Математички факултет

Универзитет у Београду

Линеарни статистички модели

Хијерархијски линеарни модели

Аутори:

Кристина Матовић 180/2014

Марија Костић 286/2014

Ана Антић 113/2015

Професор:

Бојана Милошевић

Асистент:

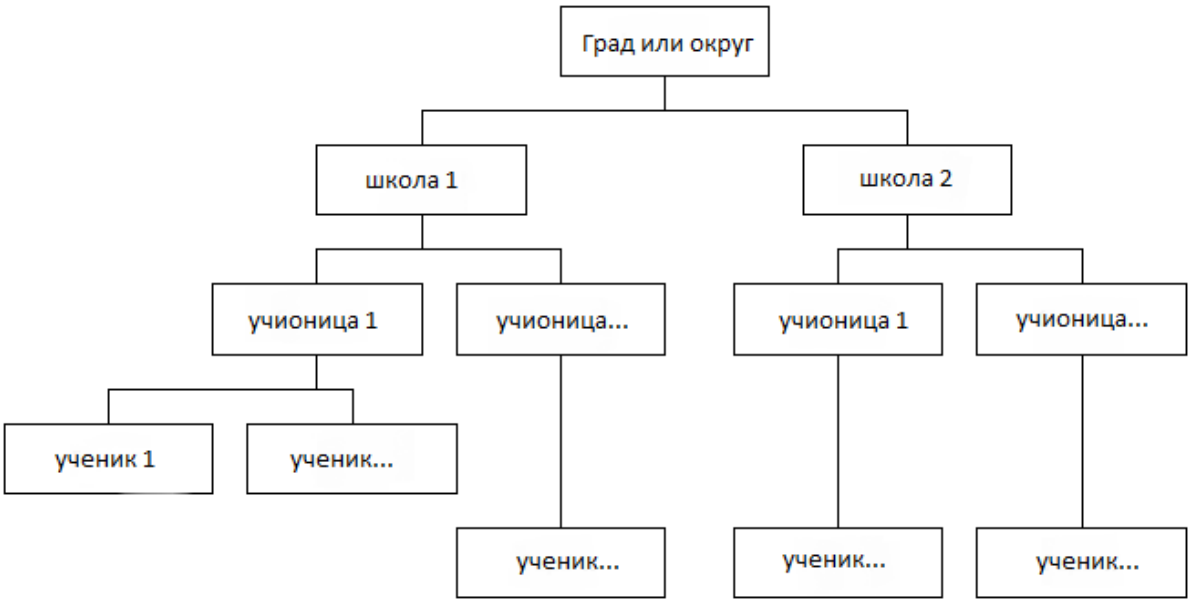
Благоје Ивановић

децембар 2018.
Београд

Садржај

Резиме

Хијерархијски линеарни модели данас имају широку примену у различитим областима науке. Интересовање за њиховом употребу на пољу образовања временом све више расте. Груписање ученика у учионице, или пак разреде, а учионице по школама које су уклопљене унутар неке управне јединице, као што је школски округ или град, представља један од основних примера хијерархијске структуре над подацима везаним за образовање. Погледајмо слику 1. Овако груписана организација се назива *хијерархијска* или *вишестепена структура*, а модели које ћемо користити за статистичку анализу ове врсте података су *хијерархијски линеарни модели*. Хијерархијски линеарни модели се обично користе у анализи груписаних или кластеризованих података, где се за посматрања у кластеру не може разумно претпоставити да су независна једна од другог. Развој ових модела узима у обзир варијабилност података унутар и међу хијерархијским нивоима.



Слика 1: Пример хијерархијске структуре у образовању

Из поменутих разлога, рад ће се надаље базирати на проблему употребе обичне регресије приликом моделирања статистика везаних за образовање, чији подаци имају природну хијерархијску структуру, као и самој примени и ефикасности претходно објашњеног модела. Користећи поменути модел, над одговарајућим подацима можемо описати и увидети разлике унутар и између различитих школа. Истраживач је, тада, у могућности да закључи у којој се мери достигнућа ученика, на пример, на пољу математике, разликују између школа, узимајући у обзир, рецимо, одређене каратеристике ученика, њихов социоекономски статус , постигнућа током школовања, број ученика по разреду, тип школе (државна или приватна), округ или град у коме се та школа налази. Дакле, у претходно наведеном примеру *ниво један* поменутног хијерархијског модела

представљали би ученици, док би се на *нивоу два* нашле школе које дати ученици похађају. Неоспорно је да бисмо, у том случају, за предикторе на *нивоу један* одабрали карактеристике ученика, њихов социоекономски статус као и постигнућа током школовања, док би се остали предиктори, тада, нашли на *нивоу два.* Дакле, број ученика по разреду, тип школе, округ или град у коме се дата школа налази су предиктори *нивоа два*. Управо над оваквим статистикама везаним за образовање, где подаци, као што се да приметити, имају природну хијерархијску структуру, јављајају се одређени проблеми приликом употребе обичне линеарне регресије. Продискутоваћемо о два таква проблема.

1. Проблем игнорисања важности ефеката групе и претпоставка о независности

   Ученици унутар хијерархија имају тенденције да буду међусобно хомогенији од ученика насумично одабраних из дате популације. Разлог овакве тврдње садржан је у чињеници да ученици нису насумично додељени школама, већ, на пример, на основу одређених достигнућа, склоности ка одређеним занимањима и слично (Математичка и Филолошка гимназија). Такође, ученици у истим школама углавном деле одређене заједничке особине, крећу се у истом окружењу, живе на истом подручју, имају исте професоре, сличне радне навике, образовни профил, наставни план и програм, и приближно сличан социоекономски статус. Дакле, уочавамо да дате опсервације на основу студената нису у потпуности независне. Међутим, једна од претпоставки технике обичне регресије јесте управо независност опсервација. Такође, примећујемо да ће разлике између школа у одређеним карактеристима дефинитивно постојати, те свако занемаривање ефеката групе може довести до погрешних и недовољно добрих закључака приликом проучавања њихових односа.

2. Проблем са „cross-level “ ефектима

   У истраживањима везаним за образовање, често се дешава да је истраживач заинтересован за испитивање везе између фактора окружења (на пример: начин рада и понашање одређеног наставника, наставни план и програм, број ученика унутар разреда) и појединачних резултата, као што су, на пример, постигнућа ученика, његово понашање и слично. С обзиром да су дати појединачни резултати прикупљени на *нивоу један*, док се остале, горе наведене, променљиве налазе на *нивоу 2* (на пример: учионица, школа, градови), поставља се питање на који начин приступити и бавити се овим „*cross-level* “ ефектима. Дакле, најбоље решење би представљали модели који истовремено могу моделирати везе на нивоу ученика уважавајући дато груписање, док свако разједињавање и уједињавање не би довољно добро могло описати природу хијерархијских података.

Увод

У литератури је добро познато да постоји све веће интересовање за коришћењем хијерархијских линеарних модела за моделирање перформанси ученика у сврху реформе образовања. Први истраживачи у овој области користили су класичне статистичке методе на једном нивоу, као што је *линеарна регресија*, да би моделирали ове ситуације. Ипак, када подаци садрже информације на више од једног нивоа, или када се јединица анализе не поклапа са случајном јединицом у експерименту, јединица анализе може постати проблем. У класичним приступима се онда мора ограничити скуп података да би се елиминисала хијерархија, спровођењем анализе на индивидуалном или групном нивоу. То доводи до деагрегације података на нивоу школе до индивидуалног нивоа или агрегације података на нивоу појединца на ниво групе занемарујући идентитет групе или информације на индивидуалном нивоу. (*Bryk, Raudenbusg, 1992.*)

Ограничења једначине на једном нивоу у каректеристикама моделирања, посебно за податке угњеждене унутар групе у облику хијерархије, довели су истраживаче до коришћења алтернативне технике моделирања, познате као *хијерархијско линеарно моделирање*. Такав приступ моделирања има много предности за истраживача, јер нема потребе да се анализирају модели нижег нивоа (ниво ученика) и вишег нивоа (ниво школе) одвојено. Ове технике не само да олакшавају декомпозицију односа између променљивих на одвојеним нивоима (на нивоу ученика и нивоу школе), већ и препознају зависност међу исходима ученика исте школе. Ова зависност може настати, на пример, као резултат заједничког искуства ученика у вези са оцењивањем наставника. У школама, што више ученика дели заједничка искуства због блискости у простору и/или времену имају више сличности. Хијерархијски линеарни модели могу истовремено да проучавају ефекте и на нижем и на вишем нивоу. Штавише, корелисане грешке и ненула ICC (intra-class correlation, корелација унутар класа - основна мера за степен зависности у кластер посматрањима), нераздвојни у груписаним подацима, су на неки начин уграђени у хијерархијски линеарни модел, дајући тачне стандардне процене грешака и закључке. Такође, приступ на више нивоа заснива се на отпуштању претпоставки у зависности од методе, алгоритма и софтвера који се користи.

