurkius.

Struktūriniai ryšiai aprašomi lygtimis labai panašiai, kaip ir regresijos modeliai. Visos priklausomybės yra tiesinės. Pavyzdžiui, $Plan = \lambda \cdot Org + paklaida$. Jeigu vis dėlto kyla įtarimų, kad priklausomybė netiesinė, tai rekomenduotume indikatorius transformuoti. Be kėlimo laipsniu, kita tipinė teigiamų kintamųjų tranformacija yra jų natūralieji logaritmai (BVP logaritmas, eksporto apimčių logaritmas ir t. t.)

Nagrinėjant modelius, kuriuose yra latentiniai kintamieji, iškyla vadinamoji skalės problema. Aptarsime ją išsamiau. Imkime 3.2.2 pavyzdžio modelio schemos (žr. 3.2.1 pav., b) viršutinę dalį. Matome, kad organizaciniai gabumai (*Org*) atsiskleidžia diegiant naujoves (*Nauj*), planuojant darbo procesą (*Plan*) ir kontroliuojant (*Kontr*). Galime įtarti, kad naujovių diegimas priklauso nuo organizacinių gabumų, bet, žinoma, ne tik nuo jų. Visa, nuo ko dar gali priklausyti naujovių diegimas, įvardijame kaip liekamąją paklaidą e_1 ir (prisiminę, kad tiriame tik tiesines priklausomybes) užrašome tokią lygtį:

$$Nauj = \lambda_1 Org + e_1;$$

čia $e_1 \sim \mathcal{N}(0, \sigma_1^2)$, o σ_1^2 yra nežinoma liekamosios paklaidos dispersija.

Analogiškai galime užrašyti ir kintamųjų *Plan*, *Nauj* priklausomybes nuo latentinio faktoriaus. Tačiau matematiškai gautoji lygčių sistema bus neišsprendžiama vienareikšmiškai – per mažai žinome apie latentinį faktorių. Primename, kad jo reikšmės tiesiogiai nematuojamos. Tenka įvesti papildomus apribojimus. Taikomi dviejų rūšių apribojimai: 1) reikalaujama, kad latentinis faktorius turėtų konkrečią dispersiją (dažniausiai lygią vienetui), 2) vienoje iš struktūrinių lygčių koeficientą prie latentinio faktoriaus (faktorinį svorį) prilyginame vienetui.[9]

Minėti apribojimai dar vadinami skalės apribojimais. Nuo pasirinkto apribojimo priklauso rezultatų interpretavimas, keičiasi pradinė atskaitos sistema, tačiau svorių santykiai lieka nepakitę. Tarkime, kad latentinio faktoriaus dispersijos nefiksuojame ir tiriame tokią sistemą:

$$\begin{cases} Nauj = Org + e_1, \\ Kontr = \lambda_2 Org + e_2. \end{cases}$$

[9] Galima prilyginti ir 5 arba 739 ir t. t. Konkreti reikšmė nėra itin svarbi, tiesiog vienetą lengviausiai interpretuoti atitinkamose lygtyse.

Sakykime, kad kinamieji *Nauj* ir *Kontr* matuojami vienoda skale ir $\lambda_2 = 2$. Tada galima teigti, kad, *palyginti su naujovių diegimu*, organizacinio faktoriaus įtaka darbo proceso kontroliavimui yra dukart didesnė.

Jeigu fiksuotume latentinio faktoriaus dispersiją (tarkime, prilygintume ją 1), tai įvertinę sistemos

$$\begin{cases} Nauj = \lambda_1 Org + e_1, \\ Kontr = \lambda_2 Org + e_2 \end{cases}$$

parametrus (tarkime, $\hat{\lambda}_1 = 0{,}75$, $\hat{\lambda}_2 = 1{,}50$) galėtume būtent šį faktą ir paskelbti: „esant latentinio faktoriaus vienetinei dispersijai, jo svoriai...“.

Šiame vadovėlyje fiksuosime vieno iš kintamųjų faktorinį svorį. Suprantama, kad nėra labai svarbu, kurį, nes interpretavimui reikia svorių santykio. Tą pačią interpretaciją gausime ir tada, kai diegiant naujoves *Org* svoris yra 0,5, o kontroliuojant lygus 1, ir tada, kai diegiant naujoves *Org* svoris yra 1, o kontroliuojant lygus 2.

Šioje knygoje kiekvieno latentinio faktoriaus svoris bent vienoje iš lygčių lygus vienetui.

Dabar jau galime lygtimis aprašyti 3.2.2 pavyzdžio CFA modelį, kurio schema pateikta 3.2.1 paveiksle, b). Vienetinius svorius skirsime kintamiesiems *Nauj* ir *Sav*.

$$\begin{cases} Nauj = Org + e_1, \\ Kontr = \lambda_2 Org + e_2, \\ Plan = \lambda_3 Org + e_3, \\ Sav = Sprend + e_4, \\ Kom = \lambda_5 Sprend + e_5, \\ Atsak = \lambda_6 Sprend + e_6; \end{cases} \tag{3.2.1}$$

čia $e_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2, 3, 4, 5, 6$.

Norime atkreipti dėmesį, kad dažnai tiriami standartizuotų reikšmių CFA modeliai. Tokia galimybė numatyta ir statistinėse programose. Tuo atveju visų liekamųjų paklaidų ir latentinių kintamųjų dispersijos yra vienetinės, o visų kintamųjų vidurkiai lygūs nuliui. Standartizuotuose CFA modeliuose galima nustatyti, kokius indikatorius faktorius veikia stipriau, o kokius – silpniau, t. y. jie dažniausiai taikomi faktoriniams svoriams palyginti „iš akies“.

Ar latentinis kintamasis (faktorius) gali būti susijęs tik su vienu indikatoriumi? Toks modelis būtų nelogiškas, nes tuo atveju galima būtų dėti lygybės ženklą tarp indikatoriaus ir faktoriaus. O juk latentiniai faktoriai tiesiogiai nematuojami, informaciją apie juos gauname nagrinėdami koreliacijas tarp indikatorių – ieškome, kas juos sieja. Net ir du indikatoriai suteikia mažai informacijos apie latentinį faktorių. Sudarant CFA modelį, yra nusistovėję praktiniai reikalavimai kintamųjų, susietų su latentiniu faktoriumi, skaičiui:

Keturi indikatoriai, tenkantys vienam latentiniam faktoriui, – rekomenduojama, trys – priimtina, du – iš bėdos, vienas – per mažai.

Šie reikalavimai – tik orientaciniai. Net modelis su trimis indikatoriais ir vienu latentinu kintamuoju gali būti neišsprendžiamas, jei jo liekamosios paklaidos koreliuoja. Kai

modelyje yra daug latentinių kintamųjų, tai leidžiama, kad kai kurie iš jų turėtų tik du indikatorius. Nėra geras modelis, kuriame visi latentiniai faktoriai turi tik po du indikatorius. Nėra jokios taisyklės, kiek daugiausia indikatorių gali turėti latentinis faktorius. Jų skaičių dažniausiai lemia tyrimo uždaviniai. Tarkime, galima sudaryti CFA modelį, kuriame yra tik vienas latentinis kintamasis, turintis daug indikatorių.

Šio skyrelio pabaigoje aptarsime modelio identifikavimo problemą (žr. taip pat 1.3 skyrelį). Kalbėdami apie modelio identifikuojamumą, tiriame, ar duomenų aibė *teoriškai* pakankama modelio parametrams įvertinti. Tarkime, nagrinėjame paprastą vieno latentinio faktoriaus Y su dviem indikatoriais X ir Z modelį:

$$\begin{cases} X = Y + e_1, \\ Z = \beta Y + e_2; \end{cases}$$

čia $e_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2$, o X, Y, Z turi nulinius vidurkius. Nežinomi parametrai yra σ_1^2, σ_2^2, β ir $\mathbf{D}Y$. Užrašome visų *matuojamųjų* kintamųjų (indikatorių) dispersijas ir kovariacijas (jų įverčius nesunkiai rasime iš imties duomenų):

$$\begin{aligned} \mathbf{D}X &= \mathbf{D}Y + \sigma_1^2, \\ \mathbf{D}Z &= \beta^2 \mathbf{D}Y + \sigma_2^2, \\ \mathbb{C}\mathrm{ov}(X, Z) &= \beta \mathbf{D}Y. \end{aligned}$$

Taigi turime tris lygtis ir keturis vertinamus parametrus. Aišku, kad vienareikšmiškai parametrų rasti nepavyks, t. y. šis modelis neidentifikuojamas. CFA modeliams galioja ta pati identifikavimo sąlyga, kaip ir kelių analizės modeliams:

Modelis su k indikatorių identifikuojamas, jei vertinamųjų parametrų skaičius neviršija $k(k+1)/2$.

Skaičius $k(k+1)/2$ iš tikrųjų yra indikatorių kovariacijų skaičius (primename, kad literatūroje dar vadinamas duomenų taškų skaičiumi). Dažniausiai CFA modeliai yra perteklinės identifikacijos modeliai, t. y. jiems vertinamų parametrų skaičius griežtai mažesnis už $k(k+1)/2$. Tiksliai identifikuojamas modelis (t. y. modelis, turintis $k(k+1)/2$ vertinamų parametrų) naudojamas tik palyginimui su tiriamu modeliu.

CFA modelyje, aprašytame 3.2.2 pavyzdyje (3.2.1) sistema, yra 6 indikatoriai. Todėl parametrams įvertinti turėsime $6(6+1)/2 = 21$ kintamųjų kovariaciją. Vertinamųjų parametrų bus 13 – keturi faktoriniai svoriai λ_j, šešios liekamųjų paklaidų dispersijos σ_j^2, $j = 1, \ldots, 6$, dviejų latentinių faktorių dispersijos ir latentinių faktorių kovariacija $\mathbb{C}\mathrm{ov}(Org, Sprend)$. Taigi modelis yra perteklinės identifikacijos.

Skirtumas tarp $k(k+1)/2$ ir vertinamųjų parametrų skaičiaus vadinamas modelio laisvės laipsnių skaičiumi ir žymimas *df*.

Nagrinėto 3.2.2 pavyzdžio atveju $df = 21 - 13 = 8$.

2.4. CFA modelio diagrama

Geriau vieną kartą pamatyti, nei dešimt kartų išgirsti.
Operos recenzija

CFA modelį lengviau suprasti nusibraižius atitinkamą diagramą. Diagramoms braižyti naudojami tie patys elementai, kaip ir kelių analizėje, tik apskritimu arba elipse žymimos ne vien liekamosios paklaidos, bet ir latentiniai kintamieji. Primename pagrindinius diagramose naudojamus simbolius (žr. 3.2.2 lentelę).

3.2.2 lentelė. CFA modelio diagramos elementai

Žymėjimas	Reikšmė
▭	Stebimasis kintamasis (indikatorius)
○	Latentinis kintamasis, liekamoji paklaida
→	Vienpusis priežastinis ryšys
↔	Koreliacija (kovariacija)

Galime nurodyti ir formalius modelio schemos braižymo reikalavimus.

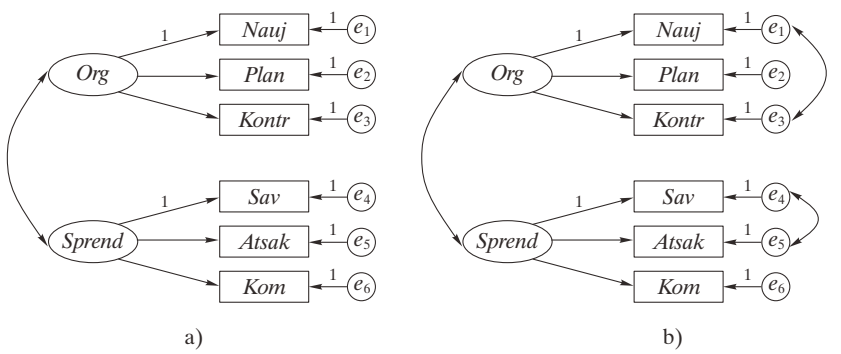
Klasikinio CFA modelio diagramos braižymo taisyklės:

1. Latentiniai faktoriai vaizduojami apskritimais (elipsėmis).
2. Indikatoriai vaizduojami stačiakampiais.
3. Liekamosios paklaidos nurodomos apskritimais.
4. Kiekvienas indikatorius tiesia rodykle susiejamas tik su vienu faktoriumi. Rodyklė brėžiama iš faktoriaus į indikatorių.
5. Kiekvieno faktoriaus rodyklės į vieną iš jo indikatorių kelio koeficientas yra lygus vienetui.
6. Kiekvienas indikatorius turi liekamąją paklaidą. Tiesios rodyklės iš paklaidos į indikatorių kelio koeficientas yra lygus vienetui.
7. Koreliuojantys latentiniai faktoriai susiejami lenkta dvipuse rodykle.
8. Koreliuojančios liekamosios paklaidos susiejamos lenkta dvipuse rodykle. Koreliuoti gali ir su skirtingais faktoriais susijusių indikatorių liekamosios paklaidos.

Penktoji ir šeštoji taisyklės yra susijusios su modelio aprašymo pasirinkimais. Penktąją taisyklę būtų galima pakeisti reikalavimu, kad latentinis faktorius turėtų vienetinę dispersiją. Analogiškai šeštojoje taisyklėje vietoj vienetinio kelio koeficiento galima pareikalauti vienetinės liekamosios paklaidos dispersijos. Be to, šeštoje taisyklėje sakoma, kad kiekvienas indikatorius turi liekamąją paklaidą. Tikslesnis reikalavimas būtų toks: liekamąsias paklaidas turi turėti visi indikatoriai, į kuriuos nukreiptos tiesios rodyklės. Tiesog klasikiniuose CFA modeliuose kitokių indikatorių nebūna. Be to, klasikiniuo-

se CFA modeliuose latentiniai faktoriai gali tik koreliuoti, bet tarp jų nėra tiesioginių (priežastinių) ryšių, t. y. jų nejungia tiesios rodyklės. Primename ankstesnio skyrelio rekomendaciją, kad kiekvienas faktorius turėtų būti susijęs su 3–4 (iš didelės bėdos – 2) indikatoriais.

3.2.3 pavyzdys. Atsižvelgdami į minėtas taisykles, galime sudaryti tokią (3.2.1) modelio diagramą, kokia pavaizduota 3.2.3 paveiksle, a). Papildykime vadybininko darbinių gebėjimų modelį dviem koreliacijomis: 1) sprendimų svarankiškumo ir atsakomybės jausmo, 2) naujovių diegimo ir gebėjimo kontroliuoti darbo procesą. Modelio su koreliuojančiomis liekamosiomis paklaidomis diagrama pavaizduota 3.2.3 paveiksle, b).



3.2.3 pav. Dvi galimos vadybininko veiklos CFA modelių diagramos: a) 3.2.2 pavyzdžio, b) 3.2.3 pavyzdžio

2.5. Matricinė CFA modelio išraiška

Iš „Matricos“ žinome, kad matrica – tai kažkas baisaus. Šis skyrelis tai dar kartą patvirtina.

Nagrinėtą (3.2.1) lygčių sistemą galima užrašyti matriciniu pavidalu:

$$\begin{pmatrix} Nauj \\ Kontr \\ Plan \\ Sav \\ Kom \\ Atsak \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ 0 & 1 \\ 0 & \lambda_5 \\ 0 & \lambda_6 \end{pmatrix} \begin{pmatrix} Org \\ Sprend \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{pmatrix}.$$

Šis užrašas tinka abiem 3.2.3 paveikslo modeliams. Todėl, norėdami modelį aprašyti tiksliau, įvedame latentinių faktorių kovariacijų matricą $\mathbf{\Phi}$ ir liekamųjų paklaidų matricą $\mathbf{\Theta}$. Modelio, pavaizduoto 3.2.3 paveiksle, a), matricos yra tokios:

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{D}Org & \mathbb{C}\mathrm{ov}(Org, Sprend) \\ \mathbb{C}\mathrm{ov}(Org, Sprend) & \mathbf{D}Sprend \end{pmatrix},$$

$$\mathbf{\Theta} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_6^2 \end{pmatrix}.$$

Modelio, pavaizduoto 3.2.3 paveiksle, b), matrica $\mathbf{\Phi}$ lieka ta pati, tačiau

$$\mathbf{\Theta} = \begin{pmatrix} \sigma_1^2 & 0 & \mathbb{C}\mathrm{ov}(e_1, e_3) & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 & 0 \\ \mathbb{C}\mathrm{ov}(e_3, e_1) & 0 & \sigma_3^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 & \mathbb{C}\mathrm{ov}(e_4, e_6) \\ 0 & 0 & 0 & 0 & \sigma_5^2 & 0 \\ 0 & 0 & 0 & \mathbb{C}\mathrm{ov}(e_6, e_4) & 0 & \sigma_6^2 \end{pmatrix}.$$

Primename, kad kovariacija yra simetrinė kintamųjų atžvilgiu $\mathbb{C}\mathrm{ov}(e_4, e_6) = \mathbb{C}\mathrm{ov}(e_6, e_4)$.

Visiškai analogiškai užrašomas ir bendrasis CFA modelis. Tarkime, kad stebime k kintamųjų, kuriuos norime susieti su m latentinių faktorių. Dėl patogumo žymenys nurodyti 3.2.3 lentelėje.

3.2.3 lentelė. Matricinio CFA modelio žymenys

Žymėjimas	Matricos eilė	Reikšmė
k		Stebimųjų kintamųjų (indikatorių) skaičius
m		Latentinių faktorių skaičius
$\mathbf{X}$	$k \times 1$	Stebimųjų kintamųjų (indikatorių) vektorius
$\boldsymbol{\xi}$	$m \times 1$	Latentinių faktorių vektorius
$\mathbf{\Lambda}$	$k \times m$	Faktorinių svorių matrica
$\mathbf{e}$	$k \times 1$	Liekamųjų paklaidų vektorius
$\mathbf{\Phi}$	$m \times m$	Latentinių faktorių kovariacijų matrica
$\mathbf{\Theta}$	$k \times k$	Liekamųjų paklaidų kovariacijų matrica
$\mathbf{\Sigma}$	$k \times k$	Indikatorių kovariacijų matrica
$\mathbf{\Sigma}_M$	$k \times k$	Teorinė modelio indikatorių kovariacijų matrica
$\mathbf{S}$	$k \times k$	Empirinė indikatorių kovariacijų matrica

Tada galima užrašyti tokį matricinį modelį:

$$\mathbf{X} = \mathbf{\Lambda}\boldsymbol{\xi} + \mathbf{e}. \tag{3.2.2}$$

Tai šiek tiek supaprastintas modelis, tarus, kad visi stebimieji kintamieji centruoti, t. y. turi nulinius vidurkius. Visiškai nesunku užrašyti ir bendresnį modelį, papildžius jį nežinomų kintamųjų vidurkių vektoriumi $\boldsymbol{\mu}$. Modelio prielaidos yra tokios: $\mathbb{C}\mathrm{ov}(\boldsymbol{\xi}, \mathbf{e}) = 0$, o $\mathbf{e}$ skirstinys – daugiamatis normalusis. Klasikiniuose CFA modeliuose faktorinių svorių matrica $\mathbf{\Lambda}$ turi tik vieną nenulinį elementą kiekvienoje eilutėje, t. y. kiekvienas indikatorius yra priežastiniu ryšiu susietas tik su vienu latentiniu faktoriumi.

Indikatorių kovariacijų matricą $\mathbf{\Sigma}$ sudaro populiacijos matuojamųjų kintamųjų kovariacijos. Kovariacijų matrica $\mathbf{\Sigma}_M$ yra kintamųjų kovariacijų matrica, kai teisingos modelio priklausomybės:

$$\mathbf{\Sigma}_M = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta}. \tag{3.2.3}$$

Idealiu atveju, jeigu modelis visiškai teisingai atspindi kintamųjų priklausomybes, $\mathbf{\Sigma}_M = \mathbf{\Sigma}$. Tiriame tik dalį populiacijos (imtį). Todėl teorinės kovariacijų matricos tėra matematinės abstrakcijos, o praktiškai iš imties duomenų apskaičiuojame jų empirinius analogus $\mathbf{S}$ (tai tradicinis žymėjimas, vartojamas vietoj $\widehat{\mathbf{\Sigma}}$) ir $\widehat{\mathbf{\Sigma}}_M$. Reikia rasti tokius $\mathbf{\Lambda}$, $\mathbf{\Phi}$ ir $\mathbf{\Theta}$ įverčius $\widehat{\mathbf{\Lambda}}$, $\widehat{\mathbf{\Phi}}$ ir $\widehat{\mathbf{\Theta}}$, kad $\widehat{\mathbf{\Sigma}}_M$ būtų kiek galima panašesnė į $\mathbf{S}$.

2.6. CFA modelio parametrų įverčiai ir modelio tinkamumas duomenims

Vienintelė teorijos alternatyva – ji neteisinga. Modelis turi trečią galimybę – jis gali būti teisingas, bet nesvarbus.

M. Eigenas

Tiksliai identifikuojamiems modeliams teorinės kovariacijų matricos įvertis (modelio kovariacijų matrica) lygus empirinei kovariacijų matricai: $\widehat{\mathbf{\Sigma}}_M = \mathbf{S}$. Tarkime, kad turime CFA modelį, pavaizduotą 3.2.4 paveiksle.

Turime 3 indikatorius. Taigi parametrams įvertinti yra $3(3+1)/2 = 6$ kovariacijos. Kita vertus, reikia įvertinti 6 nežinomus parametrus: tris liekamųjų paklaidų dispersijas, vieną faktoriaus dispersiją ir tris faktorinius svorius (kelių koeficientus).

Žinome, kad liekamosios paklaidos nekoreliuoja nei tarpusavyje, nei su faktoriais. Todėl

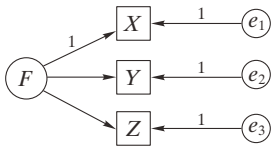
$$\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{C}\mathrm{ov}(F + e_1, \beta F + e_2) = \mathbb{C}\mathrm{ov}(F, \beta F) = \beta \mathbf{D} F.$$

Analogiškai užrašomos likusios 5 lygtys. Gauname šešių lygčių su šešiais nežinomaisiais sistemą, kurią ir sprendžiame. Vis dėlto tiksliai identifikuojami modeliai blogi tuo, kad negalima nieko pasakyti apie modelio suderinamumą su duomenimis. Todėl praktiškai daug dažniau taikomi perteklinės identifikacijos (kai parametrų yra mažiau nei indikatorių kovariacijų) modeliai. Tokius modelius toliau ir panagrinėsime.

Perteklinės identifikacijos modelių parametrų įverčiai parenkami taip, kad skirtumas tarp $\widehat{\mathbf{\Sigma}}_M$ ir $\mathbf{S}$ būtų mažiausias. Tiksliau kalbant, apibrėžiama matricų skirtumus matuojanti funkcija $F(\widehat{\mathbf{\Sigma}}_M, \mathbf{S})$ ir parametrų įverčiai parenkami taip, kad ji įgytų mažiausią reikšmę. Funkcijos $F(\widehat{\mathbf{\Sigma}}_M, \mathbf{S})$ išraiška priklauso nuo pasirinkto vertinimo metodo. Pavyzdžiui, parametrus vertinant didžiausio tikėtinumo (ML) metodu,

$$F\big(\widehat{\mathbf{\Sigma}}_M, \mathbf{S}\big) = \ln\big|\widehat{\mathbf{\Sigma}}_M\big| - \ln|\mathbf{S}| + \mathrm{tr}\big(\mathbf{S}\widehat{\mathbf{\Sigma}}_M^{-1}\big) - k;$$

čia tr žymi matricos pagrindinės įstrižainės elementų sumą, $|\mathbf{S}|$, $|\widehat{\mathbf{\Sigma}}_M|$ yra matricų determinantai, o k – analizuojamų stebimų kintamųjų (indikatorių) skaičius. Taikant apibendrintąjį



3.2.4 pav. Tiksliai identifikuojamas CFA modelis

mažiausiųjų kvadratų (GLS) metodą,

$$F(\widehat{\mathbf{\Sigma}}_M, \mathbf{S}) = 0{,}5\mathrm{tr}\big((\mathbf{S} - \widehat{\mathbf{\Sigma}}_M)\mathbf{S}^{-1}\big).$$

Abu ML ir GLS metodai nepriklauso nuo kintamųjų skalės transformacijų, t. y. transformavus kintamojo skalę (tarkime, perėjus nuo metrų prie centimetrų), lygiai taip pat keičiasi ir atitinkamas faktorinis svoris (sumažėja 100 kartų). Minėtų metodų trūkumas – jie taikytini tik normaliai pasiskirsčiusiems kintamiesiems. Šio trūkumo neturi asimptotiškai nepriklausomas nuo skirstinio (ADF[10]) metodas. Deja, ADF metodą tikslinga taikyti tik didelėms duomenų imtims.

Dažniausiai parametrų įverčiai, rasti skirtingais metodais, yra labai panašūs. Toliau pateikiamuose pavyzdžiuose visur naudojami didžiausiojo tikėtinumo metodu gauti įverčiai.

Pastaba. Konkrečiai imčiai rasti parametrų įverčius ne visada pavyksta. Įverčių skaičiavimas didžiausiojo tikėtinumo metodu – iteracinis procesas. Iš pradžių randami pradiniai įverčiai, po to jie tikslinami ir taip tęsiasi tol, kol naujieji įverčiai nuo prieš tai gautųjų praktiškai nebesiskiria. Tokie įverčių perskaičiavimai gali tęstis be galo (iteracinis procesas diverguoja). Todėl statistinėse programose yra taikomi apribojimai, fiksuojantys maksimalų leistiną iteracijų skaičių. Parametrams įvertinti tiek iteracijų ne visada pakanka.

Parametrų įverčius ne visada pavyksta rasti.

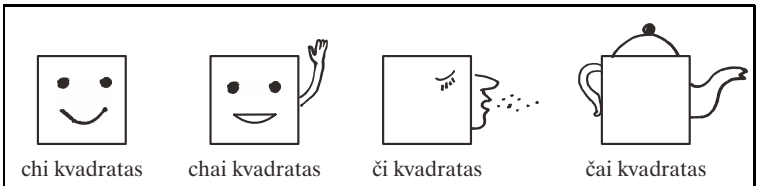
Radus parametrų įverčius, dar neaišku, ar modelis tinkamas duomenims (ar modelis ir duomenys suderinti). Iš esmės modelio tinkamumo duomenims tyrimas – tai imties kovariacijų matricos $\mathbf{S}$ lyginimas su kovariacijų matrica $\widehat{\mathbf{\Sigma}}_M$, kurią skaičiuojant atsižvelgta į modelį. Jeigu modelis teisingai aprašo kintamųjų priklausomybes, tai šios matricos nedaug skiriasi. Paprasčiausias modelio tinkamumo duomenims tikrinimo būdas – vizualus matricų skirtumo $\mathbf{S} - \widehat{\mathbf{\Sigma}}_M$ įvertinimas. Šis skirtumas vadinamas liekamųjų kovariacijų matrica. Jeigu liekamojoje kovariacijų matricoje yra daug reikšmių, kurios labai skiriasi nuo nulio, tai modelis yra netikęs. Sąvoka „labai skiriasi“ yra subjektyvi, todėl dažniausiai siūloma tirti standartizuotąją liekamąją kovariacijų matricą (t. y. liekamųjų kovariacijų matricą, kurioje kovariacijos padalytos iš standartinių paklaidų) ir žiūrėti, ar daug jos reikšmių absoliučiuoju didumu viršija 2. Tiesa, ką reiškia „daug“, paliekama spręsti pačiam tyrėjui.

3.2.4 pavyzdys. Nagrinėto 3.2.3 pavyzdyje modelio standartizuotoji liekamųjų kovariacijų matrica pateikta 3.2.4 lentelėje. Matome, kad didelių liekanų nėra.

3.2.4 lentelė. Standartizuotoji liekamųjų kovariacijų matrica

	Plan	*Kontr*	*Nauj*	*Atsak*	*Kom*	*Sav*
Plan	0,00					
Kontr	−0,10	0,00				
Nauj	0,00	0,03	0,00			
Atsak	0,11	0,56	−0,63	0,00		
Kom	−0,42	0,72	0,08	−0,65	0,00	
Sav	0,01	−0,17	0,09	0,16	0,00	0,00

[10] ADF – angl. *Asymptotically Distribution Free.*



3.2.5 pav. Kokie būna kvadratai

Vizualus liekamosios kovariacijų matricos vertinimas – pats paprasčiausias modelio tinkamumo tyrimas. Daug dažniau naudojami vadinamieji modelio suderinamumo indeksai. Nors tokių indeksų yra daugybė, beveik visi jie susiję su χ^2 (chi kvadrato, žr. 3.2.5 pav.)[11] statistika. χ^2 statistika skaičiuojama pagal formulę $\chi^2 = (n-1)F(\widehat{\mathbf{\Sigma}}_M, \mathbf{S})$; čia n – imties didumas. Maža χ^2 reikšmė yra geras rodiklis, kad modelis tinka duomenims. Tiesa, tik tada, kai duomenys yra gauti matuojant normaliai pasiskirsčiusius kintamuosius, ir duomenų pakankamai daug.

