HLM modelių analizė R programa

Knygoje "Statistika ir jos taikymai III" 2.11 skyrelyje aprašyta HLM pavyzdžių analizė naudojantis R programos funkcijų *nlme* paketu. Čia aprašysime tų pačių pavyzdžių analizę naudojantis R programos funkcijų *lme4* paketu.

Priminsime, kad pavyzdžiuose naudojami mokinių matematinių pasiekimų tyrimo (TIMSS) duomenys. Taip pat vartojami tokie žymenys:

SES_{ij} – i-ojo mokinio socialinis ir ekonominis statusas j-ojoje mokykloje;
$CSES_{ij}$ – centruotasis SES_{ij};
$MSES_j$ – j-osios mokyklos socialinis ir ekonominio statuso vidurkis;
MAT – matematikos testo rezultatas;
$IDMOK$ – mokyklos indeksas;
VK – mokyklos tipas (0 – Vilniaus mokykla, 1 – kaimo mokykla).

Socialinio ir ekonominio statuso įvertinimas (SES) yra daugelio požymių konstrukcija. Kiekvieno mokinio $CSES$ gaunamas iš jo SES atėmus visų tos mokyklos SES vidurkį. Detalesnis kintamųjų aprašymas pateikiamas knygoje.

Norint HLM analizę atlikti R programa, galima naudotis funkcijų *lme4* paketu. Standartinei jungtinio HLM modelio analizei naudojama funkcija *lmer*, kurios išraiška yra tokia:

lmer(formulė, duomenų rinkinio vardas,metodas).

Sintaksė labai panaši kaip ir knygoje aprašytos funkcijos *lme*, tačiau, turint jungtinį modelį, ji „natūralesnė".

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 norodo, kokiam antrojo lygmens elementui priklauso duomenys. Duomenų rinkinio vardas – *datafile*. Tada *lmer* išraiška yra tokia:

lmer(Y ∼1+X+W+V+(1+X|ID), data=datafile)

Funkcijos *summary, anova, print, plot* naudojamos analizės rezultatų išvedimui. Standartinei HLM analizei užtenka funkcijos *summary*.

Toliau aprašysime, kaip atlikti knygoje pateikiamų modelių analizę naudojantis *lme* ir *lmer* funkcijomis.

1. ***Besąlyginio modelio*** jungtinė lygtis:

$$MAT = \gamma_{00} + u_0 + e.$$

R programa (naudojant funkciją *lme*):

```
model.0<-lme(MAT ~ 1, data=dat, random=~ 1|IDMOK)
summary(model.0)
```

R programa (naudojant funkciją *lmer*):

```
model.00<-lmer(MAT ~ (1|IDMOK), data=dat)
summary(model.00)
```

2. ***Atsitiktinio postūmio ir posvyrio modelis.*** Šio modelio jungtinė lygtis:

$$MAT = \gamma_{00} + \gamma_{10}CSES + [u_1 CSES + u_0 + e].$$

R programa (naudojant *lme* funkciją):

```
model.2<-lme(MAT ~ 1+CSES, data=dat, random=~1+CSES|IDMOK)
summary(model.2)
```

R programa (naudojant funkciją *lmer*):

```
model.22<-lmer(MAT ~ CSES+(CSES|IDMOK), data=dat)
summary(model.22)
```

arba

```
model.22<-lmer(MAT ~ 1+CSES+(1+CSES|IDMOK), data=dat)
summary(model.22)
```

3. ***Modelio su antrojo lygmens kategoriniu kintamuoju*** jungtinė lygtis:

$$\begin{aligned} MAT &= \gamma_{00} + \gamma_{01}MSES + \gamma_{10}CSES + \gamma_{02}VK + \gamma_{12}VK * CSES \\ &\quad + [u_1 CSES + u_0 + e]. \end{aligned}$$

R programa (naudojant funkciją *lme*):

```
model.4<-lme(MAT ~ 1+CSES+MSES+VK+VK*CSES,
data=dat, random=~1+CSES|IDMOK)
summary(model.4)
```

R programa (naudojant funkciją *lmer*):

```
model.44<-lmer(MAT ~1+CSES+MSES+VK+VK*CSES
+(1+CSES|IDMOK),
data=dat)
summary(model.44)
```

Rezultatai. Iš pradžių pateikiame besąlyginio modelio aprašymą ir analizės rezultatus.

Kuo *besąlyginis* (arba nulinis) modelis 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.