Хијерархијски линеарни модели на два нивоа

Хијерархијски линеарни модел (ХЛМ) претпоставља хијерархијске податке са једном одговарајућом променљивом одговора на најнижем нивоу и објашњавајућим променљивима на свим постојећим нивоима. Концептуално модел се често посматра као хијерархијски систем регресионих једначина (Hox,1998.).

У даљем раду описиваћемо проблеме са хијерархијом на два нивоа, као што су, на пример, ученици у школама, иако је, такође, могуће вршити хијерархију на више нивоа (на пример ученици у одељењима, одељења у школама, школе у окрузима и тако даље).

Дакле, имамо податке за $\boldsymbol{j = 1, \dots, M}$ група (школа) и различит број $\boldsymbol{n_j}$ појединаца (ученика) у свакој од тих групи. Подаци не морају нужно бити уравнотежени. На нивоу ученика (*ниво један*) имамо зависну променљиву $\boldsymbol{Y_{ij}}$ и променљиву објашњења (то јест независну променљиву) $\boldsymbol{x_{ij}}$. На нивоу школе (*ниво два*) такође имамо променљиву објашњења $\boldsymbol{W_j}$. Двоструки индекс за ове променљиве указује да су опажања јединствена за сваког ученика $\boldsymbol{i}$ унутар сваке школе $\boldsymbol{j}$.

Ниво 1 регресионе једначине

Дакле, у *хијерархијском линеарном моделу на два нивоа*, можемо имати одвојене *ниво један* регресионе једначине на свакој од јединица *нивоа два*. Када постоји једна независна променљива *нивоа један*, модел *нивоа један* тј. модел унутар једне школе може бити представљен на следећи начин :

$$\boldsymbol{Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}} \qquad (1)$$

где :

- Y_{ij} представља исход за i – тог ученика у j –тој школи тј. *зависна променљива* за појединачно посматрање на *нивоу један* (индекс i односи се на индивидуалан случај, а индекс j односи се на групу)
- x_{ij} односи се на предиктор *нивоа један* тј. *објашњава променљиву* за i – тог ученика у j –тој школи
- β_{0j} је *коефицијент регресије* који се односи на пресретање зависне променљиве у j –тој групи тј. *коефицијент пресретање* за j –ту школу
- β_{1j} је *коефицијент регресије* који се односи на нагиб за j –ту школу
- ε_{ij} је *случајна грешка предвиђања* за i – тог ученика у j –тој школи са предикционе линије школе

Индекси за коефицијенте β у овој једначини показују да се они могу разликовати за сваку школу j. На *нивоу један*, коефицијент пресретања и коефицијент нагиба у групама могу

бити фиксирани (што значи да све групе имају исте вредности, у реалности је ово ретка појава), не-случајно различити (што значи да су коефицијент пресретања и/или коефицијент нагиба предвидљиви из независне променљиве на нивоу 2) или случајно различити (што значи да су коефицијент пресретања и/или коефицијент нагиба различити у различитим групама и да сваки од њих има сопствену средњу вредност и дисперзију).

Када постоји више независних променљивих *нивоа један*, модел се може проширити заменом вектора матрицама у једначини.

Ниво 2 регресионе једначине

Зависне променљиве су коефицијенти пресретања и коефицијенти нагиба за независне променљиве на *нивоу један* у групама *нивоа два*.

Коефицијент пресретања β_{0j} и коефицијент нагиба β_{1j} се моделирају помоћу објашњавајућих променљивих у *нивоу два* тј. моделу између школа на следећи начин :

$$\boldsymbol{\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}} \qquad (2)$$

$$\boldsymbol{\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}} \qquad (3)$$

где је :

- γ_{00} *просечно пресретање* када је $W_j = 0$ тј. то је средња вредност зависне променљиве у свим групама када су сви предиктори једнаки нули
- γ_{01} *укупни коефицијент регресије*, или нагиба, између зависне променљиве и предиктора *нивоа два*
- u_{0j} *случајна грешка* (одступање коефицијента пресретања од просечног пресретања) за j −ту школу
- γ_{10} *укупни коефицијент регресије*, или нагиба, између зависне променљиве и предиктора *нивоа један* када је $W_j = 0$
- u_{1j} *случајна грешка*(одступање коефицијента нагиба од просечног нагиба) за j −ту школу

γ_{00} и γ_{10} су *коефицијенти регресије* повезани са ефектима објашњења нивоа школе на структуралне односе на нивоу ученика.

Заменом (2) и (3) у (1) добија се:

$$\boldsymbol{Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} W_j + \gamma_{11} W_j x_{ij} + u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}} \qquad (4)$$

Део $\gamma_{11} W_j x_{ij}$ се назива ефекат интеракције на више нивоа $(cross - level\ interaction\ effect)$.

Када се користи више од једне променљиве на првом или другом нивоу, означићемо индексе са p $(p = 1,2, \dots, P)$ за први ниво и k $(k = 1,2, \dots, K)$ за *други ниво*. Тада (4) постаје општија једначина (Hox,1998.,2002. ;Snijders,Bosker,1999.) :

$$\boldsymbol{Y_{ij} = \gamma_{00} + \gamma_{p0} x_{pij} + \gamma_{0q} W_{qj} + \gamma_{pq} W_{qj} x_{pij} + u_{pj} x_{pij} + u_{0j} + \varepsilon_{ij}} \qquad (5)$$

Први део (5), $\gamma_{00} + \gamma_{p0} x_{pij} + \gamma_{0q} W_{qj} + \gamma_{pq} W_{qj} x_{pij}$ се зове *фиксни део модела*. Други део, $u_{pj} x_{pij} + u_{0j} + \varepsilon_{ij}$ се зове *случајни део модела*. Члан $u_{pj} x_{pij}$ се може сматрати *случајном интеракцијом* између школе и x −а.

Спецификација члана грешке на нивоу ученика (ε) и на нивоу школе (u) дозвољава хијерархијском линеарном моделу да адекватно моделира грешку у гупираним подацима.

Променљиве x и W могу бити моделиране у њиховом оригиналу, нетрансформисане метрички или могу бити центриране.(Сулливан,1999.)

Општа формулација модела на два нивоа

Модел на два нивоа се може изразити на следећи начин:

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \gamma + W_j U_j + R_j$$

где $\begin{bmatrix} R_j \\ U_j \end{bmatrix} \sim N\left(\begin{bmatrix} \emptyset \\ \emptyset \end{bmatrix}, \begin{bmatrix} \sum_j(\theta) & \emptyset \\ \emptyset & \Omega(\xi) \end{bmatrix}\right)$ и $(R_j, U_j) \perp (R_l, U_l)$ за $j \neq l$

Стандардна спецификација $\sum_j(\theta) = \sigma^2 I_{n_j}$, али могуће су и друге спецификације. Углавном је $\sum_j(\theta)$ дијагонална али и не мора да буде.