Modelio tinkamumo tikrinimo etapai:

1. Tikrinama hipotezė $\mathbf{\Sigma} = \mathbf{\Sigma}_M$.
2. Atsižvelgiama į suderinamumo indeksų reikšmes.

χ^2 reikšmė naudojama tikrinant statistinę hipotezę apie kintamųjų kovariacijų matricos ir modelio kovariacijų matricos lygybę: H_0: $\mathbf{\Sigma} = \mathbf{\Sigma}_M$ ir H_1: $\mathbf{\Sigma} \neq \mathbf{\Sigma}_M$. Primename, kad užrašas H_0: $\mathbf{\Sigma} = \mathbf{\Sigma}_M$ reiškia, jog modelis tiksliai aprašo tiriamas kintamųjų priklausomybes. Todėl patogiau šią hipotezę užrašyti žodžiais.

Bendrasis modelio tinkamumas:

$$\begin{cases} H_0\text{: modelis tinkamas duomenims,} \\ H_1\text{: modelis duomenims netinkamas.} \end{cases}$$

Jeigu p reikšmė ne mažesnė už α, čia α – reikšmingumo lygmuo, tai modelis duomenims tinkamas.

3.2.5 pavyzdys. Tegul reikšmingumo lygmuo $\alpha = 0{,}05$. Patikrinę hipotezę apie 3.2.2 pavyzdžio (žr. 3.2.3 pav., a) modelio kovariacijų matricų lygybę, gavome, kad $\chi^2 = 5{,}4$, $p = 0{,}7$. Patikrinę analogišką 3.2.3 pavyzdžio (žr. 3.2.3 pav., b) modelio hipotezę, gavome, kad $\chi^2 = 4{,}6$, $p = 0{,}6$. Darome išvadą, kad bendrasis abiejų modelių tinkamumas duomenims yra pakankamai geras.

Sakydami, kad patvirtinome gerą *bendrąjį* modelio tinkamumą duomenims, turime omenyje, jog nežinome, ar visi modelio dėmenys vienodai svarbūs ir reikalingi. Tiesiog konstatavome, kad modelyje jų yra pakankamai. Jeigu modelyje buvo nereikalingas kelias (nulinis faktorinis svoris), išvadose apie bendrąjį modelio tinkamumą tai neatsispindėjo.

[11] Ne či ir ne čai (arbatos) kvadrato! Beje, anglų kalboje taip pat tariama „chai“ arba „kai“.

Nereikia statistinių išvadų apie kovariacijų matricų lygybę suabsoliutinti. Šios išvados labai priklauso nuo stebėjimų skaičiaus n. Jeigu duomenų daug – nulinę hipotezę atmesime ir galbūt visai be reikalo „išbrokuosime“ gerą modelį. Ir priešingai, mažas duomenų skaičius suteiks neteisingą iliuziją, jog modelis geras. Pateiksime vieną *negriežtą* rekomendaciją, kada verta remtis statistinėmis išvadomis apie $\mathbf{\Sigma}$ ir $\mathbf{\Sigma}_M$ lygybę.

Statistinėmis išvadomis apie $\mathbf{\Sigma}$ ir $\mathbf{\Sigma}_M$ lygybę įprasta remtis, jei $100 \leqslant n \leqslant 200$.

Kitais atvejais gautas statistines išvadas reikia vertinti atsargiai. Žinoma, viskas dar priklauso nuo modelio sudėtingumo. Pavyzdžiui, turint paprastus modelius, galima tikrinti hipotezę apie χ^2 reikšmę, kai $n = 75$. Kita vertus, jeigu $n = 3000$, tai praktiškai neįmanoma įrodyti, kad modelio ir empirinė kovariacijų matricos statistiškai reikšmingai nesiskiria. Tuo atveju siūlytume tiesiog vizualiai palyginti pačias matricas arba pasirinkti toliau aprašytus suderinamumo indeksus *RMSEA*, *IFI*, *TLI*.

Hipotezė apie kovariacijų matricų lygybę – tik vienas modelio tinkamumo duomenims tikrinimo etapų. Dar reikia atsižvelgti į suderinamumo indeksų reikšmes. Dažniausiai siūloma indeksus vertinti „iš akies“. Bėda, kad statistikai iki šiol nesusitaria, kokios indeksų reikšmės rodo gerą tinkamumą[12]. Pačių indeksų yra daugybė. Pateikiame 3.2.5 lentelę, kurioje nurodyti tik populiariausi indeksai. Kuriuos iš jų skaičiuoti – pasirenka pats tyrėjas. Nebūtina aprašyti visus indeksus. Užtenka trijų keturių.

3.2.5 lentelė. Suderinamumo indeksai

Suderinamumo indeksas	Gera reikšmė	Priimtina reikšmė
NC	$NC \leqslant 2$	$NC \leqslant 3$
RMSEA	$RMSEA \leqslant 0{,}05$	$RMSEA \leqslant 0{,}08$
SRMR	$SRMR \leqslant 0{,}05$	$SRMR \leqslant 0{,}08$
AGFI	$AGFI \geqslant 0{,}95$	$AGFI \geqslant 0{,}90$
IFI	$IFI \geqslant 0{,}95$	$IFI \geqslant 0{,}90$
TLI	$TLI \geqslant 0{,}95$	$TLI \geqslant 0{,}90$
CFI	$CFI \geqslant 0{,}95$	$CFI \geqslant 0{,}90$

Panagrinėsime suderinamumo indeksus išsamiau. Beveik visi jie susiję su nagrinėjamo modelio χ^2 statistikos reikšme (žymėsime ją χ^2_m) ir nulinio modelio χ^2 statistikos reikšme (žymėsime ją χ^2_0). Nuliniu vadinamas modelis, kurio visos matuojamų kintamųjų kovariacijos yra lygios 0. Atitinkamai simboliais df_m ir df_0 žymėsime modelių laisvės laipsnius. Primename, kad modelio laisvės laipsnių skaičius – tai visų kovariacijų skaičiaus ir vertinamų parametrų skirtumas. Modelio, kuriame yra k matuojamų kintamųjų ir p vertinamų parametrų, $df = k(k+1)/2 - p$.

Imties didumą, t. y. visą duomenų skaičių, žymėsime simboliu n.

Visi suderinamumo indeksai turi trūkumų. Pavyzdžiui, nuliniame modelyje yra netikroviška prielaida, kad visi matuojami kintamieji nekoreliuoja. Palyginus tiriamą modelį su nerealiu nuliniu ir gavus skirtumą, nėra ko džiaugtis. Tai primena liaudyje paplitusį vedusių vyrų grožio vertinimą: „Už velnią gražesnis – jau ir gerai“. Todėl dalis tyrėjų[13]

[12] Matyt, statistikų akys labai skirtingos.

[13] Jie paprastai pasiūlo savo indeksus.

indeksus, susijusius su nuliniu modeliu, vertina skeptiškai.

Normuotasis chi kvadratas (*NC*[14]). Šis indeksas – tai χ^2_m statistikos reikšmė, padalyta iš modelio laisvės laipsnių, t. y.

$$NC = \frac{\chi^2_m}{df_m}.$$

NC gali būti didelis vien dėl didelio duomenų skaičiaus n. Jį rekomenduojama taikyti tik vidutinio didumo imtims ($n \approx 200$). Nurodytos 3.2.5 lentelėje *NC* reikšmės gana konservatyvios. Kai kurie itin liberalūs statistikai teigia, kad pakankamai gerai, jei $NC < 5$. Šiuo metu indeksas *NC* naudojamas gana retai.

Kvadratinė šaknis iš vidutinės aproksimacijos paklaidos (*RMSEA*[15]). Šis indeksas randamas pagal tokią formulę:

$$RMSEA = \sqrt{(\chi^2_m - df_m)/[(n-1)df_m]}.$$

Atkreipiame dėmesį, kad $RMSEA = 0$, kai $\chi^2_m = df_m$, t. y. ne tada, kai $\chi^2 = 0$ ($\widehat{\mathbf{\Sigma}}_M = \mathbf{S}$). *RMSEA* turi necentrinį chi kvadrato skirstinį, dažnai surandamas šio indekso pasikliautinasis intervalas. Kartais šis intervalas naudojamas kaip statistinės hipotezės apie modelio tinkamumą duomenims pakaitalas. Jei apatinis 90 % pasikliautinojo intervalo rėžis mažesnis už 0,05, o viršutinis – mažesnis už 0,10, tai interpretuojama kaip pakankamas gero modelio ir duomenų suderinamumo įrodymas.

Kai kurie tyrimai rodo, kad *RMSEA* reikšmę nelabai veikia duomenų skaičius n, nors esant mažoms imtims *RMSEA* yra „per geras“, t. y. įgyja nepagrįstai mažą reikšmę. *RMSEA* sudarytas taip, kad iš jo matyti modelio sudėtingumas. Nagrinėjant du panašaus tinkamumo duomenims modelius, paprastesniojo modelio *RMSEA* mažesnis. Vienas iš *RMSEA* indekso privalumų – jis priklauso tik nuo tiriamo modelio savybių. *RMSEA* yra vienas iš pačių populiariausių suderinamumo indeksų ir pateikiamas beveik kiekviename tyrime.

Kvadratinė šaknis iš standartizuotosios vidutinės liekanos (*SRMR*[16]). Šio indekso sudarymo logika aiški – reikia atsižvelgti į empirinės bei įvertintos modelio kovariacijų matricų skirtumus ir padaryti taip, kad kintamųjų matavimo skalės neturėtų įtakos indekso reikšmei. Nerašome tikslios indekso skaičiavimo formulės, nes ji ganėtinai gremėzdiška.

[14] NC – angl. *Normed Chi-square.*

[15] RMSEA – angl. *Root Mean Square Error of Approximation.*

[16] SRMR – angl. *Root Mean Square Residual.*

Pateiksime tik indekso skaičiavimo algoritmą: a) kiekvienas liekamosios kovariacijų matricos $\mathbf{S} - \widehat{\mathbf{\Sigma}}_M$ elementas padalijamas iš to elemento standartinės paklaidos, b) gautosios reikšmės keliamos kvadratu ir sudedamos, c) iš gautosios sumos ištraukiama kvadratinė šaknis. *SRMR* trūkumas – indeksas mažas, jeigu modelyje daug vertinamų parametrų arba didelė duomenų imtis. Privalumas – nepriklauso nuo nulinio modelio.

Pataisytasis modelio suderinamumo indeksas (*AGFI*[17]). Šis indeksas atsparus kintamųjų matavimų skalės pokyčiams ir skaičiuojamas pagal tokią formulę:

$$AGFI = 1 - \frac{\chi_m^2}{\chi_0^2} \cdot \frac{p(p+1)}{df}.$$

Nors maksimali *AGFI* reikšmė gali viršyti 1, jo reikšmes interpretuojant laikomasi nuostatos, kad modelis tinkamas duomenims, kai $AGFI > 0{,}95$. Jei $AGFI < 0$, tai modelis duomenims visiškai netinkamas. Didėjant duomenų skaičiui n, indekso *AGFI* reikšmė didėja. Esant mažoms imtims, šis indeksas gali būti nepagrįstai mažas. Šiuo metu statistiniuose tyrimuose *AGFI* naudojamas palyginti retai.

Prieaugių suderinamumo indeksas (*IFI*[18]). Šis indeksas skaičiuojamas pagal tokią formulę:

$$IFI = \frac{\chi_0^2 - \chi_m^2}{\chi_0^2 - df_m}.$$

Kartais *IFI* viršija 1. Šis indeksas mažai jautrus imties didumui. Matome, kad indeksas remiasi nulinio ir tiriamo modelių palyginimu. Nesunku šį indeksą pakeisti ir pritaikyti dviem nenuliniams CFA modeliams palyginti. Tam pakanka nulinio modelio chi kvadrato statistikos reikšmę χ_0^2 pakeisti antrojo modelio chi kvadrato statistikos reikšme.

Takerio–Liuiso indeksas (*TLI*[19]). Šis indeksas dar vadinamas *nenormuotuoju suderinamumo indeksu* (*NNFI*[20]). *TLI* taikomas dviem modeliams palyginti. Modelio ir duomenų suderinamumą jis parodo, kai tiriamasis modelis lyginamas su nuliniu modeliu. *TLI* skaičiuojamas pagal tokią formulę:

$$TLI = \frac{\chi_0^2/df_0 - \chi_m^2/df_m}{\chi_0^2/df_0 - 1}.$$

Kaip ir *IFI*, *TLI* mažai priklauso nuo imties didumo.

Sąlyginis suderinamumo indeksas (*CFI*[21]). Skaičiuojamas pagal tokią formulę:

$$CFI = 1 - \frac{\chi_m^2 - df_m}{\chi_0^2 - df_0}.$$

Didesnė idekso reikšmė rodo geresnį modelio tinkamumą duomenims, tačiau $CFI = 1$ dar neužtikrina visiško modelio ir duomenų suderinamumo (t. y., kad bus teisinga lygybė $\widehat{\mathbf{\Sigma}}_M = \mathbf{S}$). Indeksas mažai priklauso nuo imties didumo.

---

[17] AGFI – angl. *Adjusted Goodness-of-Fit Index.*

[18] IFI – angl. *Incremental Fit Index.*

[19] TLI – angl. *Tucker-Lewis Index.*

[20] NNFI – angl. *Nonnormed Fit Index.*

[21] CFI – angl. *Comparative Fit Index.*

3.2.6 pavyzdys. Tirkime 3.2.2 pavyzdžio (žr. 3.2.3 pav., a) ir 3.2.3 pavyzdžio (žr. 3.2.3 pav., b) CFA modelius. Suderinamumo indeksai, apskaičiuoti AMOS programa, pateikiami 3.2.6 lentelėje. Iš jos matyti, kad abu modeliai gerai suderinti su duomenimis.

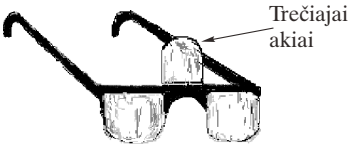
3.2.6 lentelė. Nagrinėtų 3.2.2 ir 3.2.3 pavyzdžių suderinamumo indeksai

Suderinamumo indeksas	Indekso reikšmė	
	3.2.2 modelis	3.2.3 modelis
NC	0,67	0,73
RMSEA	0,00	0,00
SRMR	0,03	0,03
AGFI	0,97	0,97
IFI	1,01	1,01
TLI	1,02	1,02
CFI	1,00	1,00

Visi modelio suderinamumo indeksai rodo tik *bendrąjį* (vidutinį) modelio ir duomenų suderinamumą (modelio tinkamumą duomenims). Net pakankamai didelė (maža) indekso reikšmė negarantuoja, kad *visos* modelio dalys bus gerai suderintos su duomenimis.

Suderinamumo indeksai atspindi tik *vidutinį* modelio tinkamumą duomenims. Jie neatskleidžia ar reikalingi *visi* modelio keliai ir visos koreliacijos.

Nors indeksai ir labai svarbūs, tačiau tinkamų jų reikšmių dar nepakanka, kad modelį laikytume geru. Pavyzdžiui, optikos salonui paskelbus prekybinę akciją ir pradėjus pardavinėti akinius su trimis stiklais už įprastų akinių kainą, naujasis modelis būtų puikiai suderintas su trumparegystės korekcija. Vis dėlto, kažin ar toks akinių modelis kam nors tiktų, išskyrus šamanų pacientus su „atverta trečiąja akimi“.



2.7. CFA modelio parametrų įverčių analizė

Net ir gerai duomenims tinkantis modelis negarantuoja, kad jame nėra nereikšmingų kintamųjų arba kelių. Prieš darant galutinę išvadą apie modelio tinkamumą, reikia ištirti ir pačių parametrų įverčius.

CFA modelio parametrų įverčių tyrimo etapai:

1. Vertinami parametrų ženklai.
2. Vertinami parametrų didumai.
3. Tikrinamos hipotezės, ar parametrai yra statistiškai reikšmingi.

Kaip jau minėta ankstesniame skyrelyje, perteklinės identifikacijos modelių parametrai skaičiuojami didžiausio tikėtinumo, apibendrintojo mažiausiųjų kvadratų ir kt. metodais. Ištyrę gautus parametrų įverčius, padarome tam tikras išvadas apie modelį.

Parametrų įverčiai privalo turėti tinkamus ženklus. Jeigu jie ne tokie, kaip nurodo atitinkamo dalyko teorija, tai arba modelis, netinkamas arba teorija neteisinga.

3.2.7 pavyzdys. Tarkime, kad sudarėme vadybinio darbo CFA modelį, kurio diagrama pavaizduota 3.2.6 paveiksle, a). Šis modelis nelabai logiškas, nes teigiame, kad organizaciniai gabumai veikia amžių (logiškiau būtų manyti priešingai). Naudodamiesi AMOS programa, didžiausiojo tikėtinumo metodu radę parametrų įverčius, gauname, kad

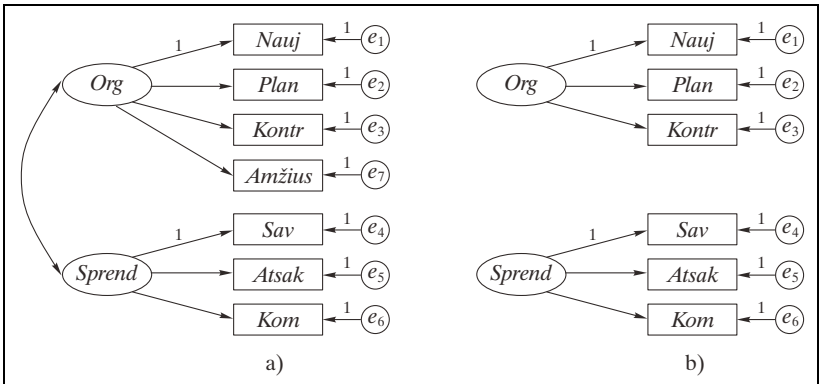
$$Amžius = -2{,}38Org + e_7.$$

Kitaip tariant, dideli organizaciniai gebėjimai vadybininką jaunina. Toks teiginys visiškai nelogiškas, taigi modelis netinkamas.

Kartais netinkami ženklai atsiranda ne todėl, kad modelis nelogiškas, o dėl to, kad indikatoriai per stipriai koreliuoja. Vis dėlto net ir tokiu atveju modelį reikia keisti. Norime atkreipti dėmesį ir į tai, kad struktūrinių lygčių parametrai skaičiuojami atsižvelgiant į kintamųjų tarpusavio ryšius, dėl to galima gauti didesnius už 1 (arba mažesnius už –1) koreliacijų įverčius arba neigiamus dispersijos įverčius. Neįprastai didelė (maža) parametro įverčio reikšmė taip pat rodo, kad modelis duomenims netinka.

Neigiami dispersijų įverčiai arba didesnės už 1 (mažesnės už –1) koreliacijos – tai ženklas, kad modelis duomenims netinka.

3.2.8 pavyzdys. Nagrinėkime vadybininko darbo CFA modelį be latentinių faktorių koreliacijos. Jo diagrama pavaizduota 3.2.6 paveiksle, b). Didžiausiojo tikėtinumo metodu gautas e_4 dispersijos įvertis $\hat{\sigma}^2 = -0{,}82$. Taigi pakeitus tik vieną kintamųjų priklausomybę, modelis iš esmės pablogėjo.



3.2.6 pav. Blogai duomenims tinkantys CFA modeliai

Gavus parametrų įverčius, tikrinamos hipotezės, ar jie skiriasi nuo nulio. Taikoma statistika – parametro įvertinys, padalytas iš to įvertinio standartinės paklaidos. Tokia statistika turi asimptotinį skirstinį. Nulinė hipotezė – parametras lygus nuliui. Taigi neatmesta nulinė hipotezė rodo, kad modelį, ko gero, vertėtų tobulinti.

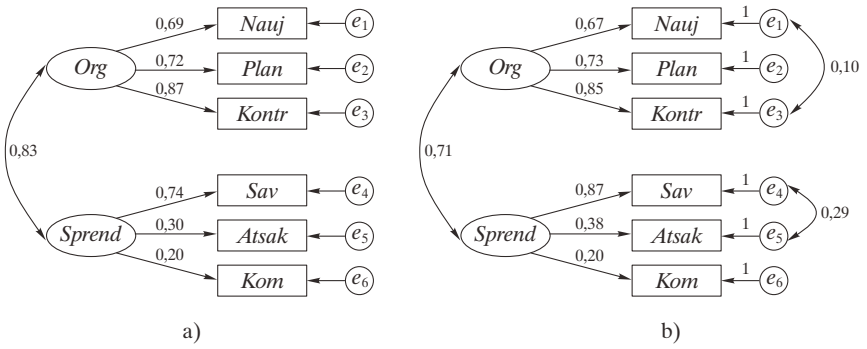
Hipotezė apie parametrą:

$$\begin{cases} H_0\text{: parametras} = 0, \\ H_1\text{: parametras} \neq 0. \end{cases}$$

Parametras statistiškai reikšmingai skiriasi nuo 0, jei p reikšmė mažesnė už α; čia α – pasirinktasis reikšmingumo lygmuo.

Primename, kad dažniausiai $\alpha = 0{,}05$.

3.2.9 pavyzdys. Ankstesniame skyrelyje įsitikinome, kad 3.2.2 ir 3.2.3 pavyzdžiuose nagrinėtų CFA modelių bendrieji suderinamumo indeksai praktiškai yra vienodi. Jie parodo, kad modeliai duomenims tinka. Standartizuotieji modeliai su faktoriniais svoriais ir koreliacijomis pavaizduoti 3.2.7 paveiksle. Ištyrę parametrų įverčių statistinį reikšmingumą, įsitikiname, kad 3.2.2 pavyzdžio modelio visi parametrai statistiškai reikšmingai (kai reikšmingumo lygmuo yra 0,05) skiriasi nuo nulio. Tuo tarpu 3.2.3 pavyzdžio modelio statistiškai nereikšmingos buvo abi liekamųjų paklaidų koreliacijos (tarp e_1 ir e_3 bei tarp e_4 ir e_6) ir kelio tarp latentinio faktoriaus *Sprend* ir indikatoriaus *Kom* koeficientas (faktorinis svoris). Todėl galima padaryti išvadą, kad 3.2.3 pavyzdžio CFA modelis nėra geras.



3.2.7 pav. Standartizuotieji CFA modeliai: a) 3.2.2 pavyzdžio, b) 3.2.3 pavyzdžio

Statistinių hipotezių išvados labai priklauso nuo stebėjimų skaičiaus. Jeigu stebėjimų pakankamai daug, tai labai tikėtina gauti statistiškai reikšmingą parametro skirtumą nuo nulio. Nagrinėto 3.2.7 pavyzdyje modelio visų parametrų įverčiai statistiškai reikšmingai skyrėsi nuo nulio.

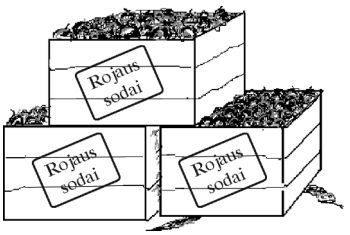
Statistiškai reikšmingi parametrų įverčiai dar negarantuoja modelio gerumo.

2.8. CFA modelio interpretavimas ir tobulinimas

Ten, kur nėra prasmės nė lašo,
Ją kuo puikiausiai žodis pavaduos.
Mefistofelis[22]

Interpretuojant gautus rezultatus, reikia nepamiršti, kad: a) išvados daromos apie populiaciją, b) išvados neturi prieštarauti logikai, c) norint palyginti faktoriaus įtaką skirtingiems indikatoriams, reikia atsižvelgti į standartizuotojo modelio svorius.

Tyrėjus domina apibendrinančios išvados, pavyzdžiui, „vadybininkų organizaciniai gebėjimai planuoti“, o ne išvados vien apie tirtus respondentus – „Adomo ir Ievos organizaciniai gebėjimai planuoti vaisių surinkimą rojaus sode“ (o kiti tegul išsitiria patys).



Jeigu duomenys neprieštarauja pasirinktame modelyje aprašytiems priežastiniams ryšiams, tai dar nereiškia, kad jie tokie ir yra. Formaliai kalbant, modelyje buvo numatytos įvairios koreliacijos ir duomenys jų nepaneigė. Vis dėlto koreliacija dar nereiškia priežastingumo. Pavyzdžiui, jeigu sumodeliuosime, kad ankstyvas infarktas senatvėje sąlygoja antsvorį jaunystėje, tai duomenys greičiausiai tam neprieštaraus, nes abu reiškiniai koreliuos. Nesunku suprasti, kad priežastinis ryšys yra visiškai priešingas, t. y. antsvoris jaunystėje sąlygoja infarkto riziką senatvėje. Taigi kaip priežastinius ryšius pagrįsti – visų pirma yra modelio sudarytojo, o ne statistikos (CFA modeliavimo) problema.

Faktoriniai svoriai parodo, kiek vienetų padidės (sumažės) indikatoriaus reikšmė, faktoriui padidėjus vienetu. Faktorius yra nelabai apčiuopiamas kintamasis (ne veltui pavadintas latentiniu), todėl faktoriniai svoriai dažniausiai naudojami tik palyginimams. Jie padeda nustatyti, kurį indikatorių faktorius veikia labiau. Tačiau faktoriniai svoriai tinka tik tada, kai visi indikatoriai matuojami toje pačioje skalėje. Netinka lyginti faktorinius svorius, jeigu vienas indikatorius matuojamas metrais, o kitas – milimetrais. Indikatorių skalės suvienodinamos standartizuotame CFA modelyje, t. y. kai tariama, kad visi latentiniai kintamieji ir liekamosios paklaidos turi vienetines dispersijas. Todėl sprendžiant, kuriuos indikatorius latentinis faktorius veikia stipriau, dažniausiai lyginami standartizuotieji faktoriniai svoriai.

3.2.10 pavyzdys. Atsižvelgę į 3.2.7 paveiksle, a) pateiktus standartizuotuosius faktorinius svorius, darome išvadą, kad organizaciniai gabumai gerai atsiskleidžia diegiant naujoves, planuojant ir kontroliuojant darbą (visi faktoriniai svoriai dideli – nuo 0,69 iki 0,87). Tuo tarpu sprendimų priėmimo faktorius labai susijęs su savarankiškais sprendimais (faktorinis svoris yra 0,74) ir gana silpnai su komunikabilumu (svoris lygus 0,20)

[22] V. Gėtė, *Faustas*.

bei atsakomybe už priimtus sprendimus (svoris yra 0,30). Peršasi mintis, kad galbūt reikėtų iš modelio pašalinti komunikabilumą.

Jeigu modelis netinka, dažniausiai kuriame naują ir tikriname viską iš pradžių, t. y. modelį tobuliname. Teoriškai (bet ne realybėje) taip elgdamiesi neturėtume atsižvelgti į ankstesnių modelių tikrinimo rezultatus. Kodėl? Tiesiog iškreiptume suderinamumo testus. Aiškiaregys atspėja būsimų rungtynių rezultatą. Žmogus, atspėjantis praėjusių rungtynių rezultatą, vadintinas kitaip. Mūsų tikslas – atspėti galimas kintamųjų priklausomybes dar prieš surenkant duomenis. Jei spėjame jau žinomą surinktų duomenų koreliaciją, tai, žinoma, pataikome, bet negalime teigti, kad tokia koreliacija būdinga *visai* populiacijai.

Realiuose tyrimuose modelis tobulinamas tol, kol gerai aprašo turimus duomenis. Yra keletas formalių metodų, palengvinančių modelio modifikavimą. Juos išsamiau aptarsime nagrinėdami bendrųjų struktūrinių lygčių modelius. Vis dėlto indikatoriai turėtų būti parenkami vadovaujantis logika, o ne formaliomis statistinėmis išvadomis. Be to, reikėtų turėti atsarginę imtį (imties dalį), su kuria ir galima patikrinti, ar sudarytasis modelis tinkamai aprašo tiriamą reiškinį.

Tobulinimo procesas neįmanomas be modelių palyginimo. Lyginami tiems patiems duomenims pritaikyti modeliai.

Lyginant CFA modelius, atsižvelgiama į jų:

1) bendrąjį tinkamumą duomenims,
2) parametrų įverčius ir modelio logiškumą,
3) Akaikės ar kito informacinio indekso reikšmę,
4) statistinės hipotezės apie skirtingą modelių suderinamumo laipsnį rezultatus.

Pirmieji du reikalavimai – akivaizdūs. Sudarę naują modelį, turime patikrinti, ar jam galioja nulinė hipotezė $\mathbf{\Sigma} = \mathbf{\Sigma}_M$. Be to, reikia įsitikinti, ar suderinamumo indeksai pakankamai geri, o parametrų įverčiai neprieštarauja logikai. Jeigu vieno modelio indeksų reikšmės blogos (pvz., *RMSEA* = 0,20), o kito – ne (pvz., *RMSEA* = 0,01), tai nesunku suprasti, kuris modelis tinkamesnis. Primename, kad geresnis suderinamumas su duomenimis – modelio privalumas, bet ne savitikslis. Galutinai nuspręsti, kuris modelis tinkamiausias, galima tik vadovaujantis logika ir tiriamo dalyko teorija.