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 aprašomas mokyklos lygmuo:

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. Pažymėkime i-ojo mokinio iš j-osios mokyklos matematikos testo rezultatus simboliu Y_{ij}. 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:

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

Antrasis (mokyklos) lygmuo:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad j = 1, \ldots, J. \tag{2}$$

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 visada regresiniuose modeliuose, tariama, kad visos atsitiktinės paklaidos e_{ij} nepriklausomos ir normaliai pasiskirsčiusios su ta pačia dispersija 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 fiksuotojo poveikio parametru,
> τ_{00} ir σ^2 vadinami atsitiktinio poveikio parametrais.

Nagrinėjamo pavyzdžio parametrai yra tokie:

- β_{0j} yra j-osios mokyklos rezultatų vidurkis;
- e_{ij} – kiekvieno mokinio rezultatų skirtumai nuo mokyklos vidurkio;
- kuo didesnė dispersijos σ^2 reikšmė, 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) už kitų mokyklų;
- kuo didesnė dispersijos τ_{00} reikšmė, tuo mokyklų rezultatai labiau skiriasi.

Besąlyginį modelį galima aprašyti ir viena jungtine lygtimi, įstačius (1) išraišką į (2) formulę:

$$Y_{ij} = \gamma_{00} + [u_{0j} + e_{ij}]. \tag{3}$$

Atkreipiame dėmesį, kad jungtinio besąlyginio (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 uždavinio sprendimui naudojamos programos. Pavyzdžiui, naudojant analizei R programos *lmer* funkciją, programą lengviau parašyti jungtinio modelio analizei.

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

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:

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

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).

Toliau aprašysime besąlyginio modelio analizės, naudojantis funkcijomis lme ir lmer, rezultatus (žr. 1 pav.).

Kaip matome, 1 pav. rezultatai, gauti naudojantis abiem funkcijomis yra identiški.

1) Vidutinio visų mokinių matematikos testo γ_{00} įvertis $\hat{\gamma}_{00} = 499{,}47$.

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 galima buvo tikėtis, kad visi mokiniai – absoliutūs nemokšos), daug svarbesnis to vidurkio skaitinis įvertis.

3) Mokinių (pirmojo lygmens elementų) rezultatų skirtumus apibūdinančios dispersijos σ^2 įvertis $\hat{\sigma}^2 = 5117{,}19$.

4) Mokyklų (antrojo lygmens elementų) skirtumus apibūdinančios dispersijos τ_{00} įvertis $\hat{\tau}_{00} = 2356{,}11$.

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. Kartu tai rodiklis, kad 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[1], kuris parodo, kaip stipriai

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

```
> model.0<-lme(MAT~1,data=dat,random=~1|IDMOK)
> summary(model.0)
Linear mixed-effects model fit by REML
 Data: dat
      AIC      BIC    logLik
  6421.181 6434.154 -3207.590
Random effects:
 Formula: ~1 | IDMOK
        (Intercept) Residual
StdDev:    48.53975 71.53455

Fixed effects: MAT ~ 1
               Value      Std.Error  DF  t-value     p-value
(Intercept) 499.4717  9.876646 532 50.57098       0

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.13763014 -0.68631097  0.01061007  0.60462241
3.05751149

Number of Observations: 559
Number of Groups: 27
> model.00<-lmer(MAT~1+(1|IDMOK),data=dat)
> summary(model.00)
 Linear mixed-effects model fit by REML
 Formula: MAT ~ 1 + (1 | IDMOK)
    Data: dat
      AIC      BIC    logLik MLdeviance REMLdeviance
  6419.181 6427.833 -3207.590   6421.598     6415.181
 Random effects:
  Groups   Name        Variance Std.Dev.
  IDMOK    (Intercept) 2356.1   48.540
  Residual             5117.2   71.535
 number of obs: 559, groups: IDMOK, 27

 Fixed effects:
              Estimate  Std. Error t value
 (Intercept) 499.4717    9.8766  50.571
```

1 pav. Rezultatai, gauti naudojant *lme* ir *lmer* funkcijas.

skiriasi grupių (mokyklų) rezultatai, palyginti su rezultatais grupėse. Tikslesnis ICC apibrėžimas, 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 spėti, 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 atskleidžia koreliaciją tarp dviejų tos pačios mokyklos mokinių rezultatų, 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 yra požymis, kad verta tirti HLM modelius.

Pakartosime svarbiausią 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 tai ženklas, 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 – tai ženklas, kad vertėtų paieškoti kintamųjų, lemiančių mokinių ir mokyklų skirtumus.