Тада модел можемо написати као: $\boldsymbol{Y}_j \sim N(\boldsymbol{X}_j \gamma, W_j \Omega(\xi) W_j' + \sum_j(\theta))$

Типови модела

Пре спровођења анализе модела на више нивоа, истраживач мора да одлучи о неколико аспеката, укључујући, такође, и то које предикторе треба да укључи у анализу, ако их има. Такође, истраживач мора да одлучи да ли ће вредности параметара (елементи који ће бити процењени) бити фиксиране $(fixed)$ или случајне $(random)$. Фиксирани параметри се састоје од константе над свим групама, док случајни параметар има различиту вредност за сваку од група. Поред тога, истраживач мора одлучити да ли ће применити процену методом максималне веродостојности или, пак, неком другом методом.

Модел случајног пресретања

Модел случајног пресретања је модел у којем је дозвољено да се коефицијенти пресретања мењају, па стога су оцене за зависну променљиву за свако појединачно посматрање предвиђене пресретањем које се разликује по групама. Овај модел претпоставља да су коефицијенти нагиба фиксирани. Поред тога, овај модел пружа информације о интракласним корелацијама, које су корисне за одређивање да ли је пре свега потребан модел на више нивоа.

Модел случајних нагиба

Модел случајног нагиба је модел у којем је дозвољено да се коефицијенти нагиба мењају, тј. нагиби су различити међу групама. Претпоставке овог модела су да су коефицијенти пресретања фиксирани.

Модел случајних пресретања и нагиба

Модел који укључује и случајно пресретање и случајне нагибе је вероватно најреалнији тип модела, али је и најсложенији. У овом моделу, и коефицијенти нагиба и коефицијенти пресретања могу да се мењају по групама.

Развој модела на више нивоа

Да би се спровела анализа модела на више нивоа, могло би се почети са фиксираним коефицијентима (нагиба и пресретања). Један аспект би се могао мењати у исто време и упоређивати са претходним моделом како би се оценило боље уклапање модела. Постоје три различита питања на која би истраживач требао да одговори.

1) Да ли је модел добар?
2) Да ли је сложенији модел бољи?
3) Какав допринос поједини предиктори дају моделу?

Претпоставке модела

Претпоставке хијерархијског линеарног модела су исте као и претпоставке линеарног модела ограничене на један ниво ОЛС (ordinary least squares-метода најмањих квадрата) регресије.

- Линеарност
- Нормираност
- Хомоскедастичност $D(\varepsilon_{ij}) = \sigma^2 > 0$
- x_{ij} и ε_{ij} су независни

У нашем моделу, једначина (1), Y_{ij} је непрекидна зависна променљива, стога претпостављамо да су грешке нивоа 1 модела нормалне случајне величине са математичким очекивањем 0 и дисперзијом σ^2:

$$E(\varepsilon_{ij}) = 0 \ , \ D(\varepsilon_{ij}) = \sigma^2$$

У нивоу 2 модела, једначине (2) и (3), претпостављамо да су параметри модела β_{0j} и β_{1j} независне и једнако расподељене случајне величине са вишедимензионом нормалном расподелом :

$$E(\beta_{0j}) = \gamma_{00} \ , \ var(\beta_{0j}) = \tau_{00} = {\tau_0}^2$$

$$E(\ \beta_{1j}) = \gamma_{10} \ , \ var(\ \beta_{1j}) = \tau_{11} = {\tau_1}^2$$

Грешке нивоа 1 и нивоа 2 су хомогене и некорелисане тј. $E(\varepsilon_i \varepsilon_j) = 0$ за $i \neq j$

У наставку ћемо сумирати матемазичке изразе претпоставки које ће нам требати:

$$E(u_{0j}) = 0 \ , \ E(u_{1j}) = 0$$

$$var(\beta_{0j}) = var(u_{0j}) = \tau_{00} = {\tau_0}^2$$

$$var(\beta_{1j}) = var(u_{1j}) = \tau_{11} = {\tau_1}^2$$

$$cov(\beta_{0j}, \beta_{1j}) = cov(u_{0j}, u_{1j}) = \tau_{01}$$

$$cov(u_{0j}, \varepsilon_{ij}) = cov(u_{1j}, \varepsilon_{ij}) = 0$$

Претпоставља се да су коефицијенти који зависе од групе (u_{0j}, u_{1j}) независни за различито j. (u_{0j}, u_{1j}) нису појединачни параметри у статистичком смислу, већ су само њихове дисперзије и коваријације параметри. Према томе имамо линеарани модел за средњу вредност и коваријациону матрицу параметара унутар група са независношћу између група.

Испитивање података на основу базе

Примену поменутог хијерархијског линеарног модела , у овом случају на два нивоа, приликом моделирања статистика везаних за образовање, демонстрираћемо над подацима из база ***MathAchieve*** и ***MathAchSchool*.** Наведене базе су део *nlme* пакета и садрже податке за 7185 средњошколаца који похађају, управо, једну од датих 160 школа. Дакле, за сваку од школа, поседујемо информације о, у просеку, 45 ученика који је у том тренутку похађају. *Ниво један* поменутог хијерархијског линеарног модела представљају ученици, док се на нивоу два, према томе, налазе школе у које су они пријављени. Прва база података односи се на средњошколце, где, дакле, сваки ред представља управо једног од седам хиљада сто осамдесет пет наведених ученика и његове одређене карактеристике. Прикажимо првих десет редова овог скупа података. Запазимо да средњошколци представљени у првих десет редова овог скупа похађају школу под идентификационим бројем ” *1224* ”.

```
library(nlme)

dim(MathAchieve)

## [1] 7185    6

head(MathAchieve,10)

## Grouped Data: MathAch ~ SES | School
##    School Minority    Sex    SES MathAch MEANSES
## 1    1224       No Female -1.528   5.876  -0.428
## 2    1224       No Female -0.588  19.708  -0.428
## 3    1224       No   Male -0.528  20.349  -0.428
## 4    1224       No   Male -0.668   8.781  -0.428
## 5    1224       No   Male -0.158  17.898  -0.428
## 6    1224       No   Male  0.022   4.583  -0.428
## 7    1224       No Female -0.618  -2.832  -0.428
## 8    1224       No   Male -0.998   0.523  -0.428
## 9    1224       No Female -0.888   1.527  -0.428
## 10   1224       No   Male -0.458  21.521  -0.428

summary(MathAchieve)

##      School     Minority       Sex            SES
##  2305   :  67   No :5211   Male  :3390   Min.   :-3.758000
##  5619   :  66   Yes:1974   Female:3795   1st Qu.:-0.538000
##  4292   :  65                            Median : 0.002000
##  8857   :  64                            Mean   : 0.000143
##  4042   :  64                            3rd Qu.: 0.602000
##  3610   :  64                            Max.   : 2.692000
##  (Other):6795
##     MathAch          MEANSES
##  Min.   :-2.832   Min.   :-1.188000
##  1st Qu.: 7.275   1st Qu.:-0.317000
##  Median :13.131   Median : 0.038000
##  Mean   :12.748   Mean   : 0.006138
##  3rd Qu.:18.317   3rd Qu.: 0.333000
##  Max.   :24.993   Max.   : 0.831000
##
```

Друга база података односи се на школе, по којима су наведени ученици , из прве базе података, груписани, и сваки ред представља једну од 160 наведених школа. Прикажимо првих десет редова , то јест школа, овог скупа података.

```
dim(MathAchSchool)
```

```
## [1] 160   7
```

```
head(MathAchSchool,10)
```

```
##      School Size   Sector PRACAD DISCLIM HIMINTY MEANSES
## 1224   1224  842   Public   0.35   1.597       0  -0.428
## 1288   1288 1855   Public   0.27   0.174       0   0.128
## 1296   1296 1719   Public   0.32  -0.137       1  -0.420
## 1308   1308  716 Catholic   0.96  -0.622       0   0.534
## 1317   1317  455 Catholic   0.95  -1.694       1   0.351
## 1358   1358 1430   Public   0.25   1.535       0  -0.014
## 1374   1374 2400   Public   0.50   2.016       0  -0.007
## 1433   1433  899 Catholic   0.96  -0.321       0   0.718
## 1436   1436  185 Catholic   1.00  -1.141       0   0.569
## 1461   1461 1672   Public   0.78   2.096       0   0.683
```