Jeigu nagrinėdami tuos pačius duomenis turime du pakankamai gerai tinkančius modelius – renkamės mažiau sudėtingą. Vis dėlto nagrinėjant vien modelių diagramas tai padaryti nėra lengva. Pavyzdžiui, kaip nuspręsti, ar tapo sudėtingesnis modelis, koreliacija tarp indikatorių pakeitus koreliacija tarp latentinių faktorių. Kiekybiškai modelio sudėtingumą ir tinkamumą duomenims aprašo informaciniai indeksai. Juos jau taikėme antroje šios knygos dalyje HLM modeliams palyginti. Populiariausias iš šių indeksų – Akaikės informacinio kriterijaus (*AIC*) reikšmė.

Kuo mažesnė *AIC* reikšmė – tuo geriau duomenims tinka modelis.

3.2.11 pavyzdys. Nagrinėto 3.2.2 pavyzdžio $AIC = 31{,}39$, 3.2.3 pavyzdžio – $AIC = 34{,}36$. Taigi dar kartą įsitikiname, kad 3.2.2 pavyzdžio modelis geresnis.

Gana dažnai taikomas tikėtino kryžminio pagrįstumo (*ECVI*[23]) indeksas. Tai faktiškai tas pats *AIC*, tik padalytas iš imties didumo, todėl jo atskirai nenagrinėsime. Mažesnė *ECVI* reikšmė atitinka geriau su duomenimis suderintą modelį.

Žinome, kad bendrąjį modelio tinkamumą duomenims parodo maža χ^2 statistikos reikšmė. Galime laikyti vieną modelį statistiškai reikšmingai geresniu už kitą, jeigu jo χ^2 reikšmė statistiškai reikšmingai mažesnė. Tokią hipotezę galima patikrinti, kai vienas modelis yra kito modelio dalis.

Tariama, kad vienas modelis yra kito modelio dalis (redukuotasis modelis), jeigu jis gautas pašalinus dalį modelio kelių ir koreliacijų (arba fiksavus atitinkamas parametrų reikšmes). Pavyzdžiui, 3.2.2 modelis yra 3.2.3 modelio dalis (žr. 3.2.7 pav.).

Aptarsime tikrinamąją statistinę hipotezę išsamiau. Tarkime, kad pirmasis (redukuotasis) modelis yra antrojo modelio dalis. Modelių chi kvadrato statistikų reikšmės yra atitinkamai χ_1^2 ir χ_2^2, o laisvės laipsniai – df_1 ir df_2. Tikrinama statistinė hipotezė:

$$\begin{cases} H_0\colon \text{ abu modeliai duomenims tinka vienodai,} \\ H_1\colon \text{ abu modeliai duomenims tinka nevienodai.} \end{cases}$$

Priimamas sprendimas priėmimas priklauso nuo to, ar galima iš imties duomenų apskaičiuoti p reikšmę. Tegul pasirinktasis reikšmingumo lygmuo lygus α. Tada H_0 atmetame ir darome išvadą, kad modeliai tinka duomenims nevienodai, jeigu $p < \alpha$. Jeigu $p \geqslant \alpha$, tai nusprendžiame, kad abu modeliai duomenims tinka vienodai.

Jeigu p reikšmės negalime apskaičiuoti, tai reikia ieškoti chi kvadrato su $df_1 - df_2$ laisvės laipsnių α lygmens kritinės reikšmės $\chi_\alpha^2(df_1 - df_2)$. Jeigu $\chi_D^2 = \chi_1^2 - \chi_2^2 > \chi_\alpha^2(df_1 - df_2)$, darome išvadą, kad modeliai tinka duomenims nevienodai.

Jeigu sudėtingesnio modelio χ^2 reikšmė statistiškai reikšmingai nesiskiria nuo paprastenio modelio χ^2 reikšmės, tai negalime teigti, kad sudėtingesnis modelis geresnis.

Žinoma, kalbame apie formalų statistinį modelių palyginimą. Gali būti, kad atliekant kitus tyrimus buvo nustatytas sudėtingesnio modelio pagrįstumas, o bendrojo suderinamumo lygių skirtumas statistiškai nereikšmingas tik dėl mažos imties.

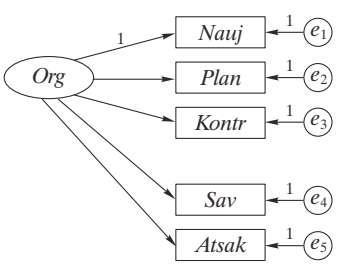
3.2.12 pavyzdys. Tegul reikšmingumo lygmuo $\alpha = 0{,}05$. Palyginsime, ar 3.2.2 ir 3.2.3 pavyzdžių modeliai vienodai gerai tinka duomenims. 3.2.2 pavyzdžio modelio $\chi_1^2 = 5{,}4$, $df_1 = 8$, o 3.2.3 pavyzdžio modelio $\chi_2^2 = 4{,}4$, $df_2 = 6$. Iš kritinių reikšmių lentelių randame $\chi_{0,05}^2(2) = 5{,}9$. Kadangi $\chi_D^2 = 5{,}4 - 4{,}4 = 1 < 5{,}9$, tai darome išvadą, jog 3.2.3 modelis nėra duomenims tinkamesnis.

Visos statistinės išvados priklauso nuo imties didumo. Esant didelėms imtims, net maži modelių skirtumai bus statistiškai reikšmingi. Šito nederėtų užmiršti ir tikrinant hipotezę apie bendrąjį modelio tinkamumą duomenims.

Sudarant CFA modelius, patariama tyrimą pradėti nuo vieno latentinio faktoriaus modelio. Jeigu duomenys neprieštarauja tokiam modeliui, siūloma pamąstyti, ar tikrai reikia sudėtingesnio daugiafaktorinio modelio. Beje, tokiu atveju sudėtingesnio modelio latentiniai faktoriai stipriai koreliuos.

[23] ECVI – angl. *Expected Cross-Validation Index.*

3.2.13 pavyzdys. Vieno latentinio faktoriaus vadybininko darbo modelis su standartizuotais svoriais pateiktas 3.2.8 paveiksle. Visos suderinamumo su duomenimis charakteristikos yra geros ($\chi^2 = 6{,}7$ statistiškai reikšmingai nesiskiria nuo nulio, $p = 0{,}67$, $NC = 0{,}75$, $RMSEA = 0{,}00$, $TLI = 1$, $AGFI = 0{,}97$ ir t. t.). Taigi galima būtų pamąstyti ir apie vienfaktorinį CFA modelį.



3.2.8 pav. Vieno faktoriaus modelis

Vis dėlto, ar vienfaktorinį CFA modelį rinktis, ar nesirinkti, priklauso tik nuo tyrėjo. Modeliai sudaromi tam tikro *vieno* reiškinio įvairiems aspektams tyrinėti, todėl neatmestas vienfaktorinis CFA modelis gali tai patvirtinti. Beje, yra situacijų, kai vienfaktorinį CFA modelį numato teorija. Pavyzdžiui, modeliuojant latentinio šešėlinės ekonomikos faktoriaus įtaką BVP, nedarbo lygiui ir kitiems ekonominiams rodikliams.

kelio koeficientas	modelio identifikavimas	suderinamumo indeksai
liekamoji paklaida	modelio laisvės laipsniai	patvirtinančioji faktorinė analizė
latentinis faktorius	standartizuotas modelis	

UŽDAVINIAI

1. Nagrinėjame tiksliai identifikuojamą CFA modelį, kurio diagrama pavaizduota 3.2.4 pav. Kintamųjų X, Y ir Z vidurkiai yra lygūs 0, o kovariacijų matrica pateikta 3.2.7 lentelėje. Surasti nežinomų parametrų įverčius.

3.2.7 lentelė. Kintamųjų X, Y, Z kovariacijų matrica

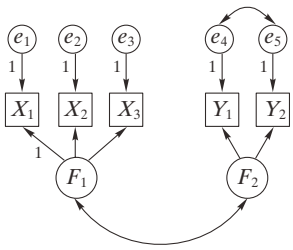
	X	Y	Z
X	73	63	53
Y	63	71	78
Z	53	78	104

2. Nagrinėjame tiksliai identifikuojamą CFA modelį, kurio diagrama pavaizduota 3.2.4 paveiksle. Kintamųjų X, Y ir Z, koreliacijų matrica pateikta 3.2.8 lentelėje.

3.2.8 lentelė. Kintamųjų X, Y, Z koreliacijų matrica

	X	Y	Z
X	1	0,88	0,61
Y	0,88	1	0,91
Z	0,61	0,91	1

3. Matricomis užrašykite CFA modelį, kurio diagrama pavaizuota 3.2.9 paveiksle.

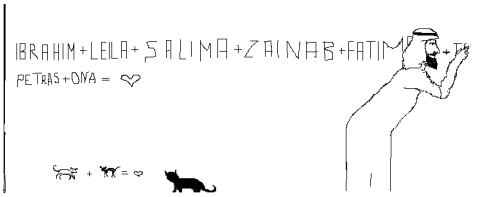


3.2.9 pav. Dviejų faktorių modelis

4. Modelio, kurio diagrama pavaizuota 3.2.9 paveiksle, $\chi^2_m = 4$, $n = 101$. Raskite suderinamumo indeksus NC, $RMSEA$, IFI, TLI ir CFI, jeigu $\chi^2_0 = 200$.

3. BENDRIEJI STRUKTŪRINIŲ LYGČIŲ MODELIAI

„Sieninė tautosaka“ sieka Antikos laikus. Ją tyrinėdami, archeologai susidarė gana išsamų gyvenimo Pompėjoje vaizdą. Kaip pasakytų statistikas, rašinėtojai, sąmoningai[24] visiškai to nesiekdami, sudarė pompėjiečių gyvenimo modelį. Suprastintą ir subjektyvų. Struktūrinių lygčių modeliuotojai ant sienų nerašinėja, bet iš esmės daro tą patį: kuria, jų nuomone, teisingą sudėtingo reiškinio modelį. Žinoma, suprastintą. Modelio aprašymui naudoja matematines lygtis. Beje, užrašų ant sienų forma kartais griežta matematinė, labai artima struktūrinėms lygtims, pavyzdžiui, „Jonas+Ona=KM[25]“, t. y. latentinis faktorius KM atsiskleidžia Jono ir Onos santykiuose.



Struktūrinių lygčių modeliavimas (SEM[26]) – tai modelių, aprašančių stebimų ir latentinių kintamųjų spėjamas priklausomybes, kūrimas ir analizė. SEM dar vadinamas kovariacijų struktūrine analize (CSA[27]), arba priežastiniu modeliavimu (CM[28]). Dažnai sakoma, kad SEM – tai kelių analizė su latentiniais kintamaisiais. Ankstesniuose skyreliuose nagrinėtieji kelių analizės (PA) ir patvirtinančiosios faktorinės analizės (CFA) modeliai yra atskiri SEM modelių atvejai. Todėl daug teiginių, suformuluotų PA ir CFA modeliams, tinka ir bendriesiems SEM modeliams. Vis dėlto, kad būtų lengviau skaityti, pagrindinius faktus atitinkamuose skyreliuose pakartosime.

Susidomėjimas struktūrinių lygčių metodais išaugo XX a. paskutiniame dešimtmetyje. Tada atsirado net keletas specialių kompiuterinių programų, o 1994 m pradėtas leisti vien SEM skirtas žurnalas[29]. Šiuo metu SEM yra viena iš sparčiausiai besivystančių statistikos sričių. Tokio audringo vystymosi pasekmė – nėra vieningos nuomonės, į kokius modelio rodiklius atsižvelgti, kokias tų rodiklių reikšmes laikyti pakankamomis, kiek imtyje turi būti duomenų ir pan. Todėl neretai SEM vadovėlių autoriai, per daug nesivargindami, išvardija statistikos autoritetus ir jų prieštaringus siūlymus, o kuo tikėti, palieka spręsti pačiam skaitytojui. Tokio vadovėlio skaitytojas pasijunta Buridano asilu[30]. Mes, kiek įmanoma, stengėmės nuo to skaitytoją apsaugoti.

SEM modelio kūrimo etapai: a) modelio aprašymas, b) modelio identifikavimas, c) parametrų vertinimas ir tinkamumo duomenims tyrimas, d) rezultatų interpretavimas ir

---

[24] Ant sienos rašinėti sąmoningas žmogus neina.

[25] Paauglių žargonu KM – karšta meilė.

[26] SEM – angl. *Structural Equation Modeling.*

[27] CSA – angl. *Covariance Structure Analysis.*

[28] CM – angl. *Causal Modeling.*

[29] *Structural Equation Modeling: A Multidisciplinary Journal* (anglų k.).

[30] Kartais ir ne Buridano.

modelio tobulinimas. Vis dėlto pirmiausia reikia patikrinti, ar duomenys tinka SEM modeliavimui.

3.1. SEM prielaidos

Skaitydamas respondento atsakymus, tyrėjas stebisi grafoje „Tėvo amžius (jei gyvas)“ radęs 125: „Čia turbūt klaida“. Respondentas nesutinka: „Jei būtų gyvas, jam tiek metų ir būtų.“

Įsivaizduokime vaikišką konstruktorių. Bet koks modelio konstravimas susideda iš dviejų pagrindinių etapų: a) pasirenkami objektai, iš kurių kuriamas modelis, b) jie sujungiami. Atsižvelgiant į tai, kas pasirinkta ir kaip sujungta, galima sukonstruoti ir grakštų Eifelio bokštą, ir gremėzdišką viešbutį „Novotel“. SEM modelių kūrimas visiškai analogiškas: visų pirma pasirenkami modelio konstrukciniai elementai, t. y. kintamieji, kurie bus stebimi (matuojami), ir kintamieji, kurie tiesiogiai stebimi nebus (latentiniai faktoriai). Po to jie susiejami spėjamais tiesioginiais arba netiesioginiais ryšiais. Tada tikrinama, ar pavyko sukurti gerą modelį.

SEM modelio duomenys – tai stebimųjų kintamųjų reikšmės. Latentiniai kintamieji tiesiogiai nestebimi. Tai tam tikri hipotetiniai konstruktai: gabumai, vaizduotė, tolerancija, verslumas ir pan. Sudarydami SEM modelį, manome, kad latentinių kintamųjų įtaką galima nustatyti pagal stebimųjų kintamųjų reikšmes. Stebimieji kintamieji, susiję su (paaiškinami) faktoriais, vadinami tų faktorių *indikatoriais*. Likusieji stebimieji kintamieji vadinami *tiesiogiai stebimais egzogeniniais kintamaisiais*.

SEM modelio kintamieji:

- latentiniai faktoriai,
- faktorių indikatoriai,
- tiesiogiai stebimi egzogeniniai kintamieji.

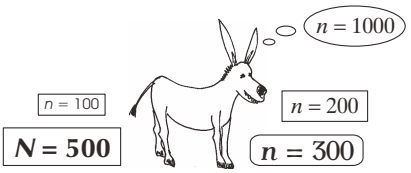
Prieš pradedant kurti SEM modelį, pirmiausia patikrinama, ar duomenys tenkina keliamus reikalavimus.

1 *Stebimieji kintamieji turi būti normaliai pasiskirstę.* Praktiniuose taikymuose šis reikalavimas dažniausiai sušvelninamas – reikalaujama, kad šie kintamieji būtų intervaliniai. Bendra taisyklė – kuo mažiau stebimieji kintamieji tenkina normalumo prielaidą, tuo didesnės imties reikia. Kartais, kai kintamojo reikšmių vidurkis visiškai aiškiai suvokiamas, į modelį įtraukiami ir ranginiai kintamieji. Pavyzdžiui, vidutinis pažymys – lengvai suvokiama ir interpretuojama sąvoka, vidutinis mokslinis laipsnis (ketvirtadaliu habilituotas daktaras) – ne. Jeigu modelyje naudojami

ranginiai kintamieji, jie turi įgyti pakankamai daug skirtingų reikšmių[31]. Klasikiniams SEM modeliams netinka kategoriniai kintamieji – lytis, tautybė, religija ir pan. Ką daryti, kai stebimieji kintamieji yra intervaliniai, bet nėra normaliai pasiskirstę, aptariama 3.9 skyrelyje.

2 *Duomenyse nėra išskirčių.* Jeigu yra labai išsiskiriančių stebėjimų, standartiniai SEM modelio parametrų vertinimo metodai (didžiausiojo tikėtinumo arba mažiausiųjų kvadratų) netinka. Jeigu duomenų yra labai daug (per 1000), galima parametrus vertinti ankstesniame skyriuje minėtu ADF metodu. Kitais atvejais SEM modelių analizė nerekomenduojama.

3 *Duomenų turi būti pakankamai daug.* Šiuo atveju Buridano asilo problemos išvengti nepavyks, nes svarstant, kiek turėtų būti stebėjimų, egzistuoja radikaliai priešingos nuomonės.



Vieningai sutariama, kad mažiau kaip 100 stebėjimų tinka tik naudojantis labai paprastais modeliais. Dažniausiai tyrimams naudojamos imtys, kurias sudaro 150–500 stebėjimų. Vis dėlto pasitaiko ir tokių SEM modelių, kuriuose stebėjimų yra keli tūkstančiai. Viena vertus, kuo daugiau duomenų, tuo tikslesni parametrų įverčiai. Kita vertus, kai imtis didelė, gali kilti problemų vertinant modelio tinkamumą duomenims ir interpretuojant rezultatus. Nagrinėjant dideles imtis, bet koks, net menkiausias parametrų skirtumas yra statistiškai reikšmingas. Norintiems griežtesnių imties didumo parinkimo taisyklių siūlome jos didumą sieti su vertinamųjų parametrų skaičiumi. Duomenų turi būti 10–20 kartų daugiau nei vertinamų parametrų.

4 *Visos kintamųjų priklausomybės yra tiesinės.* Pavyzdžiui, struktūrinė lygtis gali būti $Y = \lambda \cdot X + \beta Z + e$, bet ne $Y = \lambda X^3 + e$ arba $Y = \lambda XZ + e$. Jeigu kyla įtarimų, kad priklausomybė gali būti netiesinė, tai rekomenduojama stebimus kintamuosius transformuoti, keliant juos laipsniais, logaritmuojant ir pan.

5 *Nėra stebimųjų kintamųjų multikolinearumo.* Vadinasi, jų koreliacijos nėra beveik lygios vienetui. Kaip ir regresijoje, kintamųjų multikolinearumas gali iškreipti jų priklausomybes, pavyzdžiui, modelyje atsiras priešingo, nei turėtų būti, ženklo daugiklis. Dažniausiai duomenyse multikolinearumo pradedame ieškoti tik tada, kai susiduriame su modelio vertinimo ir interpretavimo problemomis. Jei modelyje dviejų kintamųjų koreliacija yra beveik lygi vienetui, galbūt vieną iš jų galima pašalinti.

[31] Pakankamai daug dažniausiai reiškia 5–7 reikšmes, nors, pavyzdžiui, skaičiavimams taikant LISREL programą, reikalaujama, kad ranginiai indikatoriai įgytų ne mažiau kaip 15 skirtingų reikšmių.

6 *Nedaug praleistų reikšmių.* Klasikiniai parametrų vertinimo metodai pritaikyti duomenims be praleistų reikšmių. Tuo tarpu realiuose tyrimuose ne taip dažnai pavyksta pasiekti, kad visi respondentai atsakytų į visus klausimus. Paprasčiausias problemos sprendimo būdas – iš duomenų aibės pašalinti eilutes, kuriose yra trūkstamų reikšmių. Toks metodas labai sumažina tiriamų duomenų skaičių. Todėl kartais praleistosios reikšmės keičiamos vidutine to kintamojo reikšme arba regresine tos reikšmės prognoze. Tokios operacijos galimos tik tada, kai praleistos reikšmės išsibarsčiusios visiškai atsitiktinai ir jų nėra labai daug. Praleistų reikšmių pakaitalų skaičiavimo procedūros įtrauktos į daugelį SEM skirtų statistinių programų. Tiesa, skirtingose programose taikomi skirtingi algoritmai.

Net turint visus reikalavimus tenkinančius duomenis dar neaišku, ar sudarytasis SEM modelis gerai tiks turimiems duomenims.

3.2. SEM modelio aprašymas

Kuriems galams tokia knyga <...>, kurioje nėra paveikslėlių nė pasikalbėjimų.
L. Kerolis. „Alisa stebuklų šalyje“

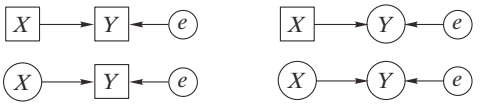
SEM modelio sudarymas – tai modelio užrašymas struktūrinėmis lygtimis arba modelio diagramos nubraižymas. Kokį modelį sudaryti – sprendžia pats tyrėjas. Būtent jis sugalvoja, kokius indikatorius įtraukti į modelį, kiek bus latentinių faktorių (latentinių kintamųjų) ir kaip juos derėtų pavadinti; su kokiais indikatoriais faktoriai bus susiję ir kokie priežastiniai ryšiai kintamuosius sies. Lengviausiai modelį nusakyti jo diagrama. Dar kartą primename naudojamus diagramose žymenis (žr. 3.3.1 lentelę). Iš jos matyti, kad latentiniai faktoriai ir stebimi kintamieji žymimi skirtingais simboliais.

3.3.1 lentelė. SEM modelio diagramos elementai

Žymėjimas	Reikšmė
▭	Stebimasis kintamasis
○	Latentinis kintamasis, liekamoji paklaida
→	Vienpusis priežastinis ryšys
↔	Koreliacija (kovariacija)

Latentiniai faktoriai vaizduojami apskritimais (elipsėmis).
Stebimieji kintamieji vaizduojami stačiakampiais.
Liekamosios paklaidos vaizduojamos apskritimais.

Kai manome, kad latentinis arba stebimasis kintamasis X daro įtaką kinatamajam Y, diagramoje brėžiame tiesią rodyklę nuo X į Y. Jeigu manome, kad Y priklauso ne vien nuo jam darančių įtaką kintamųjų, bet ir dar nuo kažko, braižome liekamąją paklaidą, kuri su Y susiejama tiesia rodykle (žr. 3.3.1 pav.). Pavyzdžiui, jei manome, kad vadybinin-

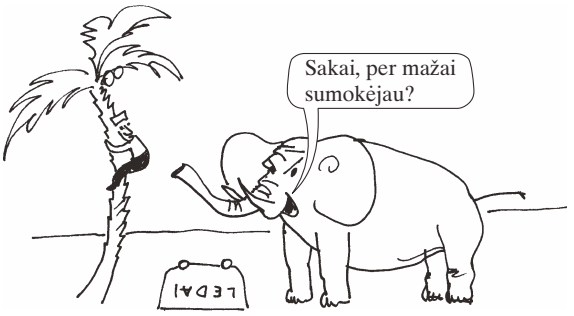


3.3.1 pav. Kintamųjų ryšių su liekamosiomis paklaidomis diagramos

ko organizaciniai gabumai Y priklauso *vien tik* nuo gebėjimo priimti sprendimus X, tai liekamosios paklaidos kintamajam Y nereikės. Jeigu manome, kad organizacinius gabumus gali lemti ir tam tikri kiti faktoriai, tai diagramoje juos pažymėsime kaip liekamąją paklaidą. Retai kada tikimasi, kad modelis idealiai aprašys kintamųjų priklausomybes. Todėl beveik visada diagramoje kintamasis, į kurį nubrėžta tiesi rodyklė, turi ir liekamąją paklaidą.

Diagramoje kintamasis, į kurį nubrėžta tiesi rodyklė iš kito kintamojo, beveik visada turi liekamąją paklaidą.

Įtraukdami į modelį latentinius faktorius ir liekamąsias paklaidas turime pasirinkti vieną iš dviejų alternatyvų: arba fiksuoti jų dispersijas ir laikyti jas lygiomis vienetui[32], arba vienetui prilyginti vieną su faktoriumi (liekamąja paklaida) susijusį kelio koeficientą. Tai vadinama skalės parinkimu. Norint gauti parametrų įverčius, skalę parinkti būtina. Pavyzdžiui, negalėtume pasakyti, kiek kainavo braškinių ledų porcija ir kiek porcijų buvo parduota, jeigu žinotume tik bendrąją už parduotus ledus gautą sumą. Jei, be bendrosios sumos, žinotume ir parduotų porcijų skaičių – porcijos kainą nustatytume be vargo. Ir atvirkščiai, jei, be bendrosios sumos, žinotume porcijos kainą, tai lengvai rastume pirkėjų skaičių.



Panašiai yra ir SEM modeliuose. Žinodami tik faktoriaus bendrąją įtaką visiems indikatoriams, negalime pasakyti, kaip stipriai kiekvieną indikatorių faktorius veikia. Kai kurie statistikai rekomenduoja fiksuoti faktorių ir liekamųjų paklaidų dispersijas. Kiti siūlo fiksuoti kelio koeficientus (faktorinius svorius). Taip darysime ir mes[33]. Patariama diagramoje vienetinio ilgio keliu sujungti su faktoriumi tą indikatorių, kurio ryšiu su faktoriumi mažiausiai abejojama. Jeigu tokie yra visi indikatoriai, tai nesvarbu, kurį kelio koeficientą prilyginti vienetui.

[32] Žinoma, galima pasirinkti ir ne vienetą, o kitą nenulinį, širdžiai mielesnį, skaičių.

[33] Išimtį sudarys keletas 3.4 skyrelyje nagrinėjamų SEM modelių pavyzdžių.

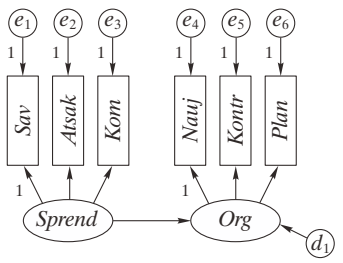
Kiekvieno faktoriaus rodyklė į vieną iš indikatorių turi fiksuotą kelio koeficientą. Tiesi rodyklė iš paklaidos į kintamąjį turi fiksuotą vienetinio ilgio kelio koeficientą.

Beveik visuomet su pasirinktuoju modeliu tikrinamas ir vadinamasis standartizuotasis modelis, kurio visų egzogeninių kintamųjų (įskaitant liekamąsias paklaidas) dispersijos prilyginamos vienetui. Tai suvienodina indikatorių skales. Būtent lygindami tarpusavyje standartizuotojo modelio kelių koeficientus, nusprendžiame, kurių kintamųjų ryšys stipresnis.

Koreliuojantys kintamieji diagramoje susiejami lenkta dvipuse rodykle. Jeigu kintamieji turi liekamąsias paklaidas, dvipusė rodyklė brėžiama tarp paklaidų.

Naudojantis minėtomis taisyklėmis, galima nubraižyti labai įvairias SEM modelių diagramas. Klasikinį SEM modelį geriausiai apibūdina terminas *latentinių faktorių PA modelis*. Tai CFA modelis, kurio latentiniai faktoriai sujungti priežastiniais ryšiais.

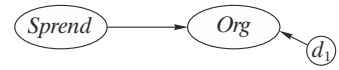
3.3.1 pavyzdys. Prisiminkime 3.2.2 pavyzdį apie vadybininką, kurio profesinei veiklai įtaką daro du latentiniai faktoriai: organizaciniai gabumai (*Org*), pasireiškiantys darbo planavimu (*Plan*), kontroliavimu (*Kontr*) ir naujovių diegimu (*Nauj*); gebėjimas priimti sprendimus (*Sprend*), pasireiškiantis vadybininko savarankiškumu (*Sav*), atsakomybės jausmu (*Atsak*) ir mokėjimu bendrauti su kitais vadybininkais (*Kom*). Tariame, kad gebėjimas priimti sprendimus daro įtaką organizaciniams gabumams. Spėjama SEM modelio schema pavaizduota 3.3.2 paveiksle.



3.3.2 pav. SEM modelio schema

Minėtą modelį galima išskaidyti į dvi sudėtines dalis: 1) struktūrinį modelį, siejantį latentinius faktorius, 2) matavimo modelius, kurie latentinius faktorius susieja su jų indikatoriais. Atkreipiame dėmesį, kad klasikinio SEM modelio visi stebimieji kintamieji yra latentinių faktorių indikatoriai, nėra egzogeninių stebimųjų kintamųjų.

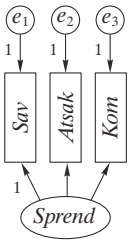
Struktūrinis modelis – tai kelių analizės modelio dalis, skirta tik latentiniams faktoriams.



Beje, dažnai literatūroje faktorių liekamosios paklaidos vadinamos trikdžiais. Šioje knygoje visas paklaidas vadiname liekamosiomis. Latentinių faktorių paklaidas žymime raide d, o matuojamųjų kintamųjų – e.