Taigi toliau aptarsime *atsitiktinio postūmio ir posvyrio modelį* ir *modelį su antrojo lygmens kategoriniu kintamuoju.* [2]

Atsitiktinio postūmio ir posvyrio modelio analizė. Tarkime, 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:

$$MAT = \gamma_{00} + \gamma_{10} CSES + [u_1 CSES + u_0 + e]. \tag{4}$$

Fiksuotasis kintamasis yra $CSES$, atsitiktiniai kintamieji – postūmis ir $CSES$. Primename, kad $\mathbf{D}(u_0) = \tau_{00}$, $\mathbf{D}(u_1) = \tau_{11}$, $\mathbb{C}\mathrm{ov}\,(u_0, u_1) = \mathbb{C}\mathrm{ov}\,(\beta_1, \beta_2) = \tau_{10} = \tau_{01}$, $\mathbf{D}e = \sigma^2$.

Gauti rezultatai pateikti 2 pav.

Palyginti su besąlyginiu modeliu, visi informaciniai indeksai sumažėjo, t.y. naujasis modelis geriau tinka turimiems duomenims. Postūmio įvertis pasikeitė labai mažai, t. y. $\hat{\gamma}_{00} = 502{,}29$. Fiksuotojo 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.

Atsitiktinės paklaidos dispersijos įvertis $\hat{\sigma}^2 = 4294,72$. 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 $\hat{\sigma}^2$ palyginti su ankstesniu modeliu, sumažėjimas parodo, ar labai naudinga buvo įtraukti naują kintamąjį.

Besąlyginio modelio mokinio lygmens paklaidos įvertis buvo $\hat{\sigma}^2 = 5117{,}19$, todėl įvertis sumažėjo

$$\frac{5117{,}19 - 4294{,}72}{5117{,}19} = 0{,}16.$$

[2] Žinomos ne kievieno mokinio CSES ir VK reikšmės, todėl analizei naudojama mažesnė duomenų aibė.

```
> model.2<-lme(MAT~1+CSES,data=dat,random=~1+CSES|IDMOK)
> summary(model.2)
Linear mixed-effects model fit by REML
 Data: dat
      AIC      BIC    logLik
  6157.056 6182.816 -3072.528
Random effects:
 Formula: ~1 + CSES | IDMOK
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 49.064947 (Intr)
CSES         2.980567 0.321
Residual    65.534094
Fixed effects: MAT ~ 1 + CSES
                Value Std.Error  DF  t-value      p-value
(Intercept) 502.2935  9.906152 515 50.70521       0
CSES          7.9256  1.220360 515  6.49452       0
> model.22<-lmer(MAT~1+CSES+(1+CSES|IDMOK),data=dat)
> summary(model.22)
Linear mixed-effects model fit by REML
Formula: MAT ~ 1 + CSES + (1 + CSES | IDMOK)
         Data: dat
     AIC      BIC    logLik MLdeviance REMLdeviance
 6155.056 6176.541 -3072.528   6153.691     6145.056
Random effects:
 Groups   Name             Variance  Std.Dev. Corr
 IDMOK    (Intercept)     2407.3460 49.0647
          CSES               8.8838  2.9806  0.321
 Residual                 4294.7186 65.5341
number of obs: 543, groups: IDMOK, 27
Fixed effects:
            Estimate Std. Error t value
(Intercept) 502.2935    9.9061  50.705
CSES          7.9257    1.2204   6.495

Correlation of Fixed Effects:
     (Intr)
CSES 0.145
```

2 pav. *lme* ir *lmer* funkcijos atsitiktinio postūmio ir posvyrio modeliui.

Taigi galima sakyti, kad įtraukus į modelį $CSES$ mokinio lygmens atsitiktinės paklaidos dispersija sumažėjo 16 %. Dažniausiai tai interpretuojama taip: palyginti su besąlyginiu modeliu, mokinių nepaaiškintų rezultatų skirtumų liko 16 % mažiau.

Analogiškai interpretuojamas ir $\hat{\tau}_{00} = 2407{,}346$ 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 kintamieji 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).

Modelio su antrojo lygmens kategoriniu kintamuoju analizė. Įtrauksime į modelį antrojo lygmens pseudokintamąjį VK, nurodantį, kad ta mokykla yra 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:

$$\begin{aligned} MAT \quad = \quad & \gamma_{00} + \gamma_{01} MSES + \gamma_{10} CSES + \gamma_{02} VK + \gamma_{12} VK \cdot CSES \\ & + [u_1 CSES + u_0 + e]. \end{aligned}$$

Fiksuotieji kintamieji yra $CSES$, $MSES$, VK ir $VK \cdot CSES$, atsitiktiniai kintamieji – postūmis ir $CSES$. Kintamasis VK kategorinis.