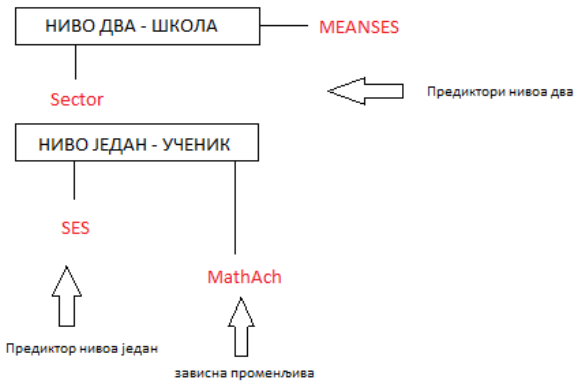
```
summary(MathAchSchool)
```

```
##      School          Size             Sector       PRACAD      
##  1224   :  1   Min.   : 100.0   Public  :90   Min.   :0.0000  
##  1288   :  1   1st Qu.: 588.5   Catholic:70   1st Qu.:0.3100  
##  1296   :  1   Median :1061.0                 Median :0.5000  
##  1308   :  1   Mean   :1097.8                 Mean   :0.5139  
##  1317   :  1   3rd Qu.:1526.0                 3rd Qu.:0.6825  
##  1358   :  1   Max.   :2713.0                 Max.   :1.0000  
##  (Other):154                                                 
##     DISCLIM          HIMINTY    MEANSES        
##  Min.   :-2.41600   0:116   Min.   :-1.1880000  
##  1st Qu.:-0.72600   1: 44   1st Qu.:-0.3012500  
##  Median :-0.07800           Median : 0.0345000  
##  Mean   :-0.01512           Mean   :-0.0001875  
##  3rd Qu.: 0.66625           3rd Qu.: 0.3200000  
##  Max.   : 2.75600           Max.   : 0.8310000  
##
```

Променљиве, из две претходно наведене базе, које ће нам бити потребне у даљем сегменту рада детаљније ћемо описати.

- **School**
  Променљива *school* представља идентификациони број школе и појављује се у обе, од две наведене, базе. Подаци унутар база су уређени (Прво су дате информације о ученицима једне, а затим друге, наредне, школе. Школе су поређане од најмањег до највећег индентификационог броја.), иако функције *lmer* и *lme*, о којима ће касније бити речи, то не захтевају. Школе дефинишу групе и припадају *нивоу два* хијерархијског линеарног модела. Такође, нелогично је претпоставити да су средњошколци унутар истих школа независни, с обзиром да имају, на пример, исте наставнике, књиге, наставни план и програм, као и средину у којој бораве.

- **SES**
  Променљива *ses* дефинише социоекономски статус ученикове породице. Такође, ову променљиву узећемо за предиктор, то јест, независну променљиву на *нивоу један* нашег хијерархијског линеарног модела.
- **MathAch**
  Променљива *MathAch* представља зависну променљиву нашег хијерархијског линеарног модела и приказује поене које је ученик постигао на завршном тесту из математике(тест показује учениковова достигнућа из математике,колико разуме и колико је савладао различите математичке теме, за погрешан одговор на тесту, такође, постоје и негативни поени).
- **Sector**
  Променљива *sector* пружа информацију да ли је школа, са одређеним идентификационим бројем државна или католичка, и као таква представља предиктор, то јест, независну променљиву нивоа два нашег хијеархијског линеарног модела. Запазимо да за ученике унутар исте школе променљива *sector* узима идентичну вредност, па ћемо, према томе, сваком ученику, то јест реду базе *MathAchieve,* касније доделити одговарајућу вредност ове променљиве.
- **MEANSES**
  Променљива *MEANSES* пружа информацију о просечном социоекономском статусу породица ученика који похађају одређену школу, те као таква представља предиктор, то јест, независну променљиву нивоа два нашег хијерархијског модела.



С обзиром да су нам предиктори *нивоа један* и зависна променљива смештени у бази *MathAchieve*, а предиктори *нивоа два* у наредној бази, *MathAchSchool* , потребно је формирати нову базу од датих података, тако да она у себи садржи како предикторе *нивоа један* и зависну променљиву, тако и предикторе *нивоа два*. Дату базу над траженим подацима назваћемо *Math*, те у првом кораку у њу сместити предикторе нивоа један и нашу зависну променљиву из базе *MathAchieve*. Прикажимо десет насумично одабраних редова овако формиране базе *Math.*

```
is.factor(Sector)

## [1] TRUE

Math <- as.data.frame(MathAchieve[, c("School", "SES", "MathAch")])
set.seed(3)
sample1 <- sort(sample(nrow(Math), 10))
Math[sample1, ]
```

```
##      School    SES MathAch
## 895    2305  0.182  12.313
## 1208   2629  0.502  15.180
## 2115   3427  0.672  20.621
## 2354   3657  1.512  18.610
## 2766   4042  0.772  22.923
## 4146   5838 -0.038  13.903
## 4324   6144 -0.808  13.247
## 4340   6170  1.002  19.951
## 4528   6415  0.942  13.091
## 5802   8150  0.902  16.405
```

У следећем кораку, новоформираној бази *Math* додаћемо и предикторе *нивоа два*, те опет приказати истих десет насумично одабраних редова.

```
sector <- MathAchSchool$Sector
names(sector) <- row.names(MathAchSchool)
mses <- with(MathAchieve, tapply(SES, School, mean))
Math <- within(Math,{ Meanses <- as.vector(mses[as.character(School)])
                      Cses <- SES - Meanses
                      Sector <- sector[as.character(School)] })
Math[sample1,]
```

```
##      School    SES MathAch   Sector       Cses    Meanses
## 895    2305  0.182  12.313 Catholic  0.8100000 -0.6280000
## 1208   2629  0.502  15.180 Catholic  0.6396491 -0.1376491
## 2115   3427  0.672  20.621 Catholic  0.5189796  0.1530204
## 2354   3657  1.512  18.610   Public  2.1611765 -0.6491765
## 2766   4042  0.772  22.923 Catholic  0.3700000  0.4020000
## 4146   5838 -0.038  13.903   Public -0.1945161  0.1565161
## 4324   6144 -0.808  13.247   Public -0.3704651 -0.4375349
## 4340   6170  1.002  19.951   Public  1.3038095 -0.3018095
## 4528   6415  0.942  13.091   Public  1.1292593 -0.1872593
## 5802   8150  0.902  16.405 Catholic  0.5961364  0.3058636
```

Приметимо да смо у претходном кораку бази додали и једну модификацију променљиве *SES*, тачније, центрирану променљиву *Cses*. Ову променљиву користићемо у даљем раду. Размотримо следећи пример. Нека зависна променљива Y даје информацију о броју речи које се налазе у вокабулару одређеног детета током прве године живота. Независна променљива, то јест предиктор X, даје информацију о броју месеци који су протекли од рођења па све до тренутка дечије прве савладане и изговорене речи. Уколико променљива X није центрирана, на основу коефицијента пресретања у моделу предвиђања броја речи које се налазе у вокабулару одређеног детета током прве године живота, поседоваћемо информацију о средњој вредности броја речи у вокабулару беба које су своју прву реч савладале и изговориле на рођењу (X=0). Примећујемо да оваква информације нема смисла, те самим тим није ни од каквог значај. Уколико бисмо дату променљиву X центрирали на основу просечне вредности протеглог броја месеци након којих је дете савладало своју прву реч, коефицијент пресретања у моделу предвиђања броја речи које се налазе у вокабулару одређеног детета током прве године живота, давао би информацију о средњој вредности броја речи у вокабулару беба које су своју прву реч савладале управо у просечној вредности броја месеци након којих дете углавном савлада своју прву реч. Дакле, запазимо, да је ово, управо идеја којом смо се водили приликом центрирања наше променљиве *SES* на основу средње вредности социоекономског статуса породице ученика унутар школа.