Matavimo modeliai – tai CFA modeliai, sudaromi kiekvienam latentiniam faktoriui. Pavyzdžiui, latentinio faktoriaus *Sprend* matavimo modelis yra susijęs su indikatoriais

Sav, *Atsak* ir *Kom*. Atkreipiame dėmesį, kad matavimo modelių diagramose rodyklės brėžiamos nuo faktoriaus į indikatorių.



Kaip ir nagrinėjant CFA modelius, matavimo modelių kintamųjų, susietų su latentiniu faktoriumi, skaičiui keliami tie patys reikalavimai.

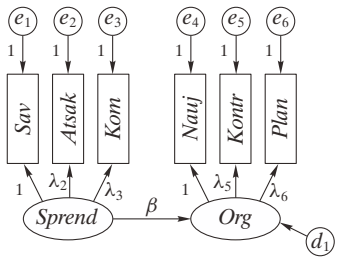
Keturi indikatoriai, tenkantys vienam latentiniam faktoriui, – rekomenduojama, trys – priimtina, du – iš bėdos, vienas – per mažai.

Primename, kad endogeniniais vadiname kintamuosius, į kuriuos diagramoje nukreipta bent viena tiesi rodyklė. Likusius vadiname egzogeniniais kintamaisiais. Klasikiniuose SEM modeliuose kalbama tik apie egzogeninius ir endogeninius faktorius, ne apie indikatorius. Nagrinėto 3.3.1 pavyzdžio modelyje faktorius *Sprend* yra egzogeninis, faktorius *Org* – endogeninis.

Diagramose pateiktą modelį galima užrašyti struktūrinėmis lygtimis. Nežinomus faktorinius svorius pažymėkime λ_j, o kelio tarp faktorių koeficientą – simboliu β (žr. 3.3.3 pav.).

Kelio koeficientas (faktorinis svoris) – tai daugiklis, iš kurio reikia padauginti kintamąjį, *nuo kurio* diagramoje brėžiama tiesi rodyklė. Jeigu kelio koeficientas fiksuotas ir lygus 1, tai dauginame iš 1. Lygtys rašomos kintamiesiems, *į kuriuos* diagramoje nukreiptos tiesios rodyklės. Paprastumo dėlei tarsime, kad visi kintamieji centruoti, t. y. modeliuose nėra laisvųjų narių (pridedamų konstantų). Pirmiausia užrašome struktūrinę modelio dalį:

$$Org = \beta Sprend + d_1.$$



3.3.3 pav. SEM modelis su pažymėtais nežinomais koeficientais

Tada pirmojo faktoriaus matavimo modelio lygtys yra tokios:

$$Sav = Sprend + e_1,$$
$$Atsak = \lambda_2 Sprend + e_2,$$
$$Kom = \lambda_3 Sprend + e_3.$$

Analogiškai užrašomos antrajo faktoriaus matavimo modelio lygtys:

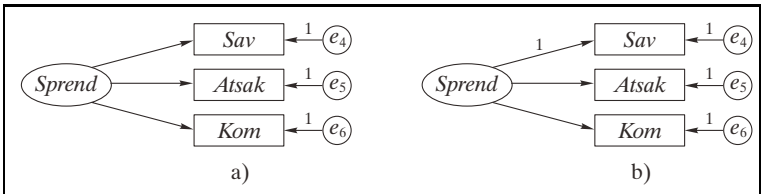
$$Nauj = Org + e_4,$$
$$Kontr = \lambda_5 Org + e_5,$$
$$Plan = \lambda_6 Org + e_6.$$

Visos liekamosios paklaidos yra normaliai pasiskirsčiusios, jų vidurkiai lygūs nuliui, o dispersijos nežinomos, $e_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2, \ldots, 6$, $d_1 \sim \mathcal{N}(0, \sigma_d^2)$. Visų latentinių faktorių vidurkiai lygūs nuliui[34]. Diagramoje koreliacijos ir kovariacijos nėra pažymėtos, todėl manome, kad visos liekamosios paklaidos nepriklausomos. Jeigu dalis kintamųjų koreliuotų, nurodytume jų koreliacijas. Beje, sudarant modelį, galima fiksuoti dalį parametrų, t. y. kelio koeficientų, dispersijų ir kovariacijų (koreliacijų).

Skyrelio pabaigoje nurodysime dažniausiai pasitaikančias sudarymo klaidas:

1 Priežastinis ryšys ne toks (ne tos krypties rodyklė), pavyzdžiui, *visų* su faktoriumi susijusių indikatorių rodyklės diagramoje nukreiptos į faktorių.

2 Užmirštama, kad faktorius turi turėti arba fiksuotą dispersiją, arba bent vieną fiksuotą faktorinį svorį. Dažniausiai faktorių dispersijų nefiksuosime, todėl vertėtų įdėmiai peržiūrėti 3.3.4 paveikslą.



3.3.4 pav. Latentinis faktorius be fiksuoto kelio koeficiento: a) tipinė modelio sudarymo klaida, b) jos ištaisymas

Sudarius modelį, reikia įvertinti nežinomus parametrus. Deja, tai ne visada įmanoma. Kokius reikalavimus turi tenkinti modelis, aptarsime kitame skyrelyje.

3.3. Modelio identifikavimas

Didžiausias iššūkis kiekvienam mąstytojui – taip suformuluoti problemą, kad ji turėtų sprendimą.

B. Raselas

[34] Sudėtingų SEM modelių latentiniai faktoriai gali ir neturėti nulinių vidurkių.

Modelis yra identifikuojamas, jei teoriškai galima įvertinti jo nežinomus parametrus. Tai jau minėjome nagrinėdami PA ir CFA modelius. Todėl, kalbėdami apie bendruosius SEM modelius, neišvengsime kai kurių pasikartojimų.

Norint modelį identifikuoti, reikia patikrinti, ar yra pakankamai informacijos parametrams įvertinti. Bendrojo SEM modelio vertinami parametrai yra šie:

1) nefiksuoti kelio koeficientai,
2) liekamųjų paklaidų dispersijos,
3) egzogeninių kintamųjų (stebimų ir latentinių) dispersijos,
4) stebimųjų kintamųjų vidurkiai, kurie vertinami tik nagrinėjant necentruotus modelius, t. y. tada, kai tie parametrai yra struktūrinėse lygtyse.

Norime atkreipti dėmesį, kad endogeninių kintamųjų dispersijos nėra nežinomi parametrai – jas išreiškiame per kitus parametrus.

Aptarsime, kokia informacija naudojama parametrų vertinimui. Tarkime, kad modelyje yra k stebimųjų kintamųjų. Tada nežinomiems parametrams įvertinti galima panaudoti $k(k+1)/2$ empirinių kovariacijų (kintamojo dispersija – tai jo kovariacija su savimi) ir k empirinių vidurkių.

Centruotų modelių parametrams įvertinti naudojamos tik kovariacijos.
Necentruotų modelių parametrams įvertinti naudojamos kovariacijos ir vidurkiai.

Visos vertinimui naudojamos kovariacijos ir vidurkiai vadinami duomenų taškais. Centruotuose modeliuose yra $k(k+1)/2$ duomenų taškų, o necentruotuose – $k(k+1)/2+k$ duomenų taškų. Apibrėžkime modelio laisvės laipsnių sąvoką.

SEM modelio laisvės laipsniai (df) yra duomenų taškų ir vertinamų parametrų skaičiaus skirtumas.

Pagal laisvės laipsnių skaičių modeliai skirstomi į nepakankamai identifikuojamus, tiksliai identifikuojamus ir perteklinės identifikacijos.

1 *Nepakankamai identifikuojamas* (*neidentifikuojamas*) *modelis* yra toks, kurio $df < 0$, t. y. vertinamųjų parametrų yra daugiau nei informacijos jiems įvertinti. Neįvedus papildomų apribojimų, tokio modelio parametrų vienareikšmiškai įvertinti negalima.

2 *Tikslai identifikuojamas modelis* yra toks, kurio $df = 0$. Šiuo atveju teoriškai informacijos pakanka parametrų įverčiams apskaičiuoti, tačiau nėra prasmės tirti modelio tinkamumo duomenims.

3 *Perteklinės identifikacijos modelis* yra toks, kurio $df > 0$. Šiuo atveju galima įvertinti ne tik nežinomus parametrus, bet ir modelio tinkamumą duomenimis.

Tyrėjams įdomiausi yra perteklinės identifikacijos modeliai. Tik juos toliau ir nagrinėsime.

Centruotų ir necentruotų modelių vertinamųjų parametrų skaičius yra skirtingas. Vis dėlto, jeigu visi latentinių faktorių vidurkiai yra lygūs nuliui (tai standartinis reikalavimas), tai tą pačią SEM modelio diagramą atitinkančių centruoto ir necentruoto modelių laisvės laipsnių skaičius yra tas pats. Todėl nepasikeičia ir modelio identifikuojamumas.

Nepasikeis ir kelių koeficientų bei kovariacijų įverčiai. Taip yra todėl, kad laisvųjų narių yra tiek, kiek ir stebimųjų kintamųjų, ir jų įverčiai lygūs tų kintamųjų empiriniams vidurkiams. Visiems kitiems parametrų įverčiams abiejuose modeliuose naudojamos stebimųjų kintamųjų kovariacijos. Tarkime, kad centruotame modelyje reikia įvertinti s parametrų. Tada $df = k(k+1)/2 - s$. Necentruotame modelyje reikia įvertinti $k+s$ parametrų, bet ir duomenų taškų yra daugiau, todėl $df = k(k+1)/2 + k - (s+k) = k(k+1)/2 - s$.

SEM modelis yra identifikuojamas, jei vertinamųjų kelio koeficientų, liekamųjų paklaidų dispersijų ir egzogeninių kintamųjų dispersijų skaičius neviršija $k(k+1)/2$; čia k – modelio stebimųjų kintamųjų skaičius.

3.3.2 pavyzdys. Užrašykime 3.3.3 diagramos centruotą ir necentruotą modelius:

$$\begin{cases} Org = \beta Sprend + d_1, \\ Sav = Sprend + e_1, \\ Atsak = \lambda_2 Sprend + e_2, \\ Kom = \lambda_3 Sprend + e_3, \\ Nauj = Org + e_4, \\ Kontr = \lambda_5 Org + e_5, \\ Plan = \lambda_6 Org + e_6, \end{cases} \qquad \begin{cases} Org = \mu_0 + \beta Sprend + d_1, \\ Sav = \mu_1 + Sprend + e_1, \\ Atsak = \mu_2 + \lambda_2 Sprend + e_2, \\ Kom = \mu_3 + \lambda_3 Sprend + e_3, \\ Nauj = \mu_4 + Org + e_4, \\ Kontr = \mu_5 + \lambda_5 Org + e_5, \\ Plan = \mu_6 + \lambda_6 Org + e_6. \end{cases}$$

Modelio latentiniai faktoriai yra *Sprend* ir *Org*, o indikatoriai – *Sav*, *Atsak*, *Kom*, *Nauj*, *Kontr*, *Plan*. Kadangi modelyje yra 6 indikatoriai (egzogeninių stebimų kintamųjų nėra), tai turime $6(6+1)/2 = 21$ kovariaciją ir 6 empirinius vidurkius. Atrodytų, kad yra 7 necentruotoe modelio vertintini vidurkiai ($\mu_0, \ldots, \mu_6$). Vis dėlto nesunku įsitikinti, kad $\mu_0 = 0$. Iš tikrųjų visų liekamųjų paklaidų vidurkiai lygūs nuliui. Standartiškai nulinius vidurkius turi ir visi latentiniai faktoriai. Todėl, viena vertus, $\mathbf{E}Org = 0$. Kita vertus,

$$\mathbf{E}Org = \mu_0 + \beta \mathbf{E}Sprend + \mathbf{E}d_1 = \mu_0 + 0 + 0 = \mu_0.$$

Taigi $\mu_0 = 0$. Analogiškai galima įsitikinti, kad $\mu_1 = \mathbf{E}Sav$ ir t. t.

3.3.3 pavyzdys. Reikia įvertinti ir centruoto, ir necentruoto 3.3.2 pavyzdžio modelių (žr. taip pat 3.3.3 diagramą) kelio koeficientus $\lambda_2, \lambda_3, \lambda_5, \lambda_6, \beta$, liekamųjų paklaidų dispersijas $\sigma_1^2, \ldots, \sigma_7^2$ ir latentinio faktoriaus *Sprend* dispersiją. Iš viso – 13 parametrų. Tam galima pasinaudoti 21 empirine indikatorių kovariacija. Dar reikia įvertinti necentruoto modelio vidurkius $\mu_0, \ldots, \mu_7$. Tam galima pasinaudoti 7 empiriniais indikatorių vidurkiais. Taigi centruotame modelyje yra 13 vertinamų parametrų ir 21 duomenų taškas, $df = 21 - 13 = 8$. Necentruotame modelyje yra $13+6 = 19$ vertinamų parametrų ir $21+6 = 27$ duomenų taškai, $df = 27-19 = 8$ (žr. 3.3.2 lentelę).

3.3.2 lentelė. Centruoto ir necentruoto 3.3.3 diagramos modelių identifikavimas

		Centruotas modelis	Necentruotas modelis
Vertinami parametrai	Kelių koeficientai	5	5
	Liekamųjų paklaidų dispersijos	7	7
	Faktorių dispersijos	1	1
	Vidurkiai	–	6
Duomenų taškai	Empirinės kovariacijos	21	21
	Empiriniai indikatorių vidurkiai	–	6
df		8	8

Šiame skyrelyje kalbėjome apie teorinį identifikuojamumą. Jį reikėtų skirti nuo praktinio identifikuojamumo. Kaip visi puikiai žinome, ugnį galima įkurti energingai trinant du pagaliukus. Teoriškai.

Taip atsirado vegetarai

Parktiškai net ir perteklinės identifikacijos modelis nebūtinai yra išsprendžiamas. Situacija primena lygčių sistemos sprendimą. Teorinis identifikuojamumo analogas būtų konstatavimas, kad norint vienareikšmiškai rasti tris nežinomuosius reikia trijų lygčių. Vis dėlto trys lygtys dar neužtikrina sprendinio egzistavimo (pvz., jei jos visos yra ekvivalenčios).

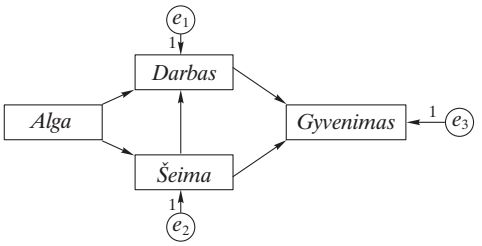
Praktiškai modelis gali būti neidentifikuojams ne tik tada, kai modelyje yra per daug vertinamųjų parametrų. Labai dažnai pamirštama, kad kiekvienas latentinis faktorius turi turėti arba fiksuotą dispersiją, arba bent vieną vienetinį su juo susijusį faktorinį svorį (žr. 3.3.4 pav.). Be to, taip gali atsitikti ir dėl per didelės indikatorių koreliacijos. Tuo atveju reikia bandyti tobulinti modelį. Beje, problemą gali pavykti išspręsti atsisakius keleto kintamųjų ir, priešingai, papildžius modelį naujais kintamųjų ryšiais.

3.4. SEM modelių pavyzdžiai

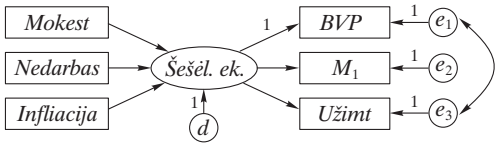
Jei negali padaryti gerai, padaryk, kad bent gerai atrodytų.

Ankstesniuose skyreliuose aptarėme klasikinį SEM modelį, kurio priežastiniai ryšiai buvo spėjami tik tarp latentinių kintamųjų. Šiuo metu egzistuoja ir daug kitokių SEM modelių. Plačiau jų neanalizuosime (analizės metodai kur kas sudėtingesni nei nagrinėjami šioje knygoje), susipažinsime su kai kuriais iš jų.

SEM modelis be latentinių kintamųjų. Minėjome, kad kelių analizės modeliai – tai atskiras SEM modelių atvejis. Tiksliai identifikuojamus PA modelius aptarėme pirmajame skyriuje. Čia pateiksime perteklinės identifikacijos kelių analizės modelio pavyzdį. Prisiminkime 3.1.1 pavyzdį. Tiriame respondento pasitenkinimą gyvenimu (kintamasis *Gyvenimas*), atsižvelgdami į santykius darbe (*Darbas*), santykius šeimoje (*Šeima*) ir respondento algą (*Alga*). Visi kintamieji yra intervaliniai. Pateikiamas 3.3.5 paveiksle SEM modelis yra 3.1.2 diagramos modifikacija. Nesunku įsitikinti, kad modelis yra perteklinės identifikacijos. Iš tikrųjų, reikia įvertinti 5 kelių koeficientus ir 3 liekamųjų paklaidų dispersijas, o turime 10 kovariacijų. Todėl $df = 10 - 8 = 2 > 0$.



3.3.5 pav. SEM modelis be latentinių kintamųjų



3.3.6 pav. MIMIC modelis

Daugiapriežastinis SEM modelis (*MIMIC*[35]). Šiame modelyje yra latentinių faktorių, kuriems daro įtaką egzogeniniai stebimieji kintamieji. MIMIC modeliai remiasi prielaida, kad ne visada latentinių faktorių apraiškas tik stebime – yra reiškinių, kurie latentinius faktorius tiesiogiai veikia. Pavyzdžiui, socialinis aktyvumas (latentinis faktorius) pasireiškia dalyvavimu bendruomenės gyvenime, bendravime su draugais ir pan. Tuo tarpu išsilavinimas nėra socialinio aktyvumo apraiška, bet jį neabejotinai veikia. MIMIC modelio diagramoje yra rodyklių, nukreiptų nuo stebimų kintamųjų į faktorius. Tipinio MIMIC modelio šešėlinei ekonomikai prognozuoti diagrama pateikta 3.3.6 paveiksle. Iš jos matyti, kas gali turėti įtakos šešėlinei ekonomikai ir kokie indikatoriai čia dalyvauja. Šešėlinė ekonomika (*Šešėl. ek.*) priklauso nuo mokesčių naštos (*Mokest*), nedarbo lygio (*Nedarbas*), infliacijos (*Infliacija*). Šešėlinė ekonomika daro įtaką BVP, tenkančiam vienam gyventojui (*BVP*), agreguotam pinigų rodikliui (M_1) ir užimtumui (*Užimt*). Žinoma, MIMIC modeliuose gali būti ir ne vienas latentinis faktorius.

Daugelio požymių – daugelio matavimo metodų modelis (*MTMM*[36]). Tai faktorinis modelis, kuriame skirtingi latentiniai faktoriai turi bendrų indikatorių. Tradiciškai toks modelis atitinka situaciją, kai tam tikros savybės matuojamos keliais skirtingais metodais. Pavyzdžiui, pagal intervalinę skalę balais kiekvieno mokinio patrauklumą, gabumus mokslui ir lyderio savybes įvertina pats mokinys, jo tėvai ir klasės draugai (imamas jų nuomonių vidurkis). Šiuos mokinio savybių indikatorius pažymėkime atitinkamai p_m, g_m, v_m, p_t, g_t, v_t ir p_k, g_k, v_k. Yra trys vertinimo būdai (mokinys, tėvai, klasė) ir trys vertinamos savybės (patrauklumas, gabumai, lyderystė). Galima įtarti, kad vertinimo būdas turi įtakos įvertinimo didumui. Todėl MTMM modelyje vertinimo būdai įtraukiami kaip latentiniai faktoriai (pažymėkime juos *Mok*, *Tėv*, *Klas*). Kita vertus, vertinamąsias

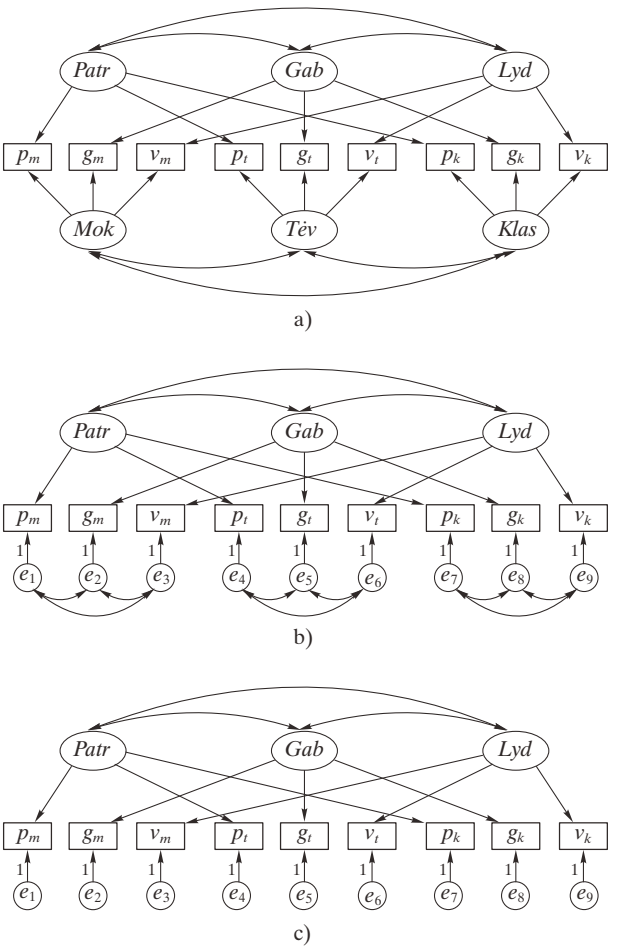
[35] MIMIC – angl. *Multiple Indicators and Multiple Causes.*

[36] MTMM – angl. *Multitrait-Multimethod Model.*

savybes taip pat galima traktuoti kaip latentinius faktorius (*Patr*, *Gab*, *Lyd*), kurie tik iš dalies atsispindi apklaustųjų nuomonėse. MTMM modelio diagrama pavaizduota 3.3.7 paveiksle, a). Atkreipiame dėmesį, kad indikatoriai neturi liekamųjų paklaidų.

MTMM modelių visų latentinių faktorių dispersijos laikomos lygiomis vienetui. Tirti MTMM modelius sunku, retai kada pavyksta gauti nežinomų parametrų įverčius.

Koreliuotųjų požymių – koreliuotųjų liekamųjų paklaidų (*CTCU*[37]) ***ir koreliuotųjų požymių*** (*CT*[38] ***modeliai***. Šie modeliai buvo sukurti kaip alternatyva MTMM modeliams, siekiant palengvinti parametrų įverčių skaičiavimą. Literatūroje CTCU modelis dažnai laikomas MTMM modelio atskiru atveju. CTCU modelyje MTMM modelyje vertinimo būdų latentiniai faktoriai pakeičiami su kiekvienam matavimo metodui koreliuojančiomis



3.3.7 pav. Mokinio savybių modeliai: a) MTMM, b) CTCU, c) CT

[37] CTCU – angl. *Correlated Trait – Correlated Uniqueness.*

[38] CT – angl. *Correlated Trait.*

liekamosiosmis paklaidomis (žr. 3.3.7 pav., b). Manoma, kad liekamųjų paklaidų koreliacija netiesiogiai įrodo vertinimo būdo įtaką. Todėl dažnai dar tiriamas ir CT modelis, t. y. CTCU modelis be koreliuojančių liekamųjų paklaidų (žr. 3.3.7 pav., c). Jeigu CT modelis blogai tinka duomenims, o CTCU modelis – pakankamai gerai, tai galima teigti, kad vertinimo būdas turėjo įtakos matavimo rezultatams. Kita vertus, liekamoji paklaida atspindi viską, kas gali daryti įtaką indikatoriams, todėl liekamosios paklaidos didumą galėjo veikti ne vien vertinimo būdas. Ir CTCU, ir CT modelių latentinių faktorių dispersijos prilyginamos vienetui.

Antrosios eilės faktorinis modelis. Tai SEM modelis, kurio vieni latentiniai faktoriai paaiškinami kitu, bendresniu, latentiniu faktoriumi.

Sudarant antrosios eilės SEM modelius, reikalaujama, kad:

1) antrosios eilės faktorius turėtų vienetinę dispersiją,
2) antrosios eilės faktorius būtų susijęs su bent trimis latentiniais pirmosios eilės faktoriais.

Pateiksime pavyzdį iš [4] straipsnio. Suaugę respondentai pildė Vekslerio[39] testą WAIS-III. Tariama, kad yra keturi (pirmosios eilės) latentiniai faktoriai, susiję su keturiolika indikatorių (testu matuojamų kintamųjų):

1 Latentinis *Verbalinio supratingumo indeksas* (*VSI*), susijęs su indikatoriais *Žodynas* (*Ž*), *Panašumai* (*P*), *Informacija* (*I*), *Supratingumas* (*S*).

2 Latentinis *Darbinės atminties indeksas* (*DAI*), susijęs su indikatoriais *Aritmetika* (*A*), *Skaičių eilė* (*SE*), *Raidžių-skaičių eilė* (*RSE*).

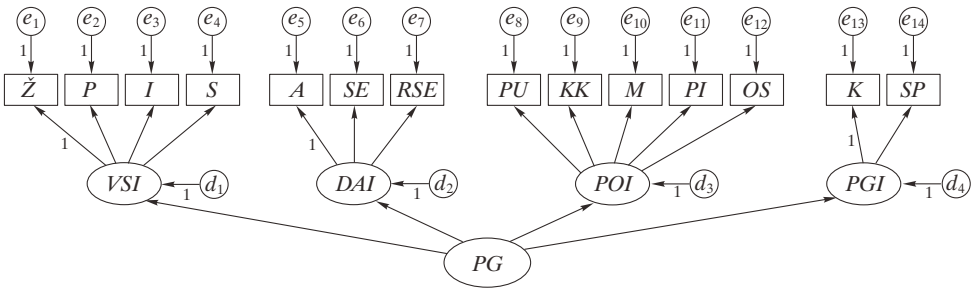
3 Latentinis *Percepcinės organizacijos indeksas* (*POI*), susijęs su indikatoriais *Paveikslėlių užbaigimas* (*PU*), *Kubelių kompozicija* (*KK*), *Matricos* (*M*), *Paveikslėlių išdėstymas* (*PI*), *Objektų sudėjimas* (*OS*).

4 Latentinis *Psichomotorinio greičio indeksas* (*PGI*), susijęs su indikatoriais *Skaičių simboliai – Kodavimas* (*K*), *Simbolių paieška* (*SP*).

Be to manoma, kad yra antrosios eilės latentinis faktorius – *Pažintiniai gebėjimai* (*PG*) – susijęs su visais latentiniais faktoriais *VSI*, *DAI*, *PO*, *PGI*. SEM modelio diagrama nubraižyta 3.3.8 paveiksle. Primename, kad visų latentinių faktorių pavadinimus sugalvoja tyrėjas. Pateiktajame pavyzdyje pirmojo lygmens faktorių pavadinimai tradiciniai, nusistovėję psichologinėje literatūroje.

Žinoma, gali būti ir daug kitokių modelių. Pavyzdžiui, SEM modeliai su tinkamai parinktais koeficientais atitinka hierarchinius augimo modelius. Be to, galima tikrinti, ar skirtingoms respondentų grupėms tinka tokie pat SEM modeliai. Šiuos modelius nagrinėsime kituose skyreliuose.

[39] *David Wechsler* (1896–1981) – JAV psichologas.



3.3.8 pav. Antrosios eilės faktorinis modelis

3.5. Matricinis SEM modelio užrašas

Matematikos suprasti negalima, bet galima prie jos priprasti.
Dž. Neimanas

Klasikinių SEM modelių matematiniai žymėjimai yra nusistovėję. Beveik jų nekeisdami, užrašysime SEM modelį. Primename, kad klasikinis SEM modelis yra toks, kuriame priežastiniai ryšiai yra tik tarp latentinių kintamųjų (jie sudaro struktūrinį modelį), o visi stebimieji kintamieji tėra latentinių faktorių indikatoriai. Visur šiame skyrelyje kalbama tik apie centruotus modelius, t. y. tariama, kad visų kintamųjų (ir latentinių, ir stebimųjų) ir jų liekamųjų paklaidų vidurkiai lygūs nuliui.