Gauti rezultatai pateikti 3 ir 4 pav. Palyginti su prieš tai buvusiu modeliu, informaciniai indeksai truputį sumažėjo[3]. Pavyzdžiui, $AIC = 6109{,}135 (6107{,}135)$. Fiksuoto poveikio parametrų įverčiai leidžia sudaryti testo rezultatų priklausomybės lygtis Vilniaus ($VK = 0$) ir kaimo ($VK = 1$) mokykloms. Paaiškinsime, kaip tai daroma. Įstatę koeficientų įverčius į jungtinę lygtį, gauname

$$MAT = -47{,}23 + 23{,}93 MSES + 7{,}27 CSES - 47{,}93 VK + 2{,}07 VK \cdot CSES.$$

Kai $VK = 1$, tai kaimo mokyklų mokinių matematikos rezultatus aprašo tokią lygtis:

$$MAT_{\text{Kaimas}} = -0{,}69 + 23{,}93 MSES + 9{,}34 CSES.$$

Iš jos matyti, kad kiekvienas mokinio SES balas, viršijantis mokyklos vidurkį, prie mokinio rezultato prideda 9,34 balo. Tuo tarpu $MSES$ padidėjus vienu balu, MAT

[3] Taikant funkcijas *lmer* ir *lme* informacinių indeksų reikšmės šiek tiek skiriasi.

```
model.4<-lme(MAT~1+CSES+MSES+VK+VK*CSES,data=dat,
random=~1+CSES|IDMOK,method='REML')
> 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
```

3 pav. *lme* funkcija modeliui su kategoriniu kintamuoju

```
model.44<-lmer(MAT~1+CSES+MSES+VK+VK*CSES+
(1+CSES|IDMOK),data=dat)
> summary(model.44)

Linear mixed-effects model fit by REML
Formula: MAT ~ 1 + CSES + MSES + VK + VK * CSES + (1 +
CSES | IDMOK)
      Data: dat
    AIC     BIC        logLik  MLdeviance  REMLdeviance
 6107.135 6141.512 -3045.567   6112.746     6091.135
Random effects:
 Groups   Name         Variance Std.Dev. Corr
 IDMOK    (Intercept)  381.218 19.5248
       CSES              10.278  3.2059   0.146
 Residual               4293.987 65.5285
number of obs: 543, groups: IDMOK, 27

Fixed effects:
             Estimate Std. Error   t value
(Intercept)   47.2330   60.8374    0.7764
CSES           7.2703    1.5584    4.6653
MSES          23.9320    3.3168    7.2154
VK           -47.9295   17.5049   -2.7381
CSES:VK        2.0715    2.5935    0.7987

Correlation of Fixed Effects:
            (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
```

4 pav. *lmer* funkcija modeliui su kategoriniu kintamuoju

išauga iki 23,93 balo. Analogiškai Vilniaus mokyklų ($VK = 0$) mokinių pasiekimus aprašo tokia lygtis:

$$MAT_{\text{Vilnius}} = 47{,}23 + 23{,}93MSES + 7{,}27CSES.$$

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 nagrinėjant Vilniaus mokyklų mokinių rezultatus, mokyklos „priedas“ yra 47,233 balo. Neatmesta hipotezė $H_0 : \gamma_{12} = 0$ rodo, kad mokyklos vietos ir socialinio ir ekonominio mokinio statuso sąveika yra statistiškai nereikšminga. Taigi nors socialinio 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.

Taigi, šio pavyzdžio modelis geriau
tinka duomenims nei ankstesnieji modeliai, tačiau daug neatmestų
hipotezių dėl parametrų lygybės nuliui reikalauja papildomo tyrimo.

Kaip dar lyginami modeliai?

Dviejų modelių palyginimas. Funkcija *Anova* pateikia tris charakteristikas, leidžiančias palyginti du modelius. Lyginamos *AIC*, *BIC* reikšmės ir tikrinama hipotezė apie dviejų modelių tikėtinumų santykių testo p reikšmę[4]. Dviejų HLM modelių palyginimui reikia naudoti didžiausiojo 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)
```