Пре самог статистичког моделирања, испитаћемо дате податке ради бољег разумевања структуре, као и прикупљања информација које ће нам у даљем раду над датим подацима бити од користи.

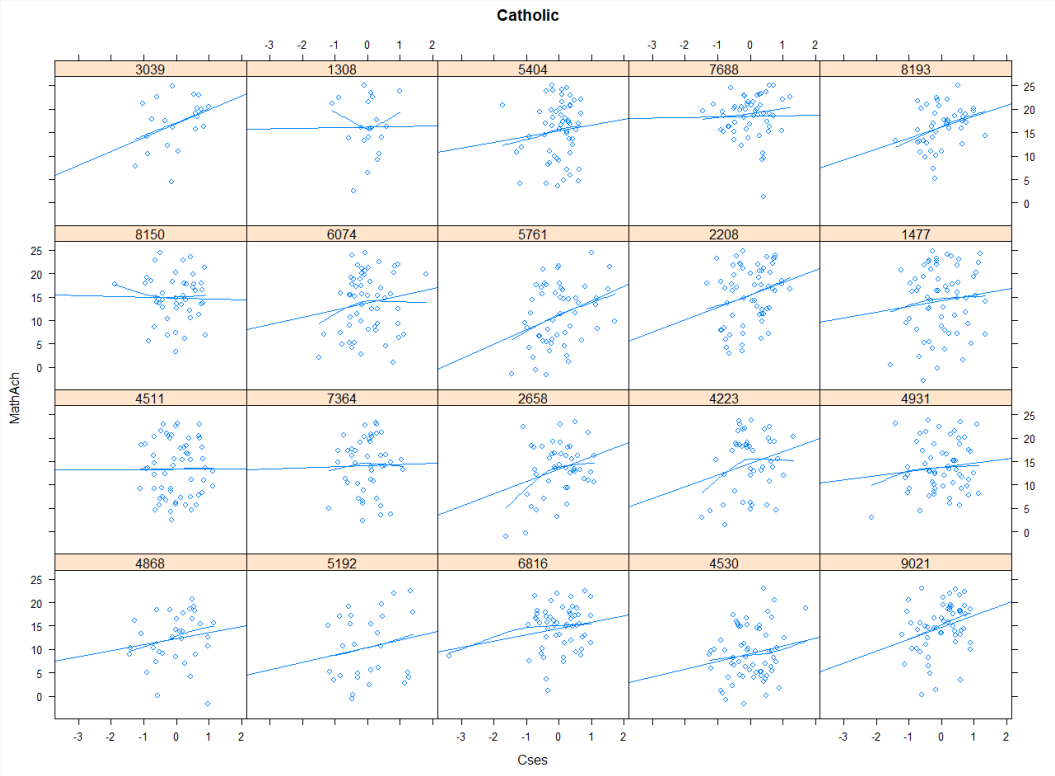
С обзиром да наша база *Math* садржи 160 школа, што је приликом испитивања сваке појединачно огроман број, ми ћемо се фокусирати, у овом тренутку, на двадесет католичких и двадесет државних школа.

```
cato <- with(Math, sample(unique(Math$School[Sector == "Catholic"]), 20))
cato2 <- Math[is.element(Math$School, cato), ]
public <- with(Math, sample(unique(Math$School[Sector == "Public"]), 20))
public2 <- Math[is.element(Math$School, public), ]
```

Дакле, база *cato2* садржи информације о двадесет насумично одабраних католичких школа, а база *public2* информације о двадесет, насумично одабраних, државаних школа. Ради визулизације и приказивања односа и зависности броја поена које је ученик постигао на задатом тесту из математике (променљива *MathAch*) са социоекономским статусом породица датих ученика (променљива *Cses*) користићемо графички приказ *lattice* из истоименог пакета.

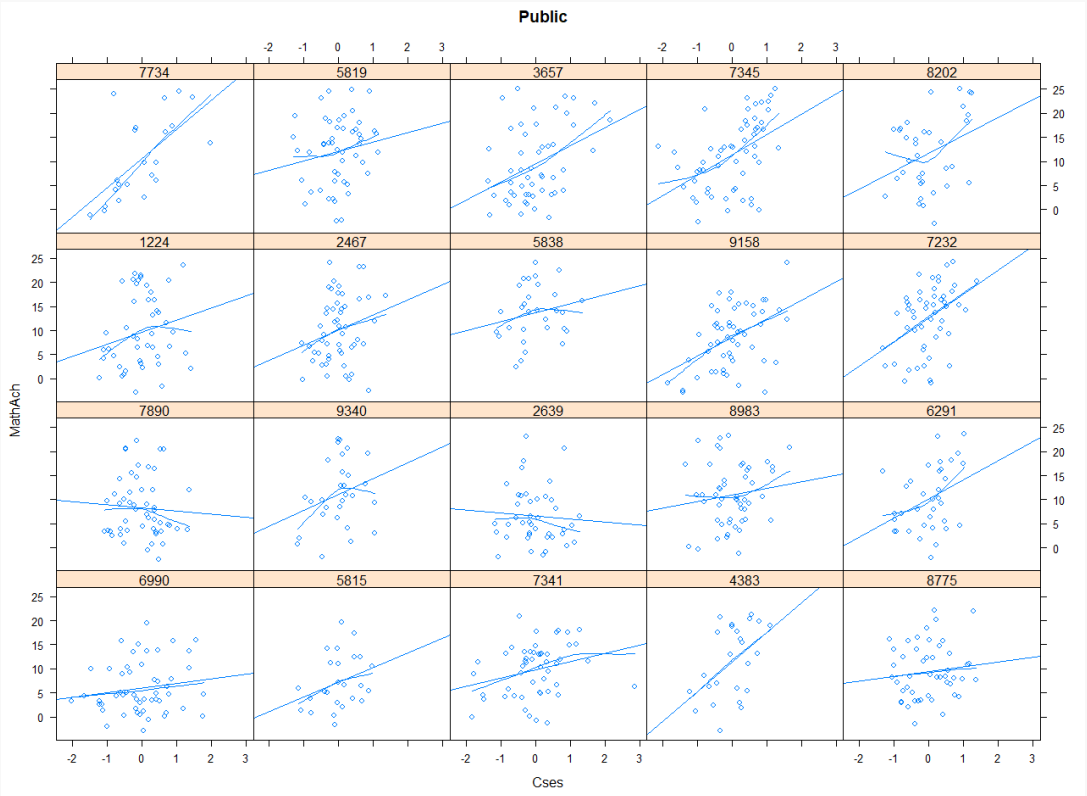
```
library(lattice)
trellis.device(color=TRUE)
xyplot(MathAch ~ Cses | School, data=cato2, main="Catholic", type=c("p", "r", "smooth"), span=1)
```

Слика 2 : Решетка приказује математичка достигнућа према социо-економском статусу за 20 случајно одабраних Католичких школа. Изломљене линије дају линеарно најмањи квадрат.



```
xyplot(MathAch ~ Cses | School, data=public2, main="Public", type=c("p", "r", "smooth"), span=1)
```

Слика 3: Решетка приказује математичка достигнућа према социо-економском статусу за 20 случајно одабраних државних школа.



На основу претходних графичких приказа примећујемо постојање слабе позитивне зависности између променљиве *MathAch* и променљиве *Cses* у већини католичких школа. Код неколико испитаних школа коефицијент нагиба је једнак нули, или пак негативан. Такође, може се запазити позитивна зависност променљиве *MathAch* и променљиве *Cses* код државних школа, са, наизглед, већим коефицијентом нагиба него код католичких. Дакле, с обзиром на ограничен број испитаних ученика, чини се да линеарне регресије могу пружити разумљив сажетак односа променљивих *MathAch* и *Cses* унутар школа.