3.3.3 lentelė. Matricinio SEM modelio žymenys

Žymėjimas	Matricos eilė	Reikšmė
s		Latentinių egzogeninių faktorių skaičius
m		Latentinių endogeninių faktorių skaičius
k		Indikatorių, susietų su egzogeniniais faktoriais, skaičius
p		Indikatorių, susietų su endogeniniais faktoriais, skaičius
$\mathbf{X}$	$k \times 1$	Indikatorių, susietų su egzogeniniais faktoriais, vektorius
$\mathbf{Y}$	$p \times 1$	Indikatorių, susietų su endogeniniais faktoriais, vektorius
$\boldsymbol{\xi}$	$s \times 1$	Latentinių egzogeninių faktorių vektorius
$\boldsymbol{\eta}$	$m \times 1$	Latentinių endogeninių faktorių vektorius
$\boldsymbol{\Lambda}_x$	$k \times s$	Kelio tarp $\boldsymbol{\xi}$ ir $\mathbf{X}$ koeficientai
$\boldsymbol{\Lambda}_y$	$p \times m$	Kelio tarp $\boldsymbol{\eta}$ ir $\mathbf{Y}$ koeficientai
$\mathbf{B}$	$m \times m$	Kelio koeficientai, siejantys tarpusavyje $\boldsymbol{\eta}$ faktorius
$\boldsymbol{\Gamma}$	$m \times s$	Kelio tarp $\boldsymbol{\xi}$ ir $\boldsymbol{\eta}$ koeficientai
$\boldsymbol{\delta}$	$k \times 1$	Indikatorių $\mathbf{X}$ liekamųjų paklaidų vektorius
$\mathbf{e}$	$p \times 1$	Indikatorių $\mathbf{Y}$ liekamųjų paklaidų vektorius
$\mathbf{d}$	$m \times 1$	Faktorių $\boldsymbol{\eta}$ liekamųjų paklaidų vektorius
$\boldsymbol{\Phi}$	$s \times s$	Egzogeninių faktorių $\boldsymbol{\xi}$ kovariacijų matrica
$\boldsymbol{\Psi}$	$m \times m$	Liekamųjų paklaidų $\mathbf{d}$ kovariacijų matrica
$\boldsymbol{\Theta}_e$	$p \times p$	Liekamųjų paklaidų $\mathbf{e}$ kovariacijų matrica
$\boldsymbol{\Theta}_\delta$	$k \times k$	Liekamųjų paklaidų $\boldsymbol{\delta}$ kovariacijų matrica
$\boldsymbol{\Sigma}$	$(k+p) \times (k+p)$	Indikatorių kovariacijų matrica
$\boldsymbol{\Sigma}_M$	$(k+p) \times (k+p)$	Teorinė modelio indikatorių kovariacijų matrica
$\mathbf{S}$	$(k+p) \times (k+p)$	Empirinė indikatorių kovariacijų matrica

Tradiciškai atskirai aprašomi egzogeniniai ir endogeniniai latentiniai faktoriai. Priminsime, kad modelio diagramoje į endogeninius latentinius faktorius yra nukreiptos rodyklės iš kitų faktorių. Tarkime, kad modelyje yra s egzogeninių ir m endogeninių latentinių faktorių. Be to, yra k indikatorių, susijusių su egzogeniniais faktoriais, ir p indikatorių, susijusių su endogeniniais faktoriais. Modelyje dar yra liekamosios paklaidos. Pagrindiniai simboliai nurodyti 3.3.3 lentelėje.

Naudodamiesi matriciniais žymenimis, pagrindines priklausomybes galime užrašyti trimis lygtimis:

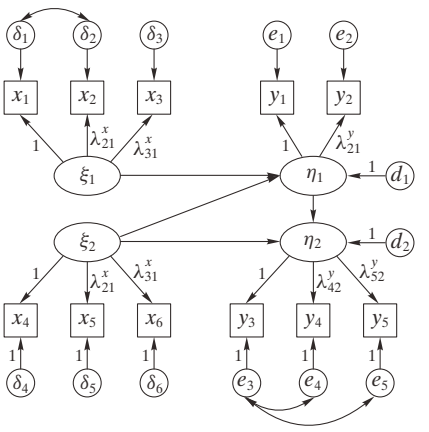
$$\begin{cases} \boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \mathbf{d}, \\ \mathbf{Y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \mathbf{e}, \\ \mathbf{X} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}. \end{cases} \tag{3.3.2}$$

Pirmoji lygtis aprašo struktūrinį modelį, o likusios lygtys – matavimo modelius. Matricos **B** pagrindinėje įstrižainėje visi elementai visada lygūs nuliui. Yra ir daugiau žinomų matricų elementų, pavyzdžiui, kelių daugikliai prie liekamųjų paklaidų lygūs vienetui. Kurie kiti parametrai žinomi, priklauso nuo konkretaus modelio.

3.3.4 pavyzdys. Užrašysime SEM modelio, pavaizduoto 3.3.9 paveiksle, matricines lygtis. Matome, kad yra du endogeniniai latentiniai faktoriai η_1 ir η_2, kurie priklauso nuo egzogeninių faktorių ξ_1 ir ξ_2. Nesunkiai nustatome, kad $s = 2$, $m = 2$, $p = 5$, $k = 6$, $d = 2$ ir

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \qquad \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \qquad \mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}, \qquad \boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}.$$



3.3.9 pav. Klasikinio SEM modelio diagrama

Iš pradžių lygtimis užrašysime struktūrinį modelį:

$$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + d_1,$$
$$\eta_2 = b_{21}\eta_1 + \gamma_{22}\xi_2 + d_2.$$

Dabar šias lygtis galime perrašyti nausodamiesi matricomis. Taigi (3.3.2) sistemos pirmojoje lygtyje

$$\mathbf{B} = \begin{pmatrix} 0 & 0 \\ b_{21} & 0 \end{pmatrix}, \qquad \mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{pmatrix}.$$

Analogiškai nustatome, kad

$$\mathbf{\Lambda}_y = \begin{pmatrix} 1 & 0 \\ \lambda_{21}^y & 0 \\ 0 & 1 \\ 0 & \lambda_{42}^y \\ 0 & \lambda_{52}^y \end{pmatrix}, \qquad \mathbf{\Lambda}_x = \begin{pmatrix} 1 & 0 \\ \lambda_{21}^x & 0 \\ \lambda_{31}^x & 0 \\ 0 & 1 \\ 0 & \lambda_{52}^x \\ 0 & \lambda_{62}^x \end{pmatrix}.$$

Belieka užrašyti modelio kintamųjų kovariacijas ir dispersijas:

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{D}\xi_1 & 0 \\ 0 & \mathbf{D}\xi_2 \end{pmatrix}, \qquad \mathbf{\Psi} = \begin{pmatrix} \mathbf{D}d_1 & 0 \\ 0 & \mathbf{D}d_2 \end{pmatrix},$$

$$\mathbf{\Theta}_y = \begin{pmatrix} \mathbf{D}e_1 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{D}e_2 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{D}e_3 & \mathbb{C}\mathrm{ov}(e_3, e_4) & \mathbb{C}\mathrm{ov}(e_3, e_5) \\ 0 & 0 & \mathbb{C}\mathrm{ov}(e_4, e_3) & \mathbf{D}e_4 & 0 \\ 0 & 0 & \mathbb{C}\mathrm{ov}(e_5, e_3) & 0 & \mathbf{D}e_5 \end{pmatrix},$$

$$\mathbf{\Theta}_x = \begin{pmatrix} \mathbf{D}\delta_1 & \mathbb{C}\mathrm{ov}(\delta_1, \delta_2) & 0 & 0 & 0 & 0 \\ \mathbb{C}\mathrm{ov}(\delta_2, \delta_1) & \mathbf{D}\delta_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{D}\delta_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{D}\delta_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{D}\delta_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{D}\delta_6 \end{pmatrix}.$$

Primename, kad kovariacija yra simetrinė funkcija, t. y. $\mathbb{C}\mathrm{ov}(\delta_1, \delta_2) = \mathbb{C}\mathrm{ov}(\delta_2, \delta_1)$.

Idealiu atveju, kai modelis teisingai aprašo kintamųjų priklausomybes, $\mathbf{\Sigma} = \mathbf{\Sigma}_M$. Pačių kovariacijų matricų nežinome, todėl tyrimui naudojame jų įverčius. Tradiciškai $\mathbf{\Sigma}$ įvertis žymimas $\mathbf{S}$. Yra žinoma, kad teorinė modelio indikatorių kovariacijų matrica yra tokios struktūros:

$$\mathbf{\Sigma}_M = \begin{pmatrix} \mathbf{\Sigma}_{yy} & \mathbf{\Sigma}_{yx} \\ \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_{xx} \end{pmatrix};$$

čia

$$\mathbf{\Sigma}_{yy} = \mathbf{\Lambda}_y(I - \mathbf{B})^{-1}(\mathbf{\Gamma\Phi\Gamma}' + \mathbf{\Psi})(I - \mathbf{B}')^{-1}\mathbf{\Lambda}_y' + \mathbf{\Theta}_e,$$
$$\mathbf{\Sigma}_{xx} = \mathbf{\Lambda}_x\mathbf{\Phi\Lambda}_x' + \mathbf{\Theta}_\delta,$$
$$\mathbf{\Sigma}_{xy} = \mathbf{\Lambda}_x\mathbf{\Phi\Gamma}'(I - \mathbf{B}')^{-1}\mathbf{\Lambda}_y'.$$

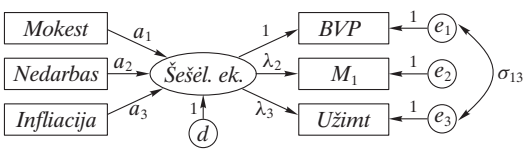
Suradę nežinomų parametrų įverčius ir įstatę juos į šias lygtis, gauname modelio indikatorių matricos įvertį $\widehat{\mathbf{\Sigma}}_M$. Kitas žingsnis būtų palyginti $\mathbf{S}$ su $\widehat{\mathbf{\Sigma}}_M$. Mažas matricų skirtumas patvirtina gerą modelio tikimą duomenims. Plačiau apie tai – kitame skyrelyje.

Pateiktas matricinis modelis tiesiogiai susijęs su istoriškai pirmąja SEM modeliavimui skirta programa LISREL. Ankstesnėse jos versijose visas matricas turėjo aprašyti vartotojai. Svarbiausias (3.3.2) užrašo trūkumas yra tai, kad jis tinka tik klasikiniams SEM modeliams. Tuo tarpu MIMIC modelio taip užrašyti nepavyks. Todėl šiuo metu vartojamos ir kitos matricinės SEM modelio formuluotės. Viena jų, vadinamoji RAM[40], formuluotė remiasi tokia logika – SEM modelį galima užrašyti regresinėmis lygtimis. Visi kintamieji, išskyrus liekamąsias paklaidas, pažymimi $\mathbf{v}$, liekamosios paklaidos ir egzogeniniai stebimieji kintamieji pažymimi $\mathbf{u}$. Tada modelį galima užrašyti taip:

$$\mathbf{v} = \mathbf{A}\mathbf{v} + \mathbf{u}.$$

Žinoma, matricoje $\mathbf{A}$ bus daug elementų, lygių nuliui. Kovariacijų matrica gaunama tiesiog apskaičiavus $\mathbf{E}(\mathbf{u}\mathbf{u}')$. RAM užrašas dažnai naudojamas parametrus vetinant R programa.

3.3.5 pavyzdys. Užrašysime matricinę 3.3.6 paveiksle pavaizduoto MIMIC modelio RAM išraišką. Nežinomus kelių koeficientus 3.3.10 diagramoje pažymėjome raidėmis a_1, a_2, a_3, λ_1 ir λ_2. Tarkime, kad σ_{13} yra nežinoma e_1 ir e_3 kovariacija, o σ_1^2, σ_2^2, σ_3^2 ir σ_d^2 žymi liekamųjų paklaidų e_1, e_2, e_3 ir d dispersijas.



3.3.10 pav. MIMIC modelio RAM formuluotės diagrama

Matricinis modelio užrašas atrodo taip:

$$\begin{pmatrix} \textit{Mokest} \\ \textit{Nedarbas} \\ \textit{Infliacija} \\ \textit{BVP} \\ M_1 \\ \textit{Užimt} \\ \textit{Šešėl. ek.} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_3 \\ a_1 & a_2 & a_3 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \textit{Mokest} \\ \textit{Nedarbas} \\ \textit{Infliacija} \\ \textit{BVP} \\ M_1 \\ \textit{Užimt} \\ \textit{Šešėl. ek.} \end{pmatrix} + \begin{pmatrix} \textit{Mokest} \\ \textit{Nedarbas} \\ \textit{Infliacija} \\ e_1 \\ e_2 \\ e_3 \\ d \end{pmatrix}.$$

Modelio kintamųjų kovariacijų matrica yra

$$\begin{pmatrix} \mathbf{D}\textit{Mokest} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{D}\textit{Nedarbas} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{D}\textit{Infliacija} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_1^2 & 0 & \sigma_{13} & 0 \\ 0 & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{13} & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_d^2 \end{pmatrix}.$$

[40] RAM – angl. *Reticular Action Model*.

3.6. SEM modelio parametrų vertinimas ir tinkamumas duomenims

Jeigu imtis iš tikrųjų didelė, tai statistiškai reikšmingas rezultatas iš esmės tepatvirtina, kad imtis didelė.

B. Tompsonas

Šiame skyrelyje nagrinėsime tik perteklinės identifikacijos SEM modelius. Be to, tarsime, kad galioja pagrindinės SEM modelių prielaidos, pavyzdžiui, kad visi kintameji normaliai pasiskirstę.

SEM modelio suderinamumo su duomenimis tyrimas beveik nesiskiria nuo CFA modelio suderinamumo tyrimo (žr. 2.6 skyrelį). Todėl tik trumpai paminėsime svarbiausius tyrimo etapus. Visų pirma, tiriant SEM modelį, įvertinami nežinomi modelio parametrai. Tai galima padaryti įvairiais būdais. Dažniausiai naudojami didžiausiojo tikėtinumo (ML), apibendrintojo mažiausiųjų kvadratų (GLS) ir asimptotiškai nepriklausomo nuo skirstinio (ADF) metodai. Pasirinktas metodas gali lemti visas tolesnes išvadas. Laimei, dažniausiai skirtingais metodais gauti įverčiai panašūs. Visuose nagrinėtuose pavyzdžiuose, kuriuose parametrų vertinimo metodas nenurodytas buvo taikomas ML metodas.

Pirmoji problema galima iškilti nepavykus rasti parametrų įverčių.

Parametrų įverčių nepavyksta rasti, jei:

a) iteracinis procesas nekonverguoja,
b) yra grubių modelio sudarymo klaidų,
c) yra labai stipriai koreliuojančių kintamųjų.

Įverčių skaičiavimas yra iteracinis procesas: iš pradžių parenkami pradiniai įverčiai, o po to jie įvertinami iš naujo. Procesas tęsiamas tol, kol naujieji įverčiai nuo prieš tai buvusiųjų beveik nesiskiria. Statistinėse programose yra nustatytas maksimalus leistinų perskaičiavimų skaičius. Ne visada tiek perskaičiavimų užtenka. Žinoma, iteracijų skaičių galima padidinti. Kartais padeda.

Vis dėlto net ir labai didelis iteracijų skaičius negelbsti padarius grubių modelio sudarymo klaidų, pavyzdžiui, pamiršus fiksuoti vieną iš latentinio faktoriaus svorių, neįtraukus endogeninių kintamųjų liekamųjų paklaidų ir pan. Be to, net ir teoriškai identifikuojamo modelio parametrų įverčių rasti nepavyks, jeigu dalis indikatorių labai stipriai koreliuos tarpusavyje (multikolinearumo problema).

Suradus parametrų įverčius, vertinamas:

1 bendrasis modelio tinkamumas duomenims,

2 kiekvieno parametro statistinis reikšmingumas ir prasmė.

Aptarsime modelio bendrojo tinkamumo duomenims idėją. Kintamųjų priklausomybę apibrėžia jų kovariacijos (koreliacijos). Žinome stebimų kintamųjų (indikatorių) kovariacijų įverčius. Apskaičiuojame, kokios tos kovariacijos turėtų būti, jeigu tiramas SEM modelis būtų teisingas. Palyginame kovariacijų įverčius. Jeigu skirtumai tarp to, kas yra, ir to, kas turėtų būti, nedideli, tai sakome, kad modelis ir duomenys gerai suderinti

(modelis tinka duomenims). Vartojant matricinius žymenis, tai reikštų, kad lyginame $\mathbf{S}$ su $\widehat{\mathbf{\Sigma}}_M$, norėdami padaryti išvadas apie $\mathbf{\Sigma}$ ir $\mathbf{\Sigma}_M$ skirtumus.

Ar modelis duomenims tinka, nustatoma:

a) peržiūrėjus standartizuotąją liekamųjų kovariacijų matricą,
b) patikrinus hipotezę apie χ^2 statistikos reikšmę,
c) atsižvelgus į suderinamumo indeksų reikšmes.

Liekamųjų kovariacijų matrica – tai matricų skirtumas $\mathbf{S} - \widehat{\mathbf{\Sigma}}_M$. Jis gaunamas iš stebimų kintamųjų kovariacijų atimus tas kovariacijas, kurios turėtų būti, jeigu sudarytasis modelis teisingas. Standartizuotoji liekamųjų kovariacijų matrica gaunama kiekvieną kovariacijų skirtumą padalijus iš jo standartinės paklaidos. Gautos reikšmės turi apytikslį standartinį normalųjį skirstinį. Vertinama vizualiai („iš akies“) ar daug skirtumų absoliučiuoju didumu viršija 2. Kas tai yra „daug“ – griežtai neapibrėžta, todėl toks metodas gana subjektyvus.

Matricų skirtumo $\mathbf{S}-\widehat{\mathbf{\Sigma}}_M$ didumą nusako χ^2 statistikos reikšmė. Kuo didesnis matricų skirtumas, tuo didesnė ir statistikos reikšmė. Esant teisingoms SEM modelio prielaidoms, yra žinomas χ^2 statistikos skirstinys, todėl galima patikrinti statistinę hipotezę, ar $\mathbf{\Sigma} - \mathbf{\Sigma}_M$ skiriasi nuo nulio. Šią hipotezę patogiau suformuluoti taip:

$$\begin{cases} H_0\text{: modelis gerai tinka duomenims,} \\ H_1\text{: modelis duomenims tinka blogai.} \end{cases}$$

Ar atmesti, ar priimti nulinę hipotezę, lemia pasirinktas reikšmingumo lygmuo α (dažniausiai $\alpha = 0{,}05$) ir gautoji p reikšmė.

Jeigu $p \geqslant \alpha$, tai H_0 neatmetame ir darome išvadą, kad modelis gerai tinka duomenims.
Jeigu $p < \alpha$, tai H_0 atmetame ir darome išvadą, kad modelis duomenims netinka.

Daugelio statistikų nuomone, šią sprendimo taisyklę tikslingiausiai taikyti tada, kai bendrasis duomenų skaičius yra 100–200. Kai duomenų labai daug, susiduriame su įprasta problema – beveik visada H_0 atmesime. Tuo tarpu ne visi statistiškai reikšmingi skirtumai yra tyrimui svarbūs. Taigi galime apibendrinti:

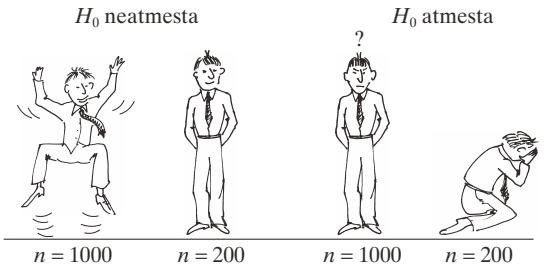
1 Jei duomenų daug, o H_0 neatmetama, tai modelis duomenims tinka puikiai.

2 Jei duomenų daug, o H_0 atmetama, tai reikia nagrinėti kitus rodiklius (pvz., liekamąją koreliacijų matricą), nes modelis duomenims gali ir tikti, ir netikti.

3 Jei duomenų mažiau nei 100, o H_0 neatmetama, tai sakyti, kad modelis duomenims tinka, galima tik turint labai paprastus modelius.

4 Jei duomenų mažiau nei 100, o H_0 atmetama, tai aišku, kad modelis duomenims netinka.

Taigi, kai $n = 1000$ ir hipotezė H_0 neatmetama, galima iš džiaugsmo saikingai pašokinėti, nes modelis puikiai tinka duomenims. Ir priešingai, jei $n = 1000$ ir hipotezė H_0 atmetama, tai modelis nebūtinai yra blogas.



Modelio tinkamumą duomenims nusako suderinamumo indeksai. Išsamiai juos aprašėme 2.6 skyrelyje. Todėl čia pateikiame tik suvestinę 3.3.4 lentelę. Nebūtina atsižvelgti į visus indeksus, užtenka pasirinkti 3–4. Atkreipiame dėmesį, kad dažnai randamas ne tik pats *RMSEA*, bet ir jo 90 % pasikliautinasis intervalas. Jeigu apatinis intervalo rėžis mažesnis už 0,05, o viršutinis rėžis mažesnis už 0,10, tai interpretuojama kaip pakankamas gero modelio tikimo duomenims įrodymas. Nagrinėjant mažas imtis, *AGFI* indeksas gali būti nepagrįstai mažas. *IFI* ir *TLI* indeksai gali būti taikomi dviem modeliams palyginti. Indeksai neatskleidžia blogo tikimo priežasčių. Be to, reikia atsiminti, kad geras tinkamumas ir stiprus kintamųjų ryšys nėra vienas ir tas pats. Galima gauti puikiai duomenims tinkantį modelį, nors kintamieji nekoreliuos.

Žinant bendrąjį modelio tinkamumą duomenims, reikia ištirti kiekvieno parametro įvertį.

Tiriant SEM modelio parametrus:

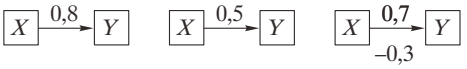
1) atsižvelgiama į parametrų ženklus,
2) atsižvelgiama į parametrų didumą,
3) tikrinamos hipotezės, ar parametrai skiriasi nuo nulio.

Parametrų ženklai turi atitikti teoriją, kuria remiantis sudarytas modelis. Modelio kintamieji koreliuoja, todėl kartais iškyla netikėtų problemų. Pavyzdžiui, atskirai nagrinėjant,

3.3.4 lentelė. SEM modelio suderinamumo indeksai

Suderinamumo indeksas	Santrumpa	Gera reikšmė	Priimtina reikšmė	Savybės
Normuotasis chi kvadratas	*NC*	$NC \leqslant 2$	$NC \leqslant 3$	Tinka vidutinio didumo imtims
Kvadratinė šaknis iš vidutinės aproksimacijos paklaidos	*RMSEA*	$RMSEA \leqslant 0{,}05$	$RMSEA \leqslant 0{,}08$	Mažoms imtims nepagrįstai mažas
Kvadratinė šaknis iš standartizuotosios vidutinės liekanos	*SRMR*	$SRMR \leqslant 0{,}05$	$SRMR \leqslant 0{,}08$	Mažas, kai yra daug vertinamų parametrų arba didelė imtis
Pataisytasis modelio suderinamumo indeksas	*AGFI*	$AGFI \geqslant 0{,}95$	$AGFI \geqslant 0{,}90$	Didėjant duomenų skaičiui, didėja
Prieaugių suderinamumo indeksas	*IFI*	$IFI \geqslant 0{,}95$	$IFI \geqslant 0{,}90$	Mažai jautrus imties didumui
Takerio–Liuiso indeksas	*TLI*	$TLI \geqslant 0{,}95$	$TLI \geqslant 0{,}90$	Mažai jautrus imties didumui
Sąlyginis suderinamumo indeksas	*CFI*	$CFI \geqslant 0{,}95$	$CFI \geqslant 0{,}90$	Mažai jautrus imties didumui

kelio nuo X iki Y koeficientas teigiamas, kelio nuo Y iki X koeficientas teigiamas, o į modelį įtraukus abudu kelius, koeficientai tampa priešingų ženklų.



Šiuo atveju būtų tik vienas patarimas – tobulinti modelį.

Parametrų didumas leidžia įsivaizduoti priklausomybių stiprumą. Beveik visada vertinami ir lyginami standartizuotojo modelio parametrai. Standartizuoto modelio visų kintamųjų skalės suvienodinamos. Tada galima palyginti kelių koeficientus. Netiesiogines kintamųjų priklausomybes rodo jau ne kovariacijos, o koreliacijos. Koreliacija, absoliučiuoju didumu mažesnė už 0,2, yra labai silpna, o didesnė už 0,8 – labai stipri. Tai palengvina interpretaciją.

Kartais gaunami su statistikos teorija nesuderinami įverčiai – neigiamos dispersijos arba koreliacijos didesnės už vienetą (mažesnės už -1). Tuomet iškart aišku, kad modelis blogas.

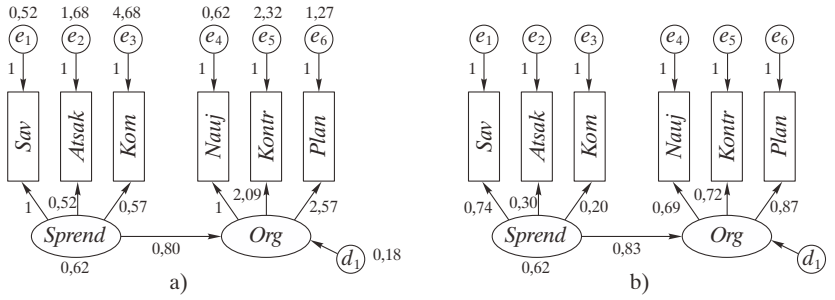
Hipotezės, ar parametrai nėra lygūs nuliui, tikrinamos turint nestandartizuotus duomenis. Statistinė hipotezė atrodo taip:

$$\begin{cases} H_0\colon & \text{parametras} = 0, \\ H_0\colon & \text{parametras} \neq 0. \end{cases}$$

Parametras statistiškai reikšmingai skiriasi nuo 0, jei p reikšmė $< \alpha$; čia α – pasirinktasis reikšmingumo lygmuo, t. y. tikimybė paskelbti, kad parametras nelygus nuliui, ir apsirikti. Dažniausiai $\alpha = 0{,}05$. Blogai, kai parametras statistiškai reikšmingai nesiskiria nuo nulio, nes tai rodiklis, kad modelyje šis parametras nereikalingas.

Jei p reikšmė ne mažesnė už α, tai parametras statistiškai nereikšmingas. Modelį reikėtų tobulinti.

Žinoma, tai rekomenduojama tik tada, kai imtyje yra pakankamai daug duomenų. Gali būti ir priešingai. Jeigu stebėjimų yra labai daug, tai net labai mažas parametras statistiškai reikšmingai skirsis nuo nulio. Todėl, tikrinant hipotezes apie parametrų reikšmes, reikia atsižvelgti ne tik į p reikšmes, bet ir į pačių parametrų didumą.



3.3.11 pav. 3.3.1 pavyzdžio modelių įverčiai: a) nestandartizuoto, b) standartizuoto

3.3.6 pavyzdys. Tiriame 3.3.1 pavyzdžio (žr. 3.3.2 pav.) SEM modelį. Imties didumas $n = 148$. Apskaičiavę įverčius didžiausiojo tikėtinumo (ML) metodu, gauname: $\chi^2 = 5{,}4$, $p = 0{,}71$. Tarkime, reikšmingumo lygmuo $\alpha = 0{,}05$. Matome, kad $p > 0{,}05$. Suderinamumo indeksai yra tokie: $RMSEA = 0{,}00$, $IFI = 1{,}01$, $TLI = 1{,}02$, $CFI = 1$. Darome išvadą, kad apskritai modelis gerai tinka duomenims. Visi parametrų įverčiai atitinka teoriją ir statistiškai reikšmingai skiriasi nuo nulio. Nestandartizuotas ir standartizuotas modeliai pavaizduoti 3.3.11 paveiksle. Nestandartizuotajame modelyje šalia kintamųjų nurodyti jų dispersijų įverčiai. Didelė kintamojo *Kom* dispersija (4,68) ir ganėtinai mažas standartizuotasis kelio koeficientas (0,20) kelia tam tikrų abejonių, ar šis kintamasis reikalingas.

Matome, kad visi 3.3.6 pavyzdyje gauti įverčiai labai panašūs į 3.2.2 pavyzdžio CFA modelio įverčius (žr. 3.2.6 lentelę ir 3.2.7 pav.). Iš tikrųjų visi modelio suderinamumo įverčiai *identiški*. Turime du ekvivalenčius modelius. Kurį laikyti teisingesniu – lemia paties tyrėjo nuomonė.

3.7. SEM modelio interpretavimas ir tobulinimas

Jei iš pirmo karto nepavyko, sunaikink visus įrodymus, kad bandei.
S. Ohanian

Modeliai sudaromi jau žinant, kaip juos interpretuoti. Jeigu modelis gerai tinka duomenims, visi parametrai statistiškai reikšmingai skiriasi nuo nulio, yra pakankamai dideli ir „teisingų“ ženklų, tai belieka nustatyti, kurie modelio ryšiai stipresni, o kurie – silpnesni. Tai galima padaryti atsižvelgus į standartizuotojo modelio parametrų reikšmes.