```
model.22B=update(model.2, method="ML")
model.44B=update(model.4,method="ML")
anova(model.22B,model.44B)
```

Palyginę *atsitiktinio postūmio ir posvyrio modelį* ir *modelį su antrojo lygmens kategoriniu kintamuoju* rezultatus, matome, kad p reikšmė maža ($p < 0{,}05$) (tik „kosmetiniai“ rezultatų skirtumai naudojant skirtingas funkcijas), žr, 5 pav. Gautas rezultatas leidžia spėti, kad paskutinis modelis geresnis.

Palyginimas su tiesinės regresijos modeliu. Dabar remdamiesi grafiku, parodysim, kad geriau taikyti jungtinį HLM modelį nei paprasčiausią tiesinę regresiją kiekvienai grupei. Toliau užrašyta matematikos pasiekimų HLM modelio su atsitiktiniais postūmio ir posvyrio parametrais ir regresijos kiekvienai mokyklai atskirai modelių analizės R programa.

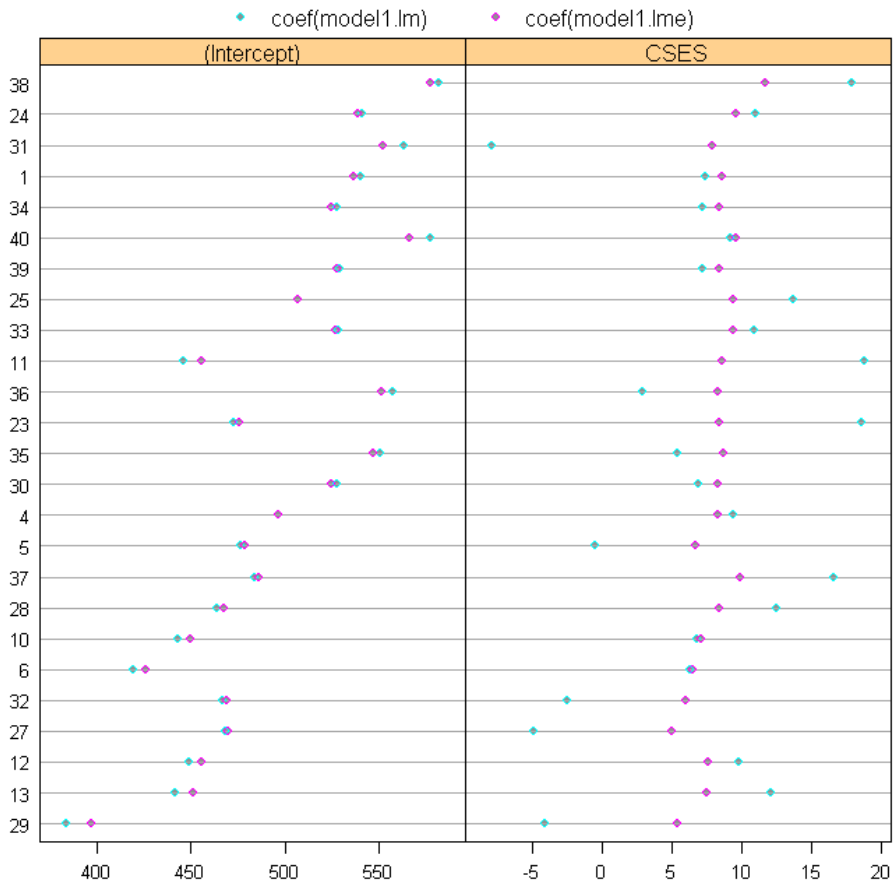
[4]Modelių informacinių indeksų reikšmės priklauso nuo to, koks įverčių skaičiavimo metodas taikomas.

```
> model.2B=update(model.2, method='ML')
> model.4B=update(model.4,method='ML')
> anova(model.2B,model.4B)
          Model  df     AIC     BIC          logLik   Test  L.Ratio    p-value
model.2B      1   6 6165.651 6191.434 -3076.826
model.4B      2   9 6130.467 6169.141 -3056.233 1 vs 2 41.18479 <.0001

> #
> #  Modelių lyginimas lmer
> #
> model.22B=update(model.22, method='ML')
> model.44B=update(model.44,method='ML')
> anova(model.22B,model.44B)

Data: dat
Models:
model.22B: MAT ~ 1 + CSES + (1 + CSES | IDMOK)
model.44B: MAT ~ 1 + CSES + MSES + VK + VK * CSES + (1 + CSES |
IDMOK)
           Df    AIC    BIC    logLik  Chisq  Chi Df  Pr(>Chisq)
model.22B   5  6163.7  6185.1 -3076.8
model.44B   8  6128.5  6162.8 -3056.2  41.185     3    5.975e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5 pav. Programa modelių lyginimui.



6 pav. Palyginimas su tiesine regresija.

```
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 įverčiai pateikiami 6 pav. Matome, kad HLM parametrų įverčių yra stabilesni, nes juos galima uždengti siauresne juostele, t. y. jie mažiau išsibarstę.