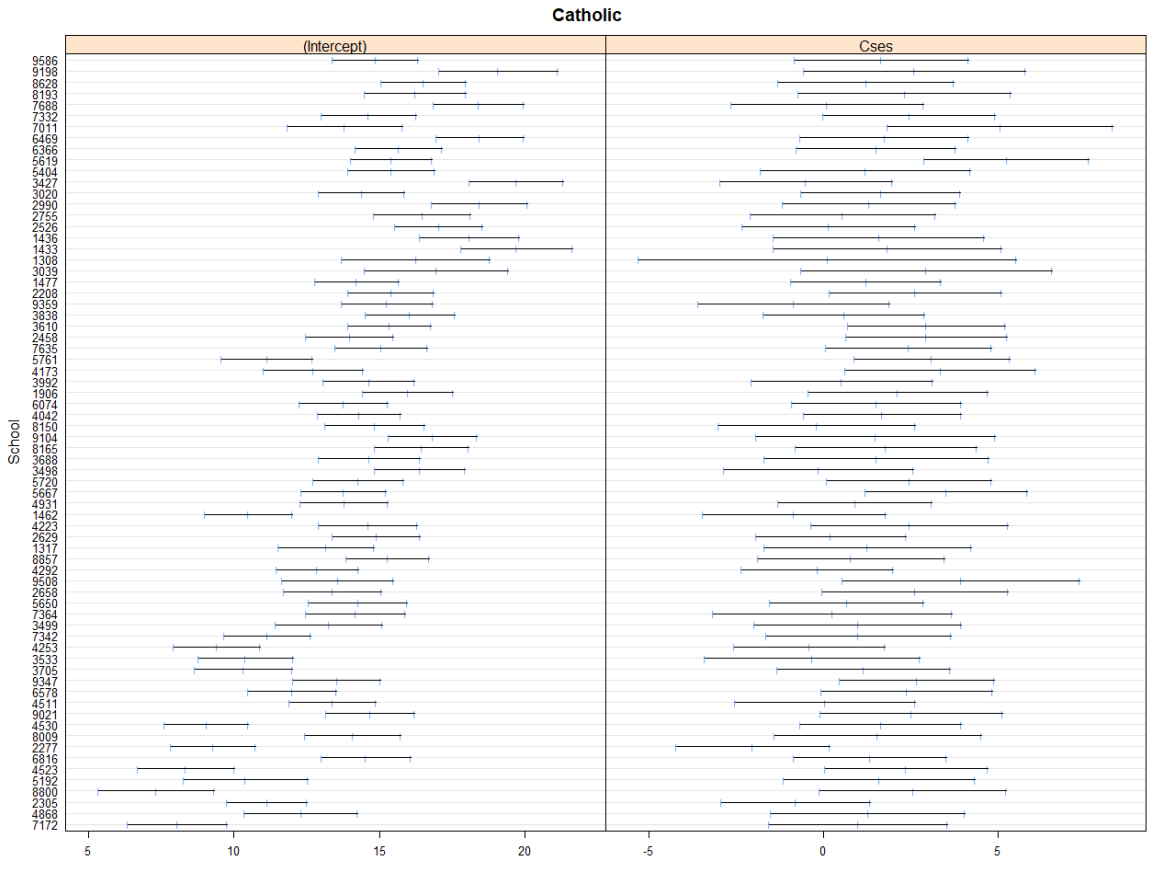
У склопу, већ поменутог, пакета *nlme* садржана је, такође, и функција *lmList* која за сваку партицију података из претходно наведене базе, груписаних на основу одређеног фактора групације, рачуна, посебно, коефицијенте линеарне регресије. Дакле, функција враћа листу објеката линеарног модела, која је сама по себи објекат класе *lmList.* На основу поменуте функције, проценимо коефицијенте линеарних регресија (***ниво два*** тј. фактор групације је школа) за модел који предвиђа број поена које је ученик постигао на задатом тесту из математике (*MathAch*) на основу социоекономског статуса породица датих ученика (*Cses*) за католичке, те, одвојено, за државне школе.

```
cat.list <- lmList(MathAch ~ Cses | School,subset =Sector == "Catholic",data = Math)
pub.list <- lmList(MathAch ~ Cses | School,subset =Sector == "Public",data = Math)
```

Ради бољег прегледа добијених резултата, као и могућности доласка до одређених закључака, прикажимо 95-опроцентне интервале поверења за коефицијенте линеарних регресија.

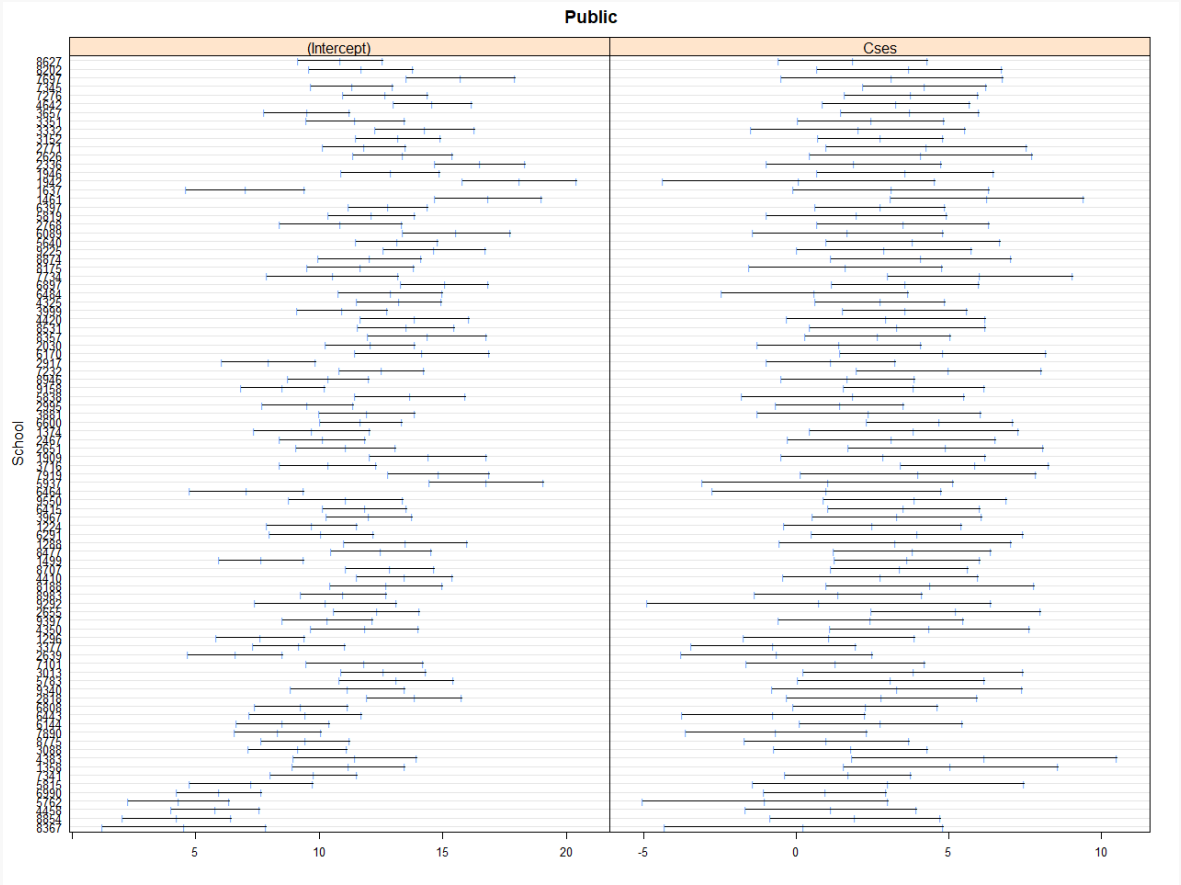
```
plot(intervals(cat.list),main="Catholic")
```

Слика 4 : 95%-ни интервал поверења за Католичке школе



```
plot(intervals(pub.list),main="Public")
```

Слика 5 : 95%-ни интервал поверења за државне школе



У тумачењу ових графика, морамо бити опрезни и узети у обзир да нисмо ограничили скале да графици буду исти. У нашем случају скале за коефицијенте пресретања и нагиба у државним школама су шире него у католичким школама. Пошто је случајна променљива SES центрирана око 0 унутар школа, коефицијенти нагиба се тумаче као просечни ниво достигнућа у математици у свакој школи. Јасно је да постоји значајна разлика коефицијената пресретања између државних и католичких школа.Интервали поверења за коефицијенте нагиба, насупрот томе, преклапају се у много већој мери, али још увек постоји очигледна разлика између школа.