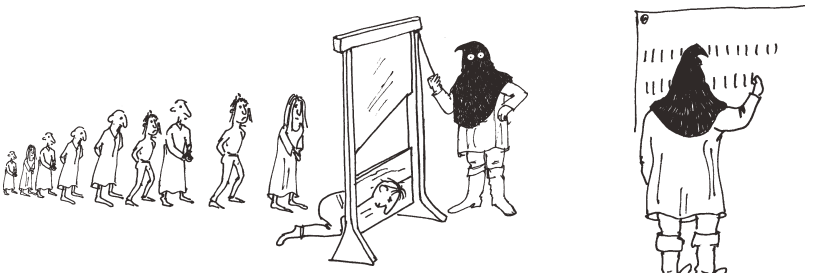
3.3.7 pavyzdys. Jau žinome, kad 3.3.1 pavyzdžio SEM modelis gerai tinka duomenims. Iš standartizuotojo modelio diagramos (žr. 3.3.11 pav., b) matome, kad organizaciniai gabumai *Org* beveik vienodai pasireiškia darbo planavimu (*Plan*), kontroliavimu (*Kontr*) ir naujovių diegimu (*Nauj*), atitinkami faktoriniai svoriai 0,69–0,87. Tuo tarpu gebėjimas priimti sprendimus *Sprend* stipriai pasireiškia vadybininko savarankiškumu *Sav* (svoris 0,74) ir kur kas silpniau atsakomybės jausmu *Atsak* (svoris 0,30) bei gebėjimu bendrauti su kitais vadybininkais *Kom* (svoris 0,20).

Daug sudėtingiau, kai pradinis modelis nėra geras. Tada prasideda kitų modelių paieškos. Blogai, jeigu jos formalios. Tyrėjas „įsijaučia “ ir pradeda bandyti įvairias kintamųjų priklausomybes, vis pasitikrindamas, ar pagerėjo suderinamumo indeksų reikšmės. Dažniausiai po kiek laiko pavyksta gauti puikiai *šiuos konkrečius duomenis* aprašantį modelį. Lieka visai neaišku, ar toks modelis tinka kitoms imtims. Šiuo atveju reikėtų išbandyti modelį bent su atsitiktiniu imties poaibiu. Be to, formali geriausiai konkretiems duomenims tinkančio modelio paieška retai kada padeda rasti modelį, atitinkantį tiriamojo dalyko teoriją.

Nebūtinai geresnis tas modelis, kuris geriau tinka duomenims.
Modelis turi atitikti tiramojo dalyko teoriją.

Dar kartą pabrėžiame, kad priimant sprendimus visų pirma reikia vadovautis savo galva. Visais atvejais labai svarbu gerai išmanyti tiriamąjį dalyką.

Modelį galima pakeisti pašalinus dalį nereikšmingų ar turinčių didelę standartinę paklaidą parametrų. Modelį galima ir išplėsti pridėjus naujų kelių arba įtraukus naujus kintamuosius. Visais atvejais ieškoma modelio, pakankamai gerai aprašančio duomenis.



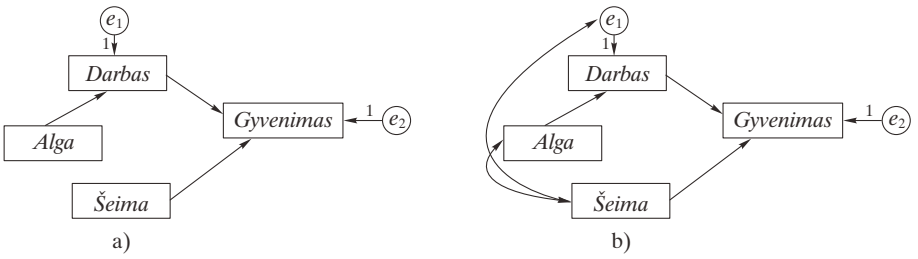
Pirma galva... Po to – statistika

Ieškant papildomų kelių, kai kuriomis statistinėmis programomis (AMOS, LISREL) skaičiuojami vadinamieji modifikavimo indeksai (MI[41]). Jie skaičiuojami kiekvienai galimai papildomai koreliacijai ir priežastiniam keliui. Vertinama, per kiek apytiksliai sumažės χ^2 statistikos reikšmė, į modelį įtraukus naująjį ryšį. Bendrasis modelio tinkamumas tuo geresnis, kuo mažesnė statistikos χ^2 reikšmė, todėl natūralu modelį papildyti tais ryšiais, kuriems esant modifikavimo indeksai didžiausi. Manoma, kad verta kreipti dėmesį tik į tokius ryšius, kurių MI ne mažesni kaip 4.

3.3.8 pavyzdys. Prisiminkime 1.1 skyrelio pavyzdį. Naudojantis specialiais klausimynais buvo įvertintas respondento pasitenkinimas gyvenimu (*Gyvenimas*), santykiai darbovietėje (*Darbas*), santykiai šeimoje (*Šeima*) ir vidutinė respondento alga (*Alga*). Suprastinta spėjamo modelio diagrama pavaizduota 3.3.12 paveiksle, a). Naudojantis AMOS programa, ML metodu rasti suderinamumo rodikliai blogi: $\chi^2 = 93$ ($p = 0{,}00$), *RMSEA* = 0,36. Modifikavimo indeksai:

1) koreliacijos $Šeima \leftrightarrow Alga$ $MI = 48{,}8$,
2) koreliacijos $Šeima \leftrightarrow e_1$ $MI = 27{,}2$,
3) kelio $Šeima \rightarrow Darbas$ $MI = 27{,}2$.

Į modelį įtraukus koreliaciją tarp kintamųjų *Šeima* ir *Alga*, suderinamumo rodikliai šiek tiek pagerėja: $\chi^2 = 38$ ($p = 0{,}00$), $RMSEA = 0{,}28$. Koreliacijos tarp kintamojo *Šeima* ir liekamosios paklaidos e_1 modifikavimo indeksas $MI = 34{,}6$ yra didžiausias. Įtraukus į modelį ir šią koreliaciją, gaunamas jau gerai duomenims tinkantis modelis (žr. 3.3.12 pav., b): $\chi^2 = 0{,}55$ ($p = 0{,}46$), $RMSEA = 0{,}00$. Tinkamos ir kitų suderinamumo indeksų reikšmės: $AGFI = 0{,}98$, $TLI = 1$, $CFI = 1$. Be to, visi parametrų įverčiai statistiškai reikšmingi. Gavome modelį, alternatyvų 3.3.5 paveikslo modeliui. Beje, pagal visus suderinamumo rodiklius abudu modeliai identiški. Taigi turime du ekvivalenčius modelius. Įdomu tai, kad 3.3.5 diagramos modelį būtume gavę, pradėję keisti ne koreliaciją, o kelią $Šeima \rightarrow Darbas$.



3.3.12 pav. Modeliai: a) pradinis, b) pakeistas

[41] MI – angl. *Modification Index.*

Kartais vietoj MI tyrimui naudojami Lagranžo daugikliai, kurie interpretuojami panašiai.

Lygindami tiems patiems duomenims pritaikytus modelius, atsižvelgiame ir į *AIC*, ir į *ECVI* informacinių indeksų reikšmes.

Esant mažesnei *AIC* arba *ECVI* reikšmei, modelis geriau tinka duomenims.

3.3.9 pavyzdys. Prisiminkime 3.3.6 pavyzdį. Pažvelgę į 3.3.11 paveikslo, b) diagramą, matome, kad silpniausia priklausomybė yra tarp latentinio sprendimų priėmimo faktoriaus *Sprend* ir kintamojo *Kom*, atspindinčio komunikavimą su kolegomis vadybininkais. Pašalinkime iš modelio kintamąjį *Kom*. Naujojo modelio visi suderinamumo indeksai panašūs į 3.3.6 pavyzdžio modelio parametrus ($\chi^2 = 1{,}98$, $p = 0{,}76$, $RMSEA = 0{,}00$, $IFI = 1{,}01$, $TLI = 1{,}02$, $CFI = 1$). Naujasis modelis paprastesnis, todėl jo $AIC = 23{,}98$, o nagrinėto 3.3.5 pavyzdyje modelio $AIC = 31{,}39$. Vis dėlto reikia patiems nuspręsti, ar pašalinti iš modelio kintamąjį *Kom*, t. y. pasilikti tą modelį, kuris atrodo logiškesnis. Pavyzdžiui, mes manytume, kad *Kom* iš modelio galima pašalinti.

Kai vienas modelis yra kito modelio dalis, galima patikrinti hipotezę, ar jie vienodai gerai tinka duomenims. Išsamiai šią procedūrą aprašėme 2.8 skyrelyje, todėl dabar tik trumpai pakartosime pagrindinius faktus. Tikrinama tokia statistinė hipotezė:

$$\begin{cases} H_0\text{: abu modeliai duomenims tinka vienodai,} \\ H_1\text{: modeliai duomenims tinka nevienodai.} \end{cases}$$

Tarkime, kad pirmasis modelis (χ_1^2, df_1) yra antrojo modelio (χ_2^2, df_2) dalis. Kuo daugiau kintamųjų ryšių aprašome, tuo mažiau lieka laisvės laipsnių. Mažesnė ir χ^2 statistikos reikšmė. Jeigu $\chi_D^2 = \chi_1^2 - \chi_2^2 > \chi_\alpha^2(df_1 - df_2)$, darome išvadą, kad modelių bendrasis tinkamumas duomenims statistiškai reikšmingai skiriasi. Čia $\chi_\alpha^2(df_1 - df_2)$ yra chi kvadrato su $df_1 - df_2$ laisvės laipsniu α lygmens kritinė reikšmė.

3.3.10 pavyzdys. Pažvelgę į 3.3.12 paveikslą, matome, kad a) modelis yra b) modelio dalis. Palyginsime, ar abudu modeliai vienodai gerai aprašo duomenis. Modelio a) $\chi_1^2 = 93{,}1$, $df_1 = 3$. Modelio b) $\chi_2^2 = 0{,}6$, $df_2 = 1$. Tada $\chi_D^2 = 93{,}1 - 0{,}6 = 92{,}5$, $df_1 - df_2 = 3 - 1 = 2$. Tegul reikšmingumo lygmuo $\alpha = 0{,}05$. Iš kritinių reikšmių lentelių randame $\chi_{0{,}05}^2(2) = 5{,}9$. Kadangi $92{,}5 > 5{,}9$, tai darome išvadą, kad vienas iš modelių statistiškai reikšmingai geriau aprašo duomenis.

3.8. Modeliai su parametrų apribojimais

Blogas planas – geriau nė jokio plano.
Babilono bokšto statytojai

Kartais tenka pareikalauti, kad SEM modelio parametrai tenkintų papildomas sąlygas. Šiame skyrelyje panagrinėsime tris būdingus atvejus:

1. kai kurios modelio parametrų reikšmės yra fiksuotos,
2. kai kurie modelio parametrų santykiai yra fiksuoti,
3. tas pats modelis taikomas kelioms imtims.

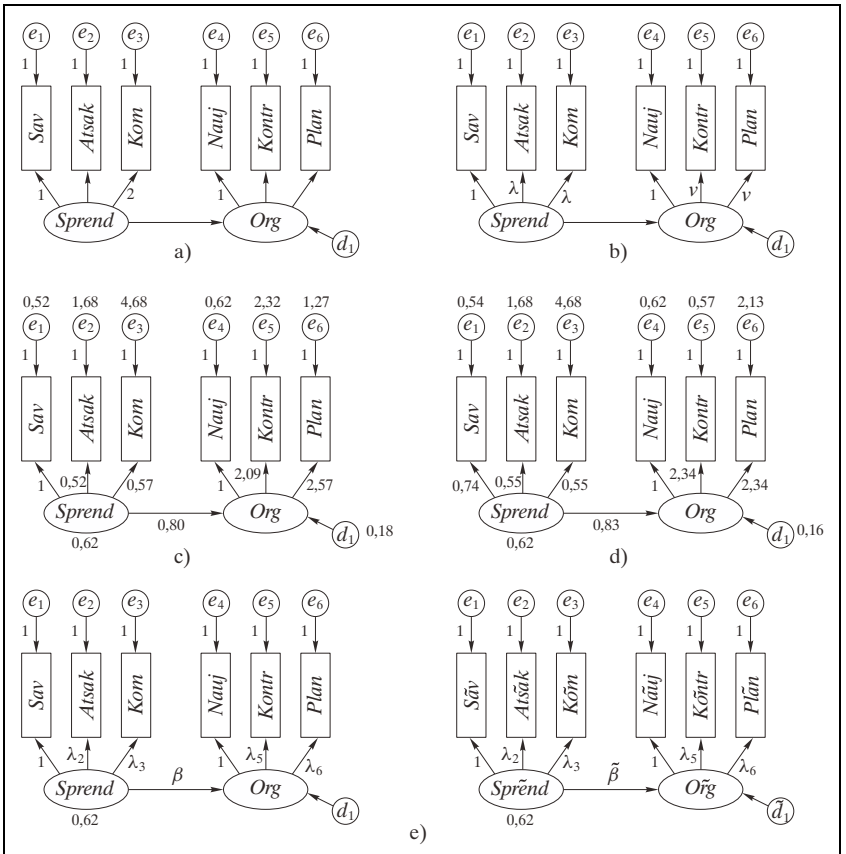
Modeliai, kurių dalies parametrų reikšmės fiksuotos. Kartais parametro reikšmė būna žinoma arba numanoma. Tuomet ją tiesiog įtraukiame į modelį. Tolesnis modelio tyrimas niekuo nesiskiria nuo aprašyto ankstesniuose skyreliuose.

3.3.11 pavyzdys. Prisiminkime 3.3.1 pavyzdžio modelį (žr. 3.3.2 pav.) apie vadybininką, kurio profesinę veiklą veikia du latentiniai faktoriai: organizaciniai gabumai (*Org*), pasireiškiantys darbo planavimu (*Plan*), kontroliavimu (*Kontr*) ir naujovių diegimu (*Nauj*), ir gebėjimas priimti sprendimus (*Sprend*), pasireiškiantis vadybininko savarankiškumu (*Sav*), atsakomybės jausmu (*Atsak*) ir gebėjimu bendrauti su kitais vadybininkais (*Kom*). Tarkime, kad dėl tam tikrų teorinių reikalavimų nusprendėme, jog *Kom* kelio koeficientas lygus 2. Modelyje jį ir fiksuojame (žr. 3.3.13 pav., a). Beje, tolesnė analizė šio modelio teisingumo nepatvirtina ($\chi^2 = 21{,}5$, $p = 0{,}01$, $RMSEA = 0{,}09$).

Modeliai, kurių dalies parametrų santykiai fiksuoti. Tokiuose modeliuose nurodoma, kurie parametrai lygūs, nefiksuojant jų reikšmių. Sulyginti galima kelių koeficientus, dispersijas, kovariacijas ir vidurkius.

3.3.12 pavyzdys. Imkime 3.3.1 pavyzdžio modelį ir įveskime apribojimus, t. y. tarkime, kad:

1) kintamųjų *Atsak* ir *Kom* faktoriniai svoriai (kelio koeficientai) yra lygūs,
2) kintamųjų *Kontr* ir *Plan* faktoriniai svoriai (kelio koeficientai) yra lygūs.



3.3.13 pav. Modeliai, kai: a) parametras yra fiksuotas, b) koeficientai lygūs, c) nėra parametrų apribojimų, d) yra parametrų apribojimų, e) yra dvi imtys

Spėjamos priklausomybės pavaizduotos 3.3.13 paveiksle, b). Rasime modelio laisvės laipsnių skaičių. Turime 6 stebimus kintamuosius. Tarkime, kad tiriame centruotą modelį. Tada vertinimams naudojame $6(6+1)/2 = 21$ kovariaciją. Įvertinti reikia liekamųjų paklaidų $e_1, \ldots, e_6$ ir d_1 dispersijas, faktoriaus *Sprend* dispersiją, faktorinius svorius λ, ν ir β – iš viso 11 parametrų. Todėl $df = 21 - 11 = 10$. Modelis yra perteklinės identifikacijos. Suradus parametrų įverčius, paaiškėja, kad modelis gerai tinka duomenims ($\chi^2 = 8{,}6$, $p = 0{,}57$, $RMSEA = 0{,}0$, $SRMR = 0{,}03$, $TLI = 1{,}01$, $CFI = 1$). Visi parametrai, išskyrus paklaidos d_1 dispersiją, statistiškai reikšmingai skiriasi nuo 0. Standartizuotojoje liekamųjų kovariacijų matricoje visos reikšmės mažos (neviršija 1). Taigi duomenys modelio prielaidoms neprieštarauja.

Naujojo modelio parametrų įverčiai pateikti 3.3.13 paveiksle, d). Palyginti su 3.3.1 pavyzdžio rezultatais (žr. 3.3.13 pav., c), naujojo modelio parametro įverčiai pasikeitė nedaug. Pavyzdžiui, 3.17 modelyje kelio tarp *Sprend* ir *Org* koeficientas buvo lygus 0,80, o apribotame modelyje – 0,84.

Beje, neatmesta nulinė hipotezė apie liekamosios paklaidos d_1 dispersiją rodo, kad reiktų ištirti modelį be šios liekamosios paklaidos.

Tas pats modelis, taikomas kelioms imtims. Galima tikrinti, ar skirtingoms imtims tinka tokia pat diagrama aprašytas modelis. Šiuo atveju kiekviena imtis tiriama atskirai ir gaunami skirtingi parametrų įverčiai. Tačiau būna situacijų, kai norime įsitikinti, kad vienodos ne tik modelių diagramos (t. y. modelio forma), bet ir parametrų įverčiai. Pavyzdžiui, sudarėme SEM modelį, kuriame pasitenkinimui gyvenimu daro įtaką santykiai šeimoje. Norime išsiaiškinti, ar ši įtaka vienodai stipri vyrams ir moterims. Analogiškos problemos kyla, kai norime palyginti, ar visos modelio priklausomybės (arba kai kurios iš jų) yra vienodos esant skirtingų tikybų respondentams, skirtingų šalių gyventojams ir pan. Žinoma, nebūtina reikalauti, kad absoliučiai visi modelio parametrai būtų identiški visoms imtims. Pavyzdžiui, galima nagrinėti, ar vienodi faktoriniai svoriai, bet netaikyti jokių apribojimų dispersijoms.

Norėdami patikrinti, ar duomenys statistiškai reikšmingai neprieštarauja spėjimui, kad parametrų reikšmės yra lygios, susiduriame su matematinėmis problemomis. Individualūs lyginamųjų parametrų skirtumai vertinami tik vizualiai („iš akies“). Šiuo metu, lyginant du modelius, statistinės hipotezės formuluojamos ir tikrinamos tik tuo atveju, kai:

a) naudojami tie patys duomenys,
b) vienas modelis yra kito modelio dalis.

Taigi turint skirtingas imtis tiesiog palyginti dviejų modelių nepavyksta. Ši problema sprendžiama taip: aprašomas bendras jungtinis modelis, kurį sudaro submodeliai, turintys tiriamojo SEM formą, t. y. tiriamasis SEM modelis pakartojamas tiek kartų, kiek yra imčių. Vertinant parametrus, kiekvienam submodeliui naudojami tik atitinkamos imties duomenys. Vertindami jungtinio modelio ir duomenų suderinamumą, kartu įvertiname, ar pasirinktoji modelio forma tinka *visoms* imtims.

Po to jungtiniame modelyje pareikalaujame, kad atitinkami submodelių parametrai būtų lygūs. Faktiškai suvedame uždavinį į pirmąjį šio skyrelio atvejį, t. y. modelį, kurio parametrai tenkina apribojimus. Jį ir tiriame.

Modelis su apribojimais yra jungtinio modelio be apribojimų dalis, todėl dar galima patikrinti, ar abu modeliai vienodai gerai aprašo duomenis (žr. 3.7 skyrelį ir 3.3.9 pavyzdį).

Taigi tikrinant, ar tas pats SEM modelis su vienodais parametrais tinka kelioms imtims, išskiriami tokie etapai:

1 Naudojant jungtinį modelį, patikrinama, ar tiriamasis SEM modelis tinka visoms imtims. Šiame etape parametrai nefiksuojami (išskyrus vienetinius koeficientus prie liekamųjų paklaidų ir pasirinktųjų indikatorių). Jeigu jungtinis modelis be apribo-

jimų blogai aprašo duomenis, tai pasirinktojo modelio forma netinka bent vienai imčiai. Klausimas apie parametrų lygybę tampa beprasmis. Jeigu modelis duomenims tinkamas, pereinama prie kito etpo.

2 Naudojant jungtinį modelį su parametrų apribojimais, patikrinama, ar tiriamojo SEM modelio parametrus galima laikyti vienodais esant skirtingoms imtims. Jeigu ne, tai ne. Vadinasi, duomenys spėjimo nepatvirtina. Jeigu taip, tai pereinama prie kito etapo.

3 Patikrinama, ar jungtinis modelis be apribojimų statistiškai reikšmingai skiriasi nuo jungtinio modelio su apribojimais. Jeigu nesiskiria, tai daroma išvada, kad papildomi apribojimai nepablogino bendrojo modelio tinkamumo duomenims. Vadinasi, esant skirtingoms imtims, tinka tas pats SEM modelis su vienodais parametrais. Toks pat patvirtinimas gaunamas, jeigu įvedus apribojimus tinkamumas statistiškai reikšmingai pagerėja. Kiek kebliau, jeigu tinkamumas statistiškai reikšmingai pablogėja. Tuo atveju galima padaryti tik tokią išvadą – modelis su apribojimais tinka, bet iš esmės blogiau nei modelis be apribojimų.

4 Atliekamas pagalbinis tyrimas – poromis palyginami modelyje be apribojimų gauti visų submodelių parametrų įverčiai. Dideli parametrų įverčių skirtumai suteikia informacijos, kokius apribojimus reikia pakeisti. Ypač šis tyrimas naudingas, jeigu antrasis etapas buvo nesėkmingas. Parametrų įverčiai lyginami taip: įverčių skirtumas padalijamas iš šio skirtumo standartinės paklaidos ir žiūrima, ar gautoji reikšmė absoliučiuoju didumu viršija 1,96 (t. y. standartinio normaliojo skirstinio 0,975 lygmens kvantilį). Parametrai, kurių įverčiai viršija 1,96, laikomi labai besiskiriančiais[42]. Tai itin grubus ir labai apytikslis metodas, kuris vis dėlto yra naudingas, norint pagerinti modelį. Standartines paklaidas apskaičiuoti nėra paprasta, todėl parametrų įverčiai lyginimi tik esant galimybei pasinaudoti tinkama programine įranga.

3.3.13 pavyzdys. Patikrinsime, ar vadybininkų ir vadybininkių organizaciniai gabumai ir gebėjimas priimti sprendimus vienodai stipriai pasireiškia stebimuose indikatoriuose. Kitaip sakant, tikrinsime, ar taikant 3.3.1 pavyzdžio modelį, vadybininkų ir vadybininkių kelio koeficientai tokie pat. Jungtinio modelio diagrama pateikta 3.3.13 paveiksle, e). Nenorėdami įvesti naujų žymenų, visus vadybininkių parametrus ir kintamuosius pažymėjome $\sim$ (tilde). Pirmajame tyrimo etape vertinami visi jungtinio modelio be apribojimų parametrai.

Antrajame etape tiriamas jungtinis modelis su tokiais apribojimais:

$$\lambda_2 = \tilde{\lambda}_2, \qquad \lambda_3 = \tilde{\lambda}_3, \qquad \lambda_5 = \tilde{\lambda}_5, \qquad \lambda_6 = \tilde{\lambda}_6, \qquad \beta = \tilde{\beta}.$$

Trečiajame etape abu jungtiniai modeliai palyginami. Dirbant AMOS programa ir taikant ML metodą gauti rezultatai pateikiami 3.3.5 lentelėje. Atkreipiame dėmesį, kad nors jungtiniame modelyje yra 12 stebimųjų kintamųjų, tačiau vertinimams turime tik 42 kovariacijas, t. y. $6 \cdot 7/2 = 21$ kovariaciją kiekvienam submodeliui. Skirtingų submodelių kintamieji nekoreliuoja, pavyzdžiui, $\mathbb{C}\mathrm{ov}(Kom, K\tilde{o}m) = 0$ ir t. t. Matome, kad abu modeliai gerai aprašo duomenis. Įdomu tai, kad geriau duomenims tinka modelis su apribojimais. Kaip ir reikėjo tikėtis, paprastesnio modelio *AIC* yra mažesnis.

Tiriant, ar statistiškai reikšmingai skiriasi bendrasis abiejų jungtinių modelių tinkamumas duomenims, surandamas χ^2 statistikų skirtumas $\chi_1^2 - \chi_2^2 = 19,23 - 17,25 = 1,98$ ir laisvės laipsnių skirtumas $df_1 - df_2 = 21 - 16 = 5$. Tegul reikšmingumo lygmuo $\alpha = 0{,}05$. Jeigu $1,98$ būtų didesnis už $\chi^2_{0,05}(5) = 5{,}9$, tai gaėtume teigti, kad modeliai duomenis aprašo nevienodai gerai. Kadangi taip nėra – darome išvadą, kad abu jungtiniai modeliai vienodai gerai tinka duomenims. Žinoma, turint p reikšmę, išvadą padaryti dar lengviau.

3.3.5 lentelė. Jungtinių modelių parametrų įverčiai ir suderinamumo indeksai

	Jungtinis modelis			Jungtinis modelis	
Parametras	be apribojimų	su apribojimais	Parametras	be apribojimų	su apribojimais
χ^2	17,25	19,23	λ_2	0,58	0,51
df	16	21	λ_3	0,73	0,55
p reikšmė	0,37	0,57	λ_5	2,15	2,49
NC	1,08	0,92	λ_6	2,06	2,08
RMSEA	0,02	0,00	$\tilde{\lambda}_2$	0,49	0,51
AGFI	0,90	0,91	$\tilde{\lambda}_3$	0,51	0,55
IFI	0,99	1,00	$\tilde{\lambda}_5$	2,52	2,49
TLI	0,98	1,01	$\tilde{\lambda}_6$	2,15	2,08
CFI	0,99	1,00	β	0,86	0,88
AIC	69,24	61,22	$\tilde{\beta}$	0,86	0,88

Naudodamiesi AMOS programa, randame $p = 0,85$. Kadangi $p > \alpha$, nulinės hipotezės apie vienodą modelių tinkamumą neatmetame.

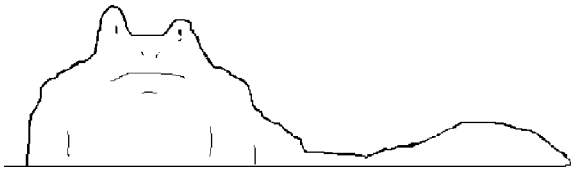
Pagalbinis tyrimas parodė, kad visi kelių koeficientų skirtumai, padalyti iš jų standartinių paklaidų, absoliučiuoju didumu yra mažesni už 1,13, todėl negalima įtarti jokio reikšmingesnio lyginamųjų parametrų skirtumo.

3.9. Nenormalieji duomenys

Gauso skirstinys „sako“ Stjudento skirstiniui: „Tu gal ir panašus į mane, tačiau vis tiek nenormalus.“

Statistikoje yra[43] terminų, kurie smagiai nuteikia su ja nesusidūrusius žmones. Įvertinys, kaip koks pianinas, gali būti ir suderintas, ir paslinktas. Skirstiniai gali turėti sunkias (kaip bebro) uodegas. Kintamieji gali būti ir normalieji, ir išsigimę[44]. Tokie baisūs, baisūs (žr. 3.3.14 pav.). Tie, kuriems teko (ne)laimė su statistika susidurti, jau niekuo nebesistebi. Jiems primename, kad kintamųjų normalumas viso labo tereiškia pasiskirstymo savybes. Normalieji kintamieji yra intervaliniai, simetriniai vidurkio atžvilgiu, o jų įgyjamos reikšmės susikoncentravusios apie vidurkį[45].

Šio skyriaus pradžioje suformuluotose SEM prielaidose paminėjome, kad stebimi kintamieji turėtų būti normaliai pasiskirstę. SEM modelių analizė atliekama ir tada, kai imties



3.3.14 pav. Nenormaliojo kintamojo tankis su sunkia uodega

[43] Toli mums iki lietuviškos kompiuterinės terminijos, bet vis šis tas.

[44] Beje, nenormalus – tai nebūtinai išsigimęs.

[45] Plačiau žr. *Statistika ir jos taikymai. I*, p. 102–105.

duomenys nedaug skiriasi nuo normaliųjų. Vis dėlto, jei kintamieji labai „nenormalūs“, tenka remtis kitokiais tyrimo metodais. Apie juos šiame skyrelyje ir pakalbėsime. Norime atkreipti dėmesį, kad nagrinėsime tik intervalinius kintamuosius. SEM modelių su kategoriniais kintamaisiais (rasė, tautybė, tikyba ir pan.) analizė dar gana menkai išplėtota.

Iš pradžių aptarsime, kaip galima nustatyti, ar duomenys yra normalieji. Tiksliai suformuluotas reikalavimas SEM modeliui skamba taip: visų stebimųjų kintamųjų vektorius turi būti pasiskirstęs pagal daugiamatį normalųjį dėsnį. Jis griežtesnis nei reikalavimas, kad kiekvienas kintamasis turėtų normalųjį skirstinį. Dažnai statistikos taikymuose šis reikalavimų skirtumas ignoruojamas ir apsiribojama atskirų kintamųjų tyrimu.