Приликом тумачења претходних графичких приказа, врло је битно приметити, да поводом представљања интервала, у обзир није узета мисао да њихов приказ буде увек на идентичном делу праве. Тачније, интервали поверења, за коефицијенте линеарних регресија код државних школа, припадају већем скупу вредности. Јасно је да постоји значајна варијација међу коефицијентима пресретања како унутар католичких, тако и државних школа. Насупрот томе, иако поред још увек очигледне варијације, коефицијенти нагиба се поклапају у много већој мери.

Сада ћемо одредити процене коефицијената(редови представљају школе):

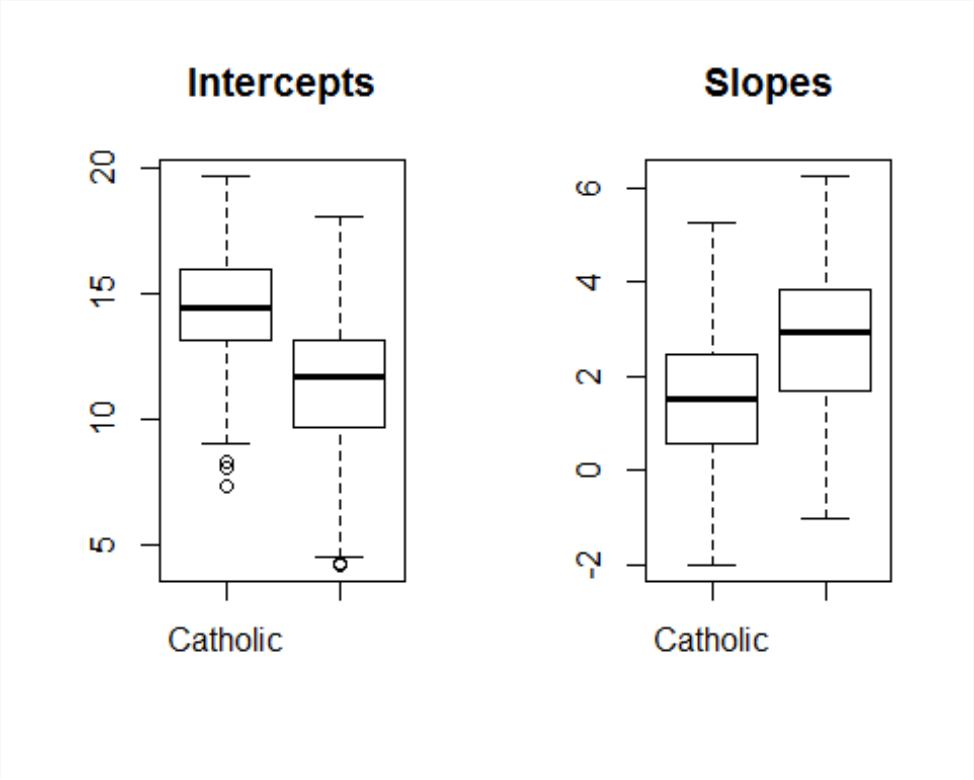
```
cat.coef <- coef(cat.list)
head(cat.coef,6)
```

```
##      (Intercept)       Cses
## 7172    8.066818  0.9944805
## 4868   12.310176  1.2864712
## 2305   11.137761 -0.7821112
## 8800    7.335937  2.5681254
## 5192   10.409500  1.6034950
## 4523    8.351745  2.3807892

pub.coef <- coef(pub.list)
head(pub.coef,6)

##      (Intercept)       Cses
## 8367    4.552786  0.2503748
## 8854    4.239781  1.9388446
## 4458    5.811396  1.1318372
## 5762    4.324865 -1.0140992
## 6990    5.976792  0.9476903
## 5815    7.271360  3.0180018
```

Паралелни *box plot*-ови пружају другачију и сажетију визуализацију процењених вредности коефицијената линеарних регресија и као такви бивају, за одређене закључке, више употребљиви.

```
par(mfrow=c(1,2))
boxplot(cat.coef[,1],pub.coef[,1],main="Intercepts",names = c("Catholic","Public"))
boxplot(cat.coef[,2],pub.coef[,2],main="Slopes",names = c("Catholic","Public"))
```

Слика 6 : *Box plot*-ови за коефицијенте пресретања и нагиба



Католичке школе остварују већи просечан број постигнутих поена на тесту из математике у односу на државне школе, док је коефицијент нагиба код државних већи у односу на коефицијент нагиба код католичких школа.

Време је да формирамо хијерархијски линеарни модел који предвиђа број постигнутих поена на задатом тесту из математике, то јест, променљива *MathAch* је зависна променљива нашег хијерархијског линеарног модела. Поменути модел, као што је већ познато, састоји се из два сета једначина. Прво, унутар школа, дефинишемо регресију на *првом* ,индивидуалном, *нивоу* са зависном променљивом *MathAch* и независном променљивом *Cses.* Променљива *Cses* је, дакле, предиктор *нивоа један*. Први сет једначина (једначине *првог нивоа*), користећи независну променљиву *Cses,* за сваку индивидуу (ученика) j у школи i гласи:

$$MathAch_{ij} = \alpha_{0i} + \alpha_{1i}Cses_{ij} + \varepsilon_{ij}$$

На другом нивоу, то јест нивоу школе, узећемо у обзир могућност да коефицијенти пресретања и нагиба зависе од типа школе (Sector), као и просечног социоекономског статуса породица ученика унутар те школе (Meanses). Дакле, променљиве Sector и Meanses су предиктори нивоа два. У том случају, други сет једначина (једначине другог нивоа) гласи:

$$\alpha_{0i} = \gamma_{00} + \gamma_{01}Meanses_i + \gamma_{02}Sector_i + u_{0i}$$

$$\alpha_{1i} = \gamma_{10} + \gamma_{11}Meanses_i + \gamma_{12}Sector_i + u_{1i}$$

С тога, замењујући коефицијенте у првој једначини другим сетом једначина, имамо следеће:

$$MathAch_{ij} = \gamma_{00} + \gamma_{01}Meanses_i + \gamma_{02}Sector_i + u_{0i} + (\gamma_{10} + \gamma_{11}Meanses_i + \gamma_{12}Sector_i + u_{1i})Cses_{ij} + \varepsilon_{ij}$$

Ради лакше прегледности приликом рада са претходном једначином реорганизоваћемо термине унутар ње.

$$MathAch_{ij} = \gamma_{00} + \gamma_{01}Meanses_i + \gamma_{02}Sector_i + \gamma_{10}Cses_{ij} + \gamma_{11}Meanses_iCses_{ij} + \gamma_{12}Sector_iCses_{ij} + u_{0i} + u_{1i}Cses_{ij} + \varepsilon_{ij}$$

Иако је интуитивно јасно, напоменимо да термине означене црвеном бојом називамо *коефицијентима ефекта*. Такође, термини означени плавом бојом, заједно са случајним грешкама предвиђања *нивоа један,* представљају *random* ефекте.

Коначно, препишимо претходну једначину у нотацији вишеструких линеарних модела (LMM).

$$MathAch_{ij} = \beta_1 + \beta_2Meanses_i + \beta_3Sector_i + \beta_4Cses_{ij} + \beta_5Meanses_iCses_{ij} + \beta_6Sector_iCses_{ij} + b_{i1} + b_{i2}Cses_{ij} + \varepsilon_{ij}$$[1]

Промена је, дакле, чисто нотацијска. Ознаке β_s коришћене су приликом записа $fixed$ ефеката, а b_s поводом записа *random*.