Ar kintamieji normalieji, galima patikrinti naudojantis Kolmogorovo–Smirnovo ir kt. kriterijais. Tiktai reikėtų nepamiršti, kad tokie kriterijai labai jautriai reaguoja į imties didumą – turint dideles imtis, net menkiausias nukrypimas nuo normaliojo dėsnio yra statistiškai reikšmingas. Todėl tikrinima dažniausiai gan grubiai – tiesiog pasižiūrima į asimetrijos ir eksceso koeficientus[46]. Vertinti galima dvejopai:

1 atsižvelgiant į pačias asimetrijos ir eksceso koeficientų reikšmes,

2 atsižvelgiant į koeficientų įvertinių ir standartinių paklaidų santykį.

Normaliai pasiskirsčiusio kintamojo asimetrijos ir eksceso koeficientai lygūs nuliui. Todėl kintamieji tuo mažiau panašūs į normaliuosius, kuo jų įverčiai yra „toliau nuo nulio“, t. y. absoliučiuoju didumu didesni. Koeficientų nuokrypis nuo nulio vertinamas vizualiai („iš akies“). Visuotinai priimta manyti, kad jei asimetrijos ir eksceso koeficientai moduliu neviršija vieneto, tai kintamasis pakankamai artimas normaliai pasiskirsčiusiajam. Jeigu bent vieno koeficiento modulis didesnis už du – jau labai tikėtina, kad kintamasis nėra normalusis. Jeigu tarp vieneto ir dvejeto – paliekama spręsti pačiam tyrėjui. Visais atvejais dar reikia atsižvelgti į įvertinio ir jo standartinės paklaidos santykį, vadinamąjį kritinį santykį.

Kai imtis yra pakankamai didelė, kritinis santykis turi apytiksliai standartinį normalųjį skirstinį. Standartinis normalusis skirstinys yra gerai žinomas, todėl lengvai nustatoma, kad tikimybė, jog kritinis santykis moduliu viršys 1,96, lygi 0,05. Tai leidžia tikrinti statistines hipotezes. Tegul reikšmingumo lygmuo $\alpha = 0{,}05$.

Kintamasis statistiškai reikšmingai skiriasi nuo normaliojo, jeigu jo asimetrijos arba eksceso kritinis santykis viršija 1,96.

Žinoma, nereikia skaičiaus 1,96 suabsoliutinti. Nieko tokio, jeigu kritinis santykis šią reikšmę viršija nedaug. O štai, jeigu jis didesnis kaip 3, tai jau tikėti kintamojo normalumu nepatartina.

Kintamasis yra „pakankamai normalus“, jeigu asimetrijos ir eksceso koeficientai absoliučiuoju didumu neviršija 1, o šių koeficientų kritiniai santykiai mažesni už 1,96.

Visais kitais atvejais, reikia patiems įvertinti, ar nuokrypiai nuo leistinų reikšmių labai dideli.

[46] Žr. ten pat, p. 43–44.

3.3.14 pavyzdys. Patikrinsime, ar 1.1 skyrelyje nagrinėto pavyzdžio kintamieji *Gyvenimas*, *Šeima*, *Darbas*, *Namai* yra normaliai pasiskirstę. Imtį sudarė 229 respondentai. Kintamųjų koeficientų įverčiai ir jų kritiniai santykiai pateikti 3.3.6 lentelėje.

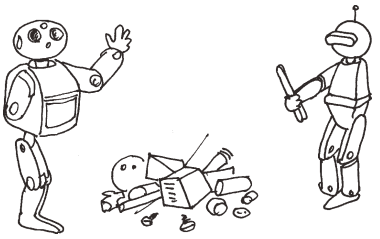
3.3.6 lentelė. Koeficientai, į kuriuos reikia atsižvelgti tikrinant normalumą

Kintamasis	Asimetrijos koeficientas	Asimetrijos koeficiento kritinis santykis	Eksceso koeficientas	Eksceso koeficiento kritinis santykis
Alga	0,39	2,4	0,10	0,32
Darbas	0,49	3,0	−0,00	−0,01
Šeima	0,32	2,0	−0,61	−1,87
Gyvenimas	−0,42	−0,42	−0,18	−0,54

Matome, kad visi eksceso įverčiai neviršija vieneto, o jų kritinės reikšmės absoliučiuoju didumu mažesnės už 1,96. Visi asimetrijos koeficientai taip pat absoliučiuoju didumu mažesni už vienetą, tačiau kintamajo *Alga* asimetrijos koeficiento kritinis santykis viršija 2, o kintamojo *Darbas* – viršija 3. Iš visų kintamųjų rimčiau galima abejoti tik kintamojo *Darbas* normalumu, tačiau apskritai duomenys pakankamai panašūs į normaliai pasiskirsčiusius ir galima taikyti standartinius SEM analizės metodus.

Ką daryti, jeigu nusprendžiame, kad duomenys nėra normaliai pasiskirstę? Nekreipiant į tai dėmesio ir taikant standartinius tyrimo metodus, gaunami iškreipti parametrų įverčiai ir nepatikimos statistinės išvados. Todėl naudojami nuo skirstinio nepriklausomi metodai. Apie vieną tokį parametrų įverčiams taikomą ADF metodą jau užsiminėme 2.6 skyrelyje. Didžiausias jo trūkumas – imtis turi būti itin didelė ($n \geqslant 1000$), o tokios imtys tyrimuose pasitaiko retai.

Kitas būdas atsižvelgti į nuokrypius nuo normaliojo skirstinio – chi kvadrato statistikai pritaikyti Satoros–Bentlerio[47] transformaciją. Transformuotoji chi kvadrato statistika vertinama ir interpretuojama, kaip ir klasikinio tyrimo. Deja, programų, kuriose būtų skaičiuojama transformuotoji Satoros–Bentlerio chi kvadrato statistika, yra labai nedaug.



Kai duomenys nėra normalieji – belieka saviranka

Dar vienas būdas išvengti normalumo reikalavimo – pasinaudoti saviranka[48]. Savirankos idėja tokia: į imtį žiūrima kaip į visos populiacijos atspindį. Iš visų imties duomenų parenkama daug imčių (tokio pat didumo, kaip ir pradinė imtis, bet renkama su grąžinimu). Skaičiuojami šių imčių parametrų įverčiai. Galutinis parametro įvertis – visų savirankos imčių įverčių vidurkis.

[47] Albert Satorra – ispanų statistikas, Peter Bentler – amerikiečių statistikas.

[48] Tai lietuviškas angliško termino *bootstrap* variantas, už kurį esame dėkingi D. Krapavickaitei ir A. Plikusui.

[49] Kenneth A. Bollen – amerikiečių sociologas, Robert A. Stine – amerikiečių statistikas.

Taikant saviranką, skaičiuojama:

1. Boleno–Staino[49] p reikšmė. Ji skaičiuojama vertinant bendrąjį modelio tinkamumą ir traktuojama kaip chi kvadrato statistikos p reikšmė. Tegul pasirinktasis reikšmingumo lygmuo $\alpha = 0{,}05$. Jei $p \geqslant 0{,}05$, tai sakoma, kad SEM modelis pkankamai gerai tinka duomenims.
2. Parametrų įverčiai ir jų standartinės paklaidos. Norint padaryti išvadą, kad parametras statistiškai reikšmingai skiriasi nuo nulio, reikia įvertį padalyti iš standartinės paklaidos (rasti įverčio kritinę reikšmę) ir patikrinti, ar ji absoliučiuoju didumu viršija 1,96. Jei taip, tai esant reikšmingumo lygmeniui $\alpha = 0{,}05$ daroma išvada, kad parametras statistiškai reikšmingai skiriasi nuo nulio (statistiškai reikšmingas).
3. Parametrų įverčių pasikliautinieji intervalai.

Formuluoti statistines išvadas, kaip nurodyta antrajame punkte, galima tik iš bėdos, nes netiesiogiai daroma prielaida, kad parametro įvertis elgiasi kaip normalusis kintamasis. Prieš tai vertėtų įsitikinti, ar savirankos būdu gauti įvertiniai nedaug skiriasi nuo apskaičiuotųjų įprastu (pvz., ML) metodu.

3.3.15 pavyzdys. Pritaikysime savirankos metodą 3.3.1 SEM pavyzdžiui. Parenkame 500 savirankos imčių. Joms tirti taikome ML metodą. Gauta Boleno–Staino statistikos p reikšmė 0,77 yra didesnė už apskaičiuotąją tiesioginiu ML metodu $p = 0{,}71$. Taigi bendrasis modelio tinkamumas duomenims labai geras. Pateikiamoje 3.3.7 lentelėje nurodyti ML būdu gauti parametrų įverčiai, savirankos būdu gautų įverčių vidurkiai, tų vidurkių standartinės paklaidos ir pasikliautinieji intervalai. Matome, kad įverčiai beveik nesiskiria. Tai neturėtų stebinti, nes visi kintamieji yra apytiksliai normalūs. Gana daug yra kintamųjų, kurių kritiniai santykiai mažesni už 1,96. Ypač mažas liekamosios paklaidos d_1 kritinis santykis, dar kartą patvirtinantis, kad šią paklaidą pašalinus modelį galima būtų patobulinti.

3.3.7 lentelė. Normalumo tyrimo koeficientai

Parametras	ML įvertis	Savirankos įvertis	Standartinė paklaida	Pasikliautinasis intervalas
λ_2	0,52	0,53	0,23	[0,13, 0,90]
λ_3	0,56	0,54	0,34	[0,06, 1,22]
λ_5	2,09	2,12	0,27	[1,67, 2,58]
λ_6	2,57	2,59	0,28	[2,20, 3,10]
β	0,80	0,87	0,49	[0,35, 1,67]
$\mathbf{D}Sprend$	0,62	0,71	0,43	[0,27, 1,34]
σ_1^2	0,52	0,41	0,42	[−0,15, 0,85]
σ_2^2	1,68	1,66	0,21	[1,41, 2,09]
σ_3^2	4,68	4,59	0,51	[3,94, 5,71]
σ_4^2	0,62	0,61	0,09	[0,50, 0,81]
σ_5^2	2,32	2,26	0,39	[1,79, 3,02]
σ_6^2	1,27	1,24	0,33	[0,73, 1,77]
σ_d^2	0,18	0,15	0,25	[−0,22, 0,48]

Nėra kokių nors griežtų taisyklių, kiek savirankos imčių reiktų atrinkti. Dažniausiai tyrimams naudojama 200, 500, 1000 savirankos imčių. Skyrelio pabaigoje norime atkreipti dėmesį, kad metodas gerai tinka tik tada, kai tiriama imtis gerai atsispindi populiacijos ypatybes, t. y. imtis yra reprezentatyvi.

3.10. SEM analizei skirtos programos

Klysti žmogiška, o suversti kaltę kompiuteriui – dar labiau.
R. Orbenas

Yra nemažai statistinių programų SEM modelių analizei atlikti. Paminėsime tik populiariausias iš jų: LISREL, AMOS, R, EQS, Mplus. Visi pavadinimai skamba kaip burtažodžiai, nors dažniausiai tėra angliškų žodžių santrumpos.

Kaip SEM modelius analizuoti naudojantis konkrečia programa, nurodyta tinklalapyje ***http://stat.vadoveliai.lt***. Šiame skyrelyje apsiribosime trumpomis pastabomis. Reikia turėti omenyje, kad atsiranda vis naujesnės programų versijos. Vis dėlto čia pateikiamos programėlės turėtų tikti ir joms.

AMOS[50] ***programa***. Ją sukūrė Dž. Arbuklas[51]. Mūsų nuomone, tai pati draugiškiausia vartotojui SEM analizės programa. Joje numatyta grafinė vartotojo sąsaja, t. y. galima tiesiog braižyti SEM modelio diagramą, o po to tikrinti, ar ji tinka. Šioje knygoje pateikiamos diagramos sutampa su vartojamomis AMOS programoje.

Šiuo metu AMOS programa yra SPSS korporacijos nuosavybė, suderinta su visais SPSS programiniais produktais. Pavyzdžiui, AMOS programa gali naudoti SPSS duomenų failus. Yra keletas AMOS programos versijų. Šios knygos pavyzdžiai analizuoti su AMOS16. Nemokama studentiška paketo versija (atitinkanti programą AMOS5) turi kai kuriuos apribojimus, tačiau visiškai tinka analizuoti daugumą šio vadovėlio pavyzdžių ir uždavinių.

Jeigu SEM modelis labai didelis, tai nubraižyti jo diagramą tampa gan sudėtinga (nebeužtenka ekrano pločio). Tuo atveju patogiau modelio priklausomybes aprašyti specialia programa. Programos sintaksė paprasta. Pateikiame pavyzdį, kaip galima aprašyti 3.3.2 paveiksle pavaizduotos diagramos koeficientus.

```
Sem.AStructure("Sav = (1)Sprend + (1)e1")
Sem.AStructure("Atsak = Sprend + (1)e2")
Sem.AStructure("Kom = Sprend + (1)e3")
Sem.AStructure("Nauj = (1)Org + (1)e4")
Sem.AStructure("Kontr = Org + (1)e5")
Sem.AStructure("Plan = Org + (1)e6")
Sem.AStructure("Org = Sprend + (1)d1")
```

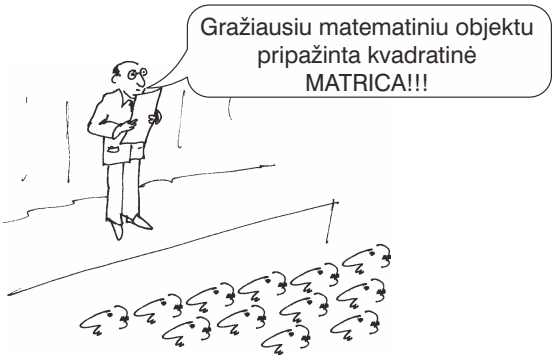
[50] AMOS – angl. *Analysis of Moment Structures.*

[51] James L. Arbuckle – amerikiečių psichologas.

Matome, kad nuosekliai aprašomi matavimo modeliai, o po to – struktūrinis modelis. Nesunku programoje aprašyti ir kintamųjų arba liekamųjų paklaidų koreliaciją. Pavyzdžiui, tarkime, kad koreliuoja liekamosios paklaidos e_1 ir e_2. Tada programą papildome sakiniu $Sem.AStructure("e1 < --- > e2")$.

Grafinė paketo aplinka turi apsaugą nuo neleistinų modelių – kai kurių kelių tiesiog nepavyks nubraižyti. Be to, AMOS programoje numatyta modelio konstrukcijos paieška, t. y. galima modelio diagramoje nurodyti dalį kelių, kurių reikalingumu abejojame. Po to automatiškai tikrinami visi modeliai be pirmojo iš pažymėtų kelių, be antrojo, be pirmojo ir antrojo ir t. t. Ši programos galimybė kelia labai didelę pagundą visą modelio paiešką automatizuoti. Tačiau taip elgiantis iškyla dvi problemos. Pirmoji problema techninė: nurodžius tikrinti daug kelių, net galingiausias kompiuteris turės darbo labai ilgam. Antroji problema kvalifikacinė: programos vartotojas turi išmanyti tiriamo reiškinio teoriją. Priešingu atveju gaus tik formaliai geriausiai *turimai* duomenų aibei tinkantį modelį, kuris gali būti ne itin prasmingas.

LISREL[52] ***programa***. Tai pati pirmoji SEM skirta programa. Ją dvidešimtojo amžiaus septintajame dešimtmetyje sukūrė K. Joreskogas ir D. Sorbomas[53]. Iš pradžių LISREL programa buvo skirta gerai statistiką išmanantiems žmonėms, suprantantiems matricinį SEM užrašą ir mokantiems visas matricas aprašyti. Tuo buvo atstumta nemaža dalis potencialių vartotojų, neįžvelgiančių matematikos (ir matricų! matricų!) grožio.



Laimei, ekonomikoje yra toks dalykas kaip konkurencija. Be to, ne visi žmonės mano, kad TV žiūrovas turi išmanyti fiziką, o omleto valgytojas turi būti susipažinęs su paukščių fiziologija. Atsiradusios naujos, daug patogesnės vartotojui programos privertė keistis ir LISREL. Programa pasipildė grafine sąsaja, o programuoti galima SIMPLIS kalba, kurioje jau nebereikia matricinių žymenų. Atskira LISREL programos dalis PRELIS yra skirta duomenims tvarkyti. Pavyzdžiui, su PRELIS galima trūkstamus duomenis pakeisti suvidurkintomis kitų respondentų duomenų reikšmėmis. Pateikiame SIMPLIS kalba užrašytą 3.3.1 pavyzdį.

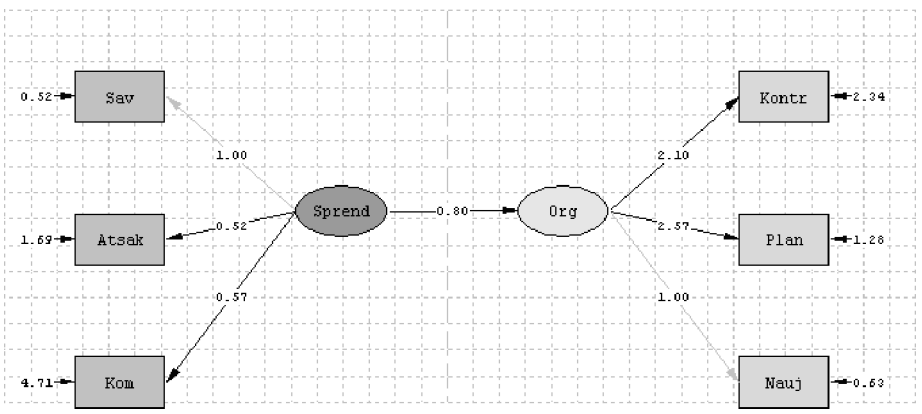
[52] LISREL – angl. *Linear Structural Relationships with Latent variables.*

[53] Karl G. Jöreskog ir Dag Sörbom – švedų statistikai.

```
SIMPLIS pavyzdys.
Lisrelio modelis - Model 1
Observed variables Kontr Plan Nauj Sav Atsak Kom
Sample size 148
Correlation Matrix
1
0.622 1
0.493 0.605 1
0.440 0.528 0.424 1
0.245 0.238 0.127 0.200 1
0.163 0.086 0.102 0.187 0.022 1 Means
3.229 11.598 5.486 5.195 6.048 8.535
Standard Deviations
2.207 2.257 1.097 1.066 1.361 2.215 Latent Variables Org Sprend
Sav = 1* Sprend
Atsak = Sprend
Kom = Sprend
Nauj = 1* Org
Kontr = Org
Plan = Org
Org = Sprend
Path Diagram
End of Problem
```

Pavyzdyje nurodėme kintamųjų koreliacijų matricą, jų vidurkius ir standartinius nuokrypius. Tą patį rezultatą gautume nurodę kovariacijų matricą. Naudojantis LISREL programa gautoji grafinė diagrama truputį skiriasi nuo AMOS programa braižomos diagramos, joje kitaip nurodytos liekamosios paklaidos (žr. 3.3.15 pav.).

Naujausios LISREL versijos numato diskrečiųjų kintamųjų modeliuose naudojimą,



3.3.15 pav. LISREL programos SEM diagrama

keičiant Pirsono koreliacijas jų diskrečiais analogais – polichorinėmis arba tetrachorinėmis koreliacijomis. Vis dėlto išliekanti tendencija – LISREL yra galingas įrankis tik labai giliai statistiką išmanančiųjų rankose, kuriems taip pat tenka pavargti, norint pasinaudoti programos teikiamomis galimybėmis. Eiliniams vartotojams LISREL programa kol kas nėra pats geriausias pasirinkimas.

R programos funkcijų biblioteka SEM. Atvirojo kodo R programa nuolatos tobulinama. Šiuo metu joje yra daugiau nei tūkstantis įvarių specializuotų funkcijų bibliotekų. Trumpai aptarsime, kaip galima atlikti SEM modelių analizę, naudojantis funkcijų biblioteka SEM. Iš pradžių nurodoma

```
library(sem)
```

Prieš atliekant SEM analizę, dar reikia nurodyti kovariacijų matricą, stebėjimų skaičių, stebimų kintamųjų vardus ir aprašyti modelį. Kaip tai daroma, paaiškinsime remdamiesi 3.3.1 pavyzdžiu[54]. Pradėkime nuo paskutinio – sudėtingiausio etapo. Kaip aprašomas modelis? Kokia sintaksė? Struktūrinis koeficientas nurodomas taip:

```
'V1 -> V2', 'name',NA,
```

Ši išraiška visiškai atitinka diagramos braižymo logiką – matome, kas kam daro įtaką. Kintamasis *V2* yra priklausomas, *V1* – nepriklausomas, o *name* yra struktūrinio koeficiento prie *V1* vardas. Struktūrinių koeficientų vardai laisvai pasirenkami. Jei koeficientas fiksuojamas (paprastai imamas lygus vienetui), tai rašome

```
'V1 -> V2',NA,1
```

Kovariacija tarp kintamųjų aprašoma taip:

```
'V1 <-> V2', 'name',NA,
```

R programos funkcijų bibliotekoje SEM paklaidų dispersijos aprašomos kaip atitinkamų kintamųjų dispersijos:

```
'V2 <-> V2', 'name',NA,
```

Jei kintamojo dispersija fiksuojama (paprastai prilyginama 1), tai rašome

```
'V2 <-> V2', NA,1,
```

Nagrinėto 3.3.1 pavyzdyje modelio aprašymo ir analizės programėlę reiktų užrašyti taip, kaip parodyta 3.3.16 paveiksle.

[54] Kaip naudotis funkcijų biblioteka SEM, išsamiau sužinoti galima surinkus *help(sem)*.

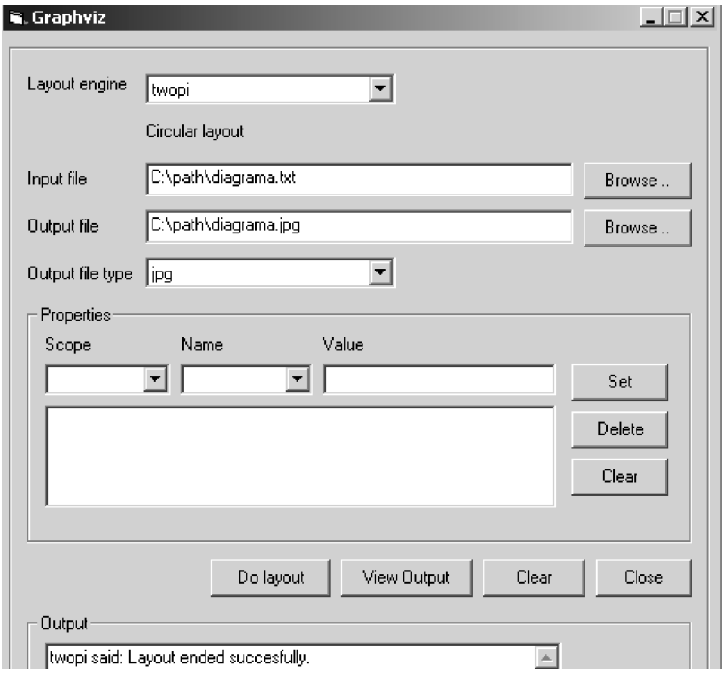
```
library(sem)
R.dlp = matrix( # kovariacijų matrica
c(1.85, 0, 0, 0,0,0,
0.07, 4.90, 0, 0,0,0,
 0.74, 0.80,4.87, 0,0,0,
0.19,0.25,1.19,1.20,0,0,
0.73,0.43,3.10,1.49,5.10,0,
 0.29,0.44, 1.04,0.50,1.27,1.14), ncol=6,byrow=T)

rownames(R.dlp) = colnames(R.dlp) = c("Atsak","Kom","Kontr","Nauj","Plan","Sav")

ram.dlp <- matrix(c(
"Sprend -> Sav", NA,1,
"Sprend->Atsak","lamx1",NA,
"Sprend -> Kom", "lamx2", NA,
"Org -> Nauj",  NA,1,
"Org -> Plan", "lamy1", NA,
"Org -> Kontr", "lamy2", NA,
"Sprend->Org","gamml",NA,
"Kontr<->Kontr","var1",NA,
"Plan<->Plan","var2",NA,
"Nauj<->Nauj","var3",NA,
"Sav<->Sav","var4",NA,
"Atsak<->Atsak","var5",NA,
"Kom<->Kom","var6",NA,
"Sprend<->Sprend","psil",NA,
"Org<->Org","phil",NA
), ncol=3, byrow=T)

sem.dlp <- sem(ram.dlp, R.dlp, N=148,rownames(R.dlp))
summary(sem.dlp)
path.diagram(sem.dlp)
```

3.3.16 pav.

Deja, R programos priemonėmis diagramos nebraižomos. Paskutinis programėlės sakinys tik sugeneruoja scenarijų, tinkamą atviro kodo grafikų braižymo programai *Graphviz*[55]. Taigi, norėdami pavaizduoti analizės rezultatus grafiškai, turime:

1) Įdiegti *Graphviz* programą.
2) Sakinio *path.diagram* vykdymo rezultatus nukopijuoti į atskirą failą (pvz., *diagrama.txt*).
3) Paleisti programos *Graphviz* modulį *twopi* ir užpildyti tris formos (žr. 3.3.17 pav.) laukus: *Input file* (diagrama.txt); *Output file* (nurodome diagramos paveikslėlio išvesties failą, pvz., *diagrama.jpg*); *Output file type* (nurodome grafinio formato tipą, pvz., jpg). Spragtelime mygtuką *Do layout*. Belieka atsidaryti failą *diagrama.jpg*.

[55] http://www.graphviz.org/.



3.3.17 pav. *Graphviz* programos langas

```
data sociol(TYPE=COV);
  _type_ = 'cov';
  input _name_ $ Atsak Kom Kontr Nauj Plan Sav;
  datalines;
  Atsak  1.85      .     .     .     .     .
  Kom    0.07   4.90    .     .     .     .
  Kontr  0.74   0.80   4.87   .     .     .
  Nauj   0.19   0.25   1.19  1.20   .     .
  Plan   0.73   0.43   3.10  1.49  5.10   .
  Sav    0.29   0.44   1.04  0.50  1.27  1.14
;
run;
proc calis cov data = sociol method = ml  nobs = 148 ;
lineqs
Sav  =              F1 + e1,
Atsak=lamx1         F1 + e2,
Kom =lamx2          F1 + e3,
Nauj =              F2     + e4,
Plan  = lamy1       F2     + e5,
Kontr= lamy2        F2     + e6,
F2   =  gam1        F1 +d1;

std
F1=phi,
e1-e6=var1-var6,
d1 =psi;
run;
```

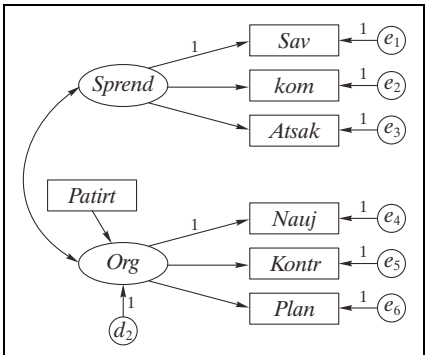
3.3.18 pav. SAS programos SEM modeliui fragmentas

Kitos programos. Programą Mplus sukūrė Linda ir Bengtas Mutenai[56]. Teigiamos šios programos ypatybės – galimybė didžiausiojo tikėtinumo metodu įvertinti praleistas duomenų reikšmes ir į modelį įtraukti diskrečius kintamuosius. Tačiau modelį tenka aprašyti sakiniais, panašiai kaip R programoje. Programa EQS yra ne mažiau patogi, nei AMOS ar LISREL programos, tik mažiau populiari. Jos pranašumas – net keletas SEM metodų, tinkančių nenormaliai pasiskirsčiusiems kintamiesiems. Gerai žinoma SAS programa taip pat turi SEM analizės procedūrą. Pateikiame 3.3.1 pavyzdžio SAS programėlę. Atkreipiame dėmesį, kad latentiniai faktoriai *Sprend* ir *Org* programoje pakeisti atitinkamai faktoriais *F1* ir *F2*.

kelio koeficientas	modelio identifikavimas	struktūrinis modelis
latentinis faktorius	modelio laisvės laipsniai	suderinamumo indeksai
liekamoji paklaida	standartizuotasis modelis	
matavimo modelis	struktūrinės lygtys	

UŽDAVINIAI

1. Užrašykite lygčių sistemą 3.3.5 paveiksle nubraižytojo modelio kelio koeficientams rasti. Visos kintamųjų kovariacijos ir vidurkiai žinomi.
2. Struktūrinėmis lygtimis užrašykite 3.3.6 paveiksle pavaizduotos diagramos MIMIC modelį.
3. Užrašykite 3.3.19 paveiksle pavaizduotą modelį matricomis.



3.3.19 pav. Alternatyvus vadybininko darbo modelis

4. Faile **stat3pa1.sav** yra tyrimo duomenys. Įvertintas respondentų pasitenkinimas gyvenimu (kintamasis *Gyvenimas*), atsižvelgiant į santykius darbe (*Darbas*), santykius šeimoje (*Seima*) ir respondento algą (*Alga*). Patikrinkite 3.3.5 paveiksle pateiktą SEM modelį.

[56] Linda Muthen, Bengt Muthen – amerikiečių statistikai.

5. Faile **st3sem1.sav** yra tyrimo duomenys. Įvertintas vadybininkų komunikabilumas (kintamasis *Kom*) bendraujant su kitais vadybininkais, sprendimų savarankiškumas (*Sav*), atsakomybės už priimtus sprendimus jutimas (*Atsak*), naujų technologijų diegimas (*Nauj*), darbo proceso planavimas (*Plan*), darbo kontroliavimas (*Kontr*) ir darbo patirtis (*Patirt*). Patikrinkite, kaip duomenims tinka 3.3.19 paveiksle pavaizduotas modelis. Modelį patobulinkite.

6. (5 uždavinio tęsinys.) Patikrinkite, ar taikant gautąjį modelį atskirai vyrams ir moterims kelio koeficientai gaunami vienodi.

PRIEDAI

Iš matricinių užrašų gausos galima susidaryti įspūdį, kad vadovėlis skirtas tik matematikams. Taip nėra. Matricos ir vektoriai kaip tik įgalina dideles lygčių sistemas užrašyti viena ir dviem eilutėmis ir taip padeda lengviau suvokti modeliuojamas priklausomybes. Prieduose primename svarbiausias sąvokas ir veiksmus su matricomis bei vektoriais. Jų nėra daug. Stengėmės vengti per didelio formalumo, todėl algebros specialistas gal ir pasiges formuluočių griežtumo. Vienas priedas skirtas daugiamačio normaliojo atsitiktinio dydžio skirstiniui.

A. Matricos

Veiksmai su matricomis. Realiųjų skaičių stačiakampę lentelę, turinčią m eilučių ir n stulpelių, vadiname $m \times n$ matmenų matrica:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Be trumpo žymėjimo $\mathbf{A}$, ši matrica dar užrašoma taip: $\mathbf{A}_{m\times n}$ arba $\mathbf{A} = (a_{ij})$; $(i = 1, \ldots, m)$, $(j = 1, \ldots n)$. Matricos žymimos didžiosiosmis raidėmis, o jų elementai – mažosiomis. Indeksai prie matricos elemento nurodo jo vietą lentelėje. Pirmosios eilutės trečiasis elementas yra žymimas simboliu a_{13}. Matricų pavyzdžiai:

$$\mathbf{B} = \begin{pmatrix} 7 & 1 \\ -3 & 2 \\ 4 & 0 \end{pmatrix}, \qquad \mathbf{C} = \begin{pmatrix} -1 & 2 & 3 \\ 0 & 4 & 15 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 1 & 5 & 1 \end{pmatrix}.$$

Sukeisdami eilutes ir stulpelius vietomis (pirmoji eilutė virsta pirmuoju stulpeliu, antroji eilutė – antruoju stulpeliu ir t. t.), gauname *transponuotą* matricą. Transponuotoji $\mathbf{A}$ matrica žymima $\mathbf{A}'$. Taigi,

$$\mathbf{A}' = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1m} & a_{2m} & \cdots & a_{mn} \end{pmatrix},$$

$$\mathbf{B}' = \begin{pmatrix} 7 & -3 & 4 \\ 1 & 2 & 0 \end{pmatrix}, \qquad \mathbf{C}' = \begin{pmatrix} -1 & 0 \\ 2 & 4 \\ 3 & 15 \end{pmatrix}, \qquad \mathbf{D}' = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 0 & 5 \\ 1 & 1 & 1 \end{pmatrix}.$$

Nesunku pastebėti, kad jei pradinės matricos matmenys yra $m \times n$, tai transponuotosios matricos matemenys yra $n \times m$.

Matricą, kurios visi elementai lygūs nuliui, vadiname *nuline matrica* ir žymėsime simboliu $\mathbf{O}$.

Matricą, kurios eilučių ir stulpelių skaičius vienodas, vadiname *kvadratine*. Kvadratinės matricos eile vadiname eilučių arba stulpelių skaičių. Kvadratinės matricos elementai $a_{11}, a_{22}, \ldots, a_{nn}$ vadinami pagrindinės įstrižainės elementais.

Pagrindinės kvadratinės matricos įstrižainės elementų suma $a_{11} + a_{22} + \cdots + a_{nn}$ vadinama matricos pėdsaku.

Kvadratinė matrica, turinti savybę $a_{ik} = a_{ki}$, $i, j = 1, \ldots, n$ vadinama *simetrine matrica*. Simetrinės matricos elementai išsidėsto simetriškai pagrindinės įstrižainės atžvilgiu. Simetrinės matricos pavyzdys:

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 5 & 3 \\ 0 & 3 & 4 \end{pmatrix}.$$

Nesunku pastebėti, kad transponuotoji simetrinė matrica sutampa su pradine matrica: $\mathbf{A}' = \mathbf{A}$.

Kvadratinę matricą, kurios visi pagrindinės įstrižainės elementai lygūs vienetui, o visi kiti elementai lygūs nuliui, vadiname vienetine. Ją žymime simboliu $\mathbf{I}$ arba $\mathbf{I}_n$ (indeksas apačioje nurodo matricos matmenis):

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Dauginant matricą ir skaičiaus, dauginami visi matricos elementai, pavyzdžiui,

$$3\mathbf{C} = 3\begin{pmatrix} -1 & 2 & 3 \\ 0 & 4 & 15 \end{pmatrix} = \begin{pmatrix} -3 & 6 & 9 \\ 0 & 12 & 45 \end{pmatrix}.$$

Dvi tokių pat matmenų matricas galima sudėti. Sudedami visi atitinkamų pozicijų elementai:

$$\begin{pmatrix} -1 & 2 & 3 \\ 0 & 4 & 15 \end{pmatrix} + \begin{pmatrix} -4 & 5 & 9 \\ 0 & 12 & 0 \end{pmatrix} = \begin{pmatrix} -5 & 7 & 12 \\ 0 & 16 & 15 \end{pmatrix}.$$

Matricų daugyba galima tik tada, kai sandaugos kairėje esančios matricos stulpelių skaičius sutampa su dešinėje esančios matricos eilučių skaičiumi, t. y. sandauga $\mathbf{AB}$ galima tik tada, jeigu $\mathbf{A} = \mathbf{A}_{m\times r}$, o $\mathbf{B} = \mathbf{B}_{r\times n}$. Matricų sandaugos $\mathbf{C} = \mathbf{AB}$ matmenys yra $m \times n$, o i-osios eilutės j-asis elementas skaičiuojamas pagal formulę

$$c_{ij} = \sum_{k=1}^{r} a_{ik}b_{kj}, \quad i = 1, \ldots, m; \; j = 1, \ldots, n.$$

Praktiškai patogiau matricų daugybą atlikti taip: mintyse pirmą kairiosios matricos eilutę pasukame (transponuojame), kad ji virstų stulpeliu. Šį stulpelį mintyse įsivaizduojame esant šalia pirmojo dešiniosios matricos stulpelio. Greta atsidūrusius skaičius sudauginame. Gautas sandaugas sudedame ir įrašome į naujosios matricos pirmosios eilutės pirmojo elemento poziciją. Analogiškai pirmąją kairiosios matricos eilutę sudauginę iš antrojo dešiniosios matricos stulpelio, gauname naujosios matricos pirmosios eilutės antrąjį elementą. Daugyba tęsiama tol, kol „išsibaigia“ dešiniosios matricos stulpeliai. Tada imame kairiosios matricos antrąją eilutę ir procedūrą tęsiame:

$$\begin{pmatrix} -1 & 2 & 3 \\ 0 & 4 & 15 \end{pmatrix}\begin{pmatrix} -4 & 5 \\ 0 & 12 \\ -2 & 3 \end{pmatrix}$$
$$= \begin{pmatrix} (-1)\cdot(-4) + 2\cdot 0 + 3\cdot(-2) & (-1)\cdot 5 + 2\cdot 12 + 3\cdot 3 \\ 0\cdot(-4) + 4\cdot 0 + 15\cdot(-2) & 0\cdot 5 + 4\cdot 12 + 15\cdot 3 \end{pmatrix} = \begin{pmatrix} -2 & 26 \\ -30 & 93 \end{pmatrix}.$$

Apskritai daugyba nėra komutatyvi $\mathbf{AB} \neq \mathbf{BA}$.

Keletas daugybos savybių:

1. $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.
2. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.
3. $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.
4. $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
5. $(\mathbf{A}')' = \mathbf{A}$.
6. $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.
7. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Determinantas. Determinantas – tai su *kvadratine* matrica susietas skaičius. Matricos $\mathbf{A}$ determinantą žymime simboliu $|\mathbf{A}|$. Tikslesnis determinanto apibrėžimas yra gana gremėzdiškas. Tarkime, kad turime kvadratinę matricą

$$\mathbf{A} = \mathbf{A}_{n\times n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Iš kiekvienos eilutės pasirenkame po vieną elementą. Visi elementai yra iš skirtingų stulpelių. Elementus sudauginame. Sandaugos ženklą lemia parinktųjų stulpelių numeriai. Tarkime, iš pirmos eilutės pasirinkome elementą a_{1j}, iš antrosios – a_{2k}, iš trečiosios – a_{3m} ir t. t. Užrašome sandaugą $a_{1j}a_{2k}a_{3m}\ldots$. Stulpelių numeriai $j, k, m, \ldots$ sudaro skaičių eilutę, kurioje yra visi skaičiai nuo 1 iki n, tik išdėstyti tam tikra tvarka. Situacija, kai skaičių eilutėje didesnis skaičius stovi prieš mažesnįjį (nebūtinai greta), vadinama inversija. Pavyzdžiui, skaičių eilutėje $2, 4, 1, 3$ yra trys inversijos ($2 > 1$, $4 > 1$, $4 > 3$). Apskaičiuojame, kiek inversijų yra eilutėje $j, k, m, \ldots$. Jeigu inversijų skaičius nelyginis, tai sandaugą $a_{1j}a_{2k}a_{3m}\ldots$ padauginame iš -1.

Tokiu būdu suradę visas įmanomas elementų sandaugas ir parinkę jų ženklus, jas sudedame. Gautasis skaičius ir vadinamas matricos determinantu.

Pirmosios eilės matricos $\mathbf{A} = (a)$ determinantu vadiname skaičių $|\mathbf{A}| = a$. Antrosios eilės matricos determinantas skaičiuojamas taip:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Trečiosios eilės kvadratinės matricos determinantas:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}.$$

N-osios eilės matricos $\mathbf{A}$ elemento a_{ij} minoru (žymime $\mathbf{M}_{ij}$) vadiname determinantą, apskaičiuotą iš matricos pašalinus i-ąją eilutę bei j-ąjį stulpelį. Matricos elemento a_{ij} adjunktu vadiname skaičių $A_{ij} = (-1)^{i+j}\mathbf{M}_{ij}$.

Determinanto savybės:

1. $|\mathbf{I}| = 1$.
2. $|\mathbf{A}'| = |\mathbf{A}|$.
3. $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.
4. Jei bent vienos matricos eilutės ar stulpelio visi elementai lygūs nuliui, tai šios matricos determinantas lygus nuliui.
5. Sukeitus matricos eilutes vietomis, determinato ženklas pasikeičia į priešingą.
6. Jei matrica turi bent dvi vienodas eilutes (stulpelius), tai šios matricos determinantas lygus nuliui.
7. Jeigu matricos kokią nors eilutę (stulpelį) padauginsime iš skaičiaus, nelygaus nuliui, ir sudėsime su kita eilute (stulpeliu), tai naujai gautos matricos determinantas bus lygus pradinės matricos determinantui.

Kvadratinė matrica, kurios determinantas lygus nuliui, vadinama išsigimusiąja matrica.

Atvirkštinė matrica. Toliau tarsime, kad $\mathbf{A}$ yra n-osios eilės kvadratinė matrica, kurios determinantas *nelygus nuliui*.

Matricą $\mathbf{A}^{-1}$ vadiname matricai $\mathbf{A}$ atvirkštine matrica, jeigu $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

Atvirkštinė matrica gali būti apskaičiuota pagal tokią formulę:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix};$$

čia A_{ij} – atitinkami adjunktai.

Tik neišsigimusios kvadratinės matricos turi atvirkštines. Jei kvadratinė matrica turi atvirkštinę, tai ji yra vienintelė.

Atvirkštinės matricos savybės:

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$. 2. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. 3. $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$.

B. Vektoriai

Realiųjų skaičių rinkinį

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ x_n \end{pmatrix}$$

vadiname n-mačiu vektoriumi (vektoriumi-stulpeliu). Vektorių galima traktuoti kaip matricą, sudarytą iš vieno stulpelio. Transponuotą vektorių $\mathbf{x}$ (vektorių-eilutę) žymime simboliu $\mathbf{x}'$:

$$\mathbf{x}' = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}' = (x_1, x_2, \cdots, x_n).$$

Pavyzdžiui,

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}' = (1, 2, 3), \qquad (1, 2, 3)' = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Aišku, kad $(\mathbf{x}')' = \mathbf{x}$. Visos kitos operacijos apibrėžiamos analogiškai matricų operacijoms. Tarkime, kad turime du n-mačius vektorius:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}.$$

Vektorių sudėtis ir daugyba iš skaičiaus apibrėžiamos taip:

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ \cdots \\ x_n + y_n \end{pmatrix}, \qquad k\mathbf{x} = \begin{pmatrix} kx_1 \\ \cdots \\ kx_n \end{pmatrix}.$$

Vektorių, kurio visos koordinatės lygios nuliui, vadiname nuliniu ir žymime raide $\mathbf{0}$ ($\mathbf{0}' = (0, 0, \ldots, 0)$).

Vektorių savybės:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$.
3. $\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x}$.
4. $k\mathbf{x} = \mathbf{x}k$.

Vektoriams tinka ir matricų sandaugos apibrėžimas:

$$\mathbf{x}'\mathbf{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n,$$

$$\mathbf{x}\mathbf{y}' = \begin{pmatrix} x_1y_1 & x_1y_2 & \ldots & x_1y_n \\ x_2y_1 & x_2y_2 & \ldots & x_2y_n \\ \cdots & \cdots & \cdots & \cdots \\ x_ny_1 & x_ny_2 & \ldots & x_ny_n \end{pmatrix}.$$

Vektoriai $\mathbf{x}$ ir $\mathbf{y}$ vadinami ortogonaliaisiais, jeigu $\mathbf{x}'\mathbf{y} = 0$.

Tarkime, kad turime $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ k-mačių vektorių. Jeigu galima rasti tokius skaičius $\alpha_1, \alpha_2, \ldots, \alpha_N$, kad bent vienas jų būtų nelygus nuliui, o suma $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_N\mathbf{x}_N = \mathbf{0}$, tai sakome, kad vektoriai $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ tiesiškai priklausomi.

Jeigu $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_N\mathbf{x}_N = \mathbf{0}$ vieninteliu atveju, kai $\alpha_1 = \alpha_2 = \cdots = \alpha_N = 0$, tai sakome, kad vektoriai $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ tiesiškai nepriklausomi.

C. Atsitiktiniai vektoriai ir atsitiktinės matricos

Atsitiktiniais vektoriais ir atsitiktinėmis matricomis vadiname vektorius ir matricas, kurių elementai yra atsitiktiniai dydžiai. Tarkime, kad $\mathbf{X}' = (X_1, X_2, \dots, X_n)$, $\mathbf{E}X_j = \mu_j$, $j = 1, 2, \dots, n$. Atsitiktinio vektoriaus vidurkis – tai vektorius sudarytas iš elementų vidurkių, t. y.

$$\mathbf{EX} = \mathbf{E}\begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_n \end{pmatrix} = \begin{pmatrix} \mathbf{E}X_1 \\ \mathbf{E}X_2 \\ \dots \\ \mathbf{E}X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix}.$$

Atsitiktiniam vektoriui apibrėžiama ir jo elementų kovariacijų matrica. Neretai ši matrica žymima $\mathrm{Var}(\mathbf{X})$, $\mathbb{C}\mathrm{ov}(\mathbf{X})$ arba $\mathbf{\Sigma}$. Pažymėkime vidurkių vektorių simboliu $\boldsymbol{\mu}$. Tada

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\
&= \begin{pmatrix} \mathbf{E}(X_1 - \mu_1)(X_1 - \mu_1) & \mathbf{E}(X_1 - \mu_1)(X_2 - \mu_2) & \dots & \mathbf{E}(X_1 - \mu_1)(X_n - \mu_n) \\ \mathbf{E}(X_2 - \mu_2)(X_1 - \mu_1) & \mathbf{E}(X_2 - \mu_2)(X_2 - \mu_2) & \dots & \mathbf{E}(X_2 - \mu_2)(X_n - \mu_n) \\ \dots & \dots & \dots & \dots \\ \mathbf{E}(X_n - \mu_n)(X_1 - \mu_1) & \mathbf{E}(X_n - \mu_n)(X_2 - \mu_2) & \dots & \mathbf{E}(X_n - \mu_n)(X_n - \mu_n) \end{pmatrix} \\
&= \begin{pmatrix} \mathbb{C}\mathrm{ov}(X_1, X_1) & \mathbb{C}\mathrm{ov}(X_1, X_2) & \dots & \mathbb{C}\mathrm{ov}(X_1, X_n) \\ \mathbb{C}\mathrm{ov}(X_2, X_1) & \mathbb{C}\mathrm{ov}(X_2, X_2) & \dots & \mathbb{C}\mathrm{ov}(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \mathbb{C}\mathrm{ov}(X_n, X_1) & \mathbb{C}\mathrm{ov}(X_n, X_2) & \dots & \mathbb{C}\mathrm{ov}(X_n, X_n) \end{pmatrix}.
\end{aligned}$$

Tarkime, kad $\mathbf{a}$ ir $\mathbf{b}$ yra n-mačiai neatsitiktiniai vektoriai, o $\mathbf{A}$ ir $\mathbf{B}$ neatsitiktinės matricos. Tada

1. $\mathbf{Ea}'\mathbf{X} = \mathbf{a}'\mathbf{EX}$.
2. $\mathbf{Ea}'\mathbf{Xb} = \mathbf{a}'\mathbf{EXb}$.
3. $\mathbf{EAXB} = \mathbf{AEXB}$.
4. $\mathbb{C}\mathrm{ov}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{X}) = \mathbf{a}'\mathbf{\Sigma}\mathbf{b}$.
5. $\mathbb{C}\mathrm{ov}(\mathbf{AXB}) = \mathbf{A\Sigma B}'$.

D. Daugiamatis normalusis skirstinys

Tarkime, kad turime atsitiktinį vektorių $\mathbf{X}' = (X_1, X_2, \dots, X_n)$ su vidurkių vektoriumi $\boldsymbol{\mu}$ ir kovariacijų matrica $\mathbf{\Sigma}$. Sakome, kad $\mathbf{X}$ turi normalųjį skirstinį, jeigu jo tankis yra

$$\varphi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\big\{-(1/2)(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\big\};$$

čia $\mathbf{x} = (x_1, x_2, \dots, x_n)$ gali įgyti reikšmes iš $\mathbb{R}^n$. Rašysime $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Daugiamatis normalusis vektorius turi tokias savybes:

1. Kiekviena vektoriaus koordinatė turi vienamatį normalųjį skirstinį.
2. Jeigu $\mathbf{a}$ yra neatsitiktinis n-matis vektorius, tai $\mathbf{a}'\mathbf{X} \sim \mathcal{N}(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\mathbf{\Sigma}\mathbf{a})$.
3. Jeigu $\mathbf{A}$ yra neatsitiktinė matrica, tai $\mathbf{AX} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A\Sigma A}')$.

Vartojamų terminų anglų–lietuvių kalbų žodynėlis

Adaptive Gaussian Quadrature (*AGQ*) *method* – Gauso adaptyvusis kvadratūrinis metodas
Adjusted Goodness-of-Fit Index (*AGFI*) – pataisytasis modelio suderinamumo indeksas
Analysis of Covariance (*ANCOVA*) – kovariančių analizė
Asymptotically Distribution free (*ADF*) *index* – asimptotiškai nepriklausomas nuo skirstinio indeksas
Best Linear Unbiased Estimator (*BLUE*) – geriausias tiesinis nepaslinktasis įvertinys
Causal Modeling (*CM*) – priežastinis modeliavimas
Comparative Fit Index (*CFI*) — sąlyginis suderinamumo indeksas
Confirmatory Factor Analysis (*CFA*) – patvirtinančioji faktorinė analizė
Correlated Trait (*CT*) *model* – koreliuotųjų požymių modelis
Correlated Trait – Correlated Uniquenness (*CTCU*) *model* – koreliuotųjų požymių – koreliuotųjų liekamųjų paklaidų modelis
Covariance Structure Analysis (*CSA*) – kovariacijų struktūrinė analizė
cross-level interactions – lygmenų sąveika
design matrix – struktūros (plano) matrica
deviation, scaled deviation – deviacija, normuotoji deviacija
endogenous variable – vidinis (endogeninis) kintamasis
estimate – įvertis
estimator – įvertinys
exogenous variable – išorinis (egzogeninis) kintamasis
Expexted Cross-Validation Index (*ECVI*) – tikėtino kryžminio pagrįstumo indeksas
Explanatory Factor Analysis (*EFA*) – tiriančioji faktorinė analizė
fixed effect – fiksuotas poveikis
Generalized Estimating Equations (*GEE*) – apibendrintosios įvertinių lygtys
Generalized Linear Model (*GLM*) – apibendrintasis tiesinis modelis
Hierarchical Linear Modeling (*HLM*) – hierarchinis tiesinis modeliavimas
Incremental Fit Index (*IFI*) – prieaugių suderinamumo indeksas
intercept – laisvasis narys, postūmio parametras
Intraclass Correlation Coeficient (*ICC*) – tarpklasinės koreliacijos koeficientas
Linear Normal Model (*LNM*) – tiesinis Gauso modelis
link function – jungties funkcija
longitudinal data – ilgalaikiai duomenys
Maximum Likelihood (*ML*) *method* (*function*) – didžiausiojo tikėtinumo metodas (funkcija)
Marginal Quasi-Likelihood (*MQL*) *method* – ribinis kvazitikėtinumo metodas
Modification Index (*MI*) — modifikacijos indeksas
Multiple Indicators and Multiple Causes (*MIMIC*) *model* – daugiapriežastinis SEM modelis
Multitrait-Multimethod Model (*MTMM*) – daugelio požymių ir daugelio metodų SEM modelis
Nonnormed Fit Index (*NNFI*) – nenormuotas suderinamumo indeksas
Normed Chi-square (*NC*) – normuotas chi kvadratas
null model likelihood ratio test – nulinio modelio tikėtinumo santykių kriterijus
overdispersion – perteklinė dispersija
panel data – paneliniai duomenys
path analysis – kelių analizė
Penalized Quasi-Likelihood (*PQL*) *method* – svorinis kvazitikėtinumo metodas
predictor – prediktorius
Random Coefficients (*RC*) – atsitiktiniai koeficientai
random effect – atsitiktinis poveikis
Restricted Maximum Likelihood (*REML*) *method* – apribotojo didžiausiojo tikėtinumo metodas
Root Mean Square Error of Approximation (*RMSEA*) – kvadratinė šaknis iš vidutinės kvadratinės paklaidos
Root Mean Square Residual (*RMSR*) – kvadratinė šaknis iš standartizuotosios vidutinės liekanos
saturated model – pilnasis (perteklinis, prisotintasis) modelis
shrinkage estimator – sutraukiantysis įvertinys
Structural Equation Modeling (*SEM*) – struktūrinių lygčių modeliavimas
Tucker-Lewis Index (*TLI*) – Takerio–Liuiso indeksas

Dalykinė rodyklė

agregavimas 69, 70
analizė
  kovariančių 10, 25, 32, 97
  patvirtinančioji faktorinė 168
  tiesinės regresijos 9, 27
  tiriančioji faktorinė 16
  vienfaktorė dispersinė 13
ANCOVA 5, 25, 40

darinys
  įvertinių 75, 88
deagregavimas 69, 70
deviacija 57, 58
diagrama
  kelių analizės 156
duomenys
  hierarchiniai 68
  paneliniai 123

eksponentinių skirstinių šeima 47

faktoriniai svoriai 172, 177, 180
faktorius
  latentinis 168

indeksai
  informaciniai (*AIC*, *AICC*, *CAIS*, *BIC*) 90
  suderinamumo (*RMSEA*, *NC*, *SRMR*, *AGFI*, *IFI*, *TLI*, *CFI*) 181–184
įvertinio patikimumas 76
įvertinys
  apibendrintas mažiausiųjų kvadratų 20
  BLUE 17
  mažiausiųjų kvadratų 17
  sutraukiantysis (svertinis) 75

jungties funkcija
  atvirkštinė 49
  identiška 49
  log 49
  logit 49

kintamasis
  atsitiktinio poveikio 85, 96
  centruotasis 79
  endogeninis 154
  egzogeninis 154, 193
  fiksuotojo poveikio 85, 96
  stebimasis 154, 193
  tarpinis 154
koeficientas
  kelio 155
  R kvadratu 20
  tarpklasinės koreliacijos 74
kontrastų matrica 24
kriterijus
  modelių tikėtinumo santykių 96

matrica kovariacijų 11, 238
metodas
  apribotojo didžiausiojo tikėtinumo 88
  didžiausiojo tikėtinumo 20
  mažiausiųjų kvadratų 17
modelio
  identifikavimas 158, 172, 200
  lygmuo 68
  matricinė formuluotė 85
modelis
  apibendrintasis tiesinis 9, 45
  atsitiktinio postūmio ir posvyrio 81
  augimo 125
  bendrasis HLM 76
  CFA 176
  CTCU 204
  daugialygis 118
  hierarchinis tiesinis 68, 76
  Gauso 10
  jungtinis 77
  LNM 11
  logit 50
  MIMIC 203
  MTMM 203
  matricinis HLM 85
  mišrusis 136
  nerekursyvusis 157
  nulinis 71
  pastovaus posvyrio 80
  perteklinės identifikacijos 159
  prežastinės analizės 151
  probit 50
  redukuotasis 91, 95
  rekursyvusis 157

SEM 206
statistinis 10
su antrojo lygmens intervaliniu kintamuoju 82, 101
su antrojo lygmens kategoriniu kintamuoju 83, 103
tiesinis 11, 45
vienfaktorės dispersinės analizės 13

paklaida
antrojo lygmens 72
pirmojo lygmens 72
liekamoji (modelio) 10, 154
parametras
atsitiktinio poveikio 72, 79
fiksuotojo poveikio 72, 79
poveikis
netiesioginis 164
tiesioginis 164
prediktorius 46
procedūra
GENLIN 60
MIXED 96, 130
PROC GENMOD 61
PROC GLM 41
PROC MIXED 109
PROC GLIMMIX 144
UNIANOVA 33

R programos funkcija
glm 60
lm 40
lme 113, 131
lmer 145
sem 227
regresija
hierarchinė logistinė 138
logistinė 49
neigiamai binominė 51, 54
Puasono 50
tiesinė 12, 13

statistinė programa
AMOS 224
LISREL 224, 225
SPSS 33, 60, 96, 128, 146
SAS 40, 60, 109, 144
R 40, 60, 113, 145
struktūrinės lygtys 192

Literatūra

1. B. M. Byrne. *Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming*. 2nd edn. Lawrence Erlbaum Associates, Publishers, 2001.
2. V. Čekanavičius, G. Murauskas, *Statistika ir jos taikymai*. I. Vilnius: TEV, 2000.
3. V. Čekanavičius, G. Murauskas, *Statistika ir jos taikymai*. II. Vilnius: TEV, 2002.
4. L. F. Garcia, M. A. Ruiz, and F. J. Abad. Factor structure of the Spanish WAIS-III. *Psicothema* **15**, 155–160, 2003.
5. R. H. Hoyle (Ed.). *Strustural Equation Modeling*. Sage, 1995.
6. R. B. Kline. *Principles and Practice of Structural Equation Modeling*. 2nd edn. Guiford Press, 2005.
7. S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models*, 2nd edn. Sage, 2002.
8. R. E. Schumacker and R. G. Lomax. *A Beginner's Guide to Structural Equation Modeling*. 2nd edn. Lawrence Erlbaum Associates, Publishers, 2004.

Vydas Čekanavičius, Gediminas Murauskas

STATISTIKA IR JOS TAIKYMAI. III

2011 08 19. 15 sp. l.
Leidykla TEV, Akademijos g. 4, LT-08412 Vilnius
Spausdino AB „Spauda“
Laisvės pr. 60, LT-05120 Vilnius
www.spauda.com