Сада, већ формирани, хијерархијски линерни модел може се анализирати позивом функције *lme,* присутне као део одавно поменутог пакета *nlme.* Запис фиксираних

[1] За случајне грешке предвиђања нивоа један важи претпоставка о независности и константној варијацији.

ефеката у позиву дате функције идентичан је запису коефицијената линеарног модела приликом позива функције *lm,* док је запис случајних ефеката дефинисан аргументом *random* дате функције, узимајући приликом тога једностране моделну формулу. Такође, пре датог позива функције, за хијерархијски линеарни модел предвиђања броја постигнух поена на тесту из математике, уредили смо, ради лакшег рада, променљиву *Sector*, тако да *contrast* дате променљиве узима вредност *0* за државне, те *1* за католичке школе.

```
Math$Sector <- factor(Math$Sector,levels = c("Public","Catholic"))
contrasts(Math$Sector)
```

```
##          Catholic
## Public          0
## Catholic        1
```

Потом, након промене референтног нивоа фактора, позивамо поменуту функцију *lme.*

```
Mathlme1 <- lme(MathAch ~ Meanses*Cses + Sector*Cses, random = ~ Cses | School, data
=Math)
summary(Mathlme1)
```

```
## Linear mixed-effects model fit by REML
##  Data: Math
##        AIC      BIC    logLik
##   46523.66 46592.45 -23251.83
##
## Random effects:
##  Formula: ~Cses | School
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 1.5426082 (Intr)
## Cses        0.3181929 0.391
## Residual    6.0597961
##
## Fixed effects: MathAch ~ Meanses * Cses + Sector * Cses
##                          Value Std.Error   DF  t-value p-value
## (Intercept)         12.127931 0.1992913 7022 60.85529   0e+00
## Meanses              5.332875 0.3691672  157 14.44569   0e+00
## Cses                 2.945041 0.1556003 7022 18.92696   0e+00
## SectorCatholic       1.226579 0.3062723  157  4.00486   1e-04
## Meanses:Cses         1.039230 0.2988967 7022  3.47689   5e-04
## Cses:SectorCatholic -1.642674 0.2397796 7022 -6.85077   0e+00
##  Correlation:
##                     (Intr) Meanss Cses   SctrCt Mnss:C
## Meanses              0.256
## Cses                 0.075  0.019
## SectorCatholic      -0.699 -0.356 -0.053
## Meanses:Cses         0.019  0.074  0.293 -0.026
## Cses:SectorCatholic -0.052 -0.027 -0.696  0.077 -0.351
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.15926142 -0.72318922  0.01704599  0.75445035  2.95822019
##
## Number of Observations: 7185
## Number of Groups: 160
```

Излаз позива *summary* нашег хијеархијског линеарног модела за предвиђање броја постигнутих поена на задатом тесту из математике (*Mathlme1*) састоји се из неколико одељака.

1. Први одељак пружа информације о AIC-у (Akaike information criterion) и BIC-у (Bayesian information criterion) . Њихова примена даје значајан допринос приликом селекције модела, тачније одабира најбољег статистичког модела из скупа понуђених кандидата (модела), заједно са logLik.
2. Следећи одељак приказује процењене вредности варијансе и коваријансе *random* ефеката. Дакле, имамо следеће:
$$\hat{\varphi}_1 = 1.543 \quad \hat{\varphi}_2 = 0.318 \quad \hat{\sigma} = 6.06 \quad \hat{\varphi}_{12} = 1.543 \times 0.318 \times 0.391 = 0.192$$ [2]
3. Наредни одељак, то јест таблица $fixed$ ефеката, сличан је излазу за *lm*.
$$\hat{\beta}_1 = 12.13 \; \widehat{\beta}_2 = 5.33 \; \hat{\beta}_3 = 1.23 \quad \hat{\beta}_4 = 2.94 \quad \hat{\beta}_5 = 1.03 \; \hat{\beta}_6 = -1.64$$
4. Неке информације о стандардизованим резидуалима унутар групе ($\hat{\varepsilon}_{ij}/\hat{\sigma}$), броју опсервација, као и броју група приказане су при крају излаза.

Тестирање да ли су неки од елемената матрице коваријансе ω [3] једнаки нули може бити од интереса за одређене проблеме. Дакле, можемо тестирати хипотезе о варијансама и коваријансама random ефеката тако што ћемо термине random ефеката обрисати из претходног хијерархијског линеарног модела. [4] Морамо бити опрезни, те приликом оваквих поређења модела обратити пажњу на потребу за идентичношћу њихових фиксираних ефеката.

Модел 1 (*Mathlme1*)

$$MathAch_{ij} = \beta_1 + \beta_2 Meanses_i + \beta_3 Sector_i + \beta_4 Cses_{ij} + \beta_5 Meanses_i Cses_{ij} + \beta_6 Sector_i Cses_{ij} + b_{i1} + b_{i2} Cses_{ij} + \varepsilon_{ij}$$

Модел 2 (*Mathlme2*)

$$MathAch_{ij} = \beta_1 + \beta_2 Meanses_i + \beta_3 Sector_i + \beta_4 Cses_{ij} + \beta_5 Meanses_i Cses_{ij} + \beta_6 Sector_i Cses_{ij} + b_{i1} + \varepsilon_{ij}$$

---

[2] $\begin{bmatrix} {\varphi_1}^2 & \varphi_{12} \\ \varphi_{12} & {\varphi_2}^2 \end{bmatrix} = V\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} = \omega$ (матрица коваријансе)

[3] $\begin{bmatrix} {\varphi_1}^2 & \varphi_{12} \\ \varphi_{12} & {\varphi_2}^2 \end{bmatrix} = V\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} = \omega$ (матрица коваријансе)

[4] Доста пажљивија формулација ових тестова дата је у књизи *Hierarchical Linear Models: Applications and Data Analysis Methods, Raudenbush and Bryk, 2002*

```
Mathlme2 <- update(Mathlme1, random = ~ 1 | School)
anova(Mathlme1, Mathlme2)
```

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## Mathlme1     1 10 46523.66 46592.45 -23251.83
## Mathlme2     2  8 46520.79 46575.82 -23252.39 1 vs 2 1.124098    0.57
```

Модел 3 (*Mathlme3*)

$$MathAch_{ij} = \beta_1 + \beta_2 Meanses_i + \beta_3 Sector_i + \beta_4 Cses_{ij} + \beta_5 Meanses_i Cses_{ij} + \beta_6 Sector_i Cses_{ij} + b_{i2} Cses_{ij} + \varepsilon_{ij}$$

```
Mathlme3<- update(Mathlme1, random = ~ Cses - 1 | School)
anova(Mathlme1, Mathlme3)
```

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## Mathlme1     1 10 46523.66 46592.45 -23251.83
## Mathlme3     2  8 46740.23 46795.26 -23362.11 1 vs 2 220.5634  <.0001
```

Дакле, примећујемо да искључивање једног од датих *random* ефеката уклања не само његову варијансу из модела, него и његову коваријансу са другим *random* ефектом . Велика p −вредност у првом тесту нам сугерише да нема потребе за *random* коефицијентом уз променљиву *cSES*, то јест, другачије речено, $\varphi_{12} = {\varphi_2}^2 = 0$.Такође, мала p −вредност у другом тесту нам, пак, сугерише да ${\varphi_1}^2 \neq 0$. Уочимо да, с тога, просечна достигнућа на датом тесту из математике варирају од школе до школе. Претходно поменути модел *Mathlme2* изостављa незначајне, на основу велике p −вредности теста, *random* ефекте за *cSES*, а процењени *fixed* ефекти су готово идентични *fixed* ефектима нашег поченог модела (*Mathlme1*) који укључује те *random* ефекте.

Такође, на основу претходног, можемо закључити да укључивање *random intercept* ефекта значајно побољшава хијеархијски линеарни модел предвиђања броја постигнутих поена на датом тесту из математике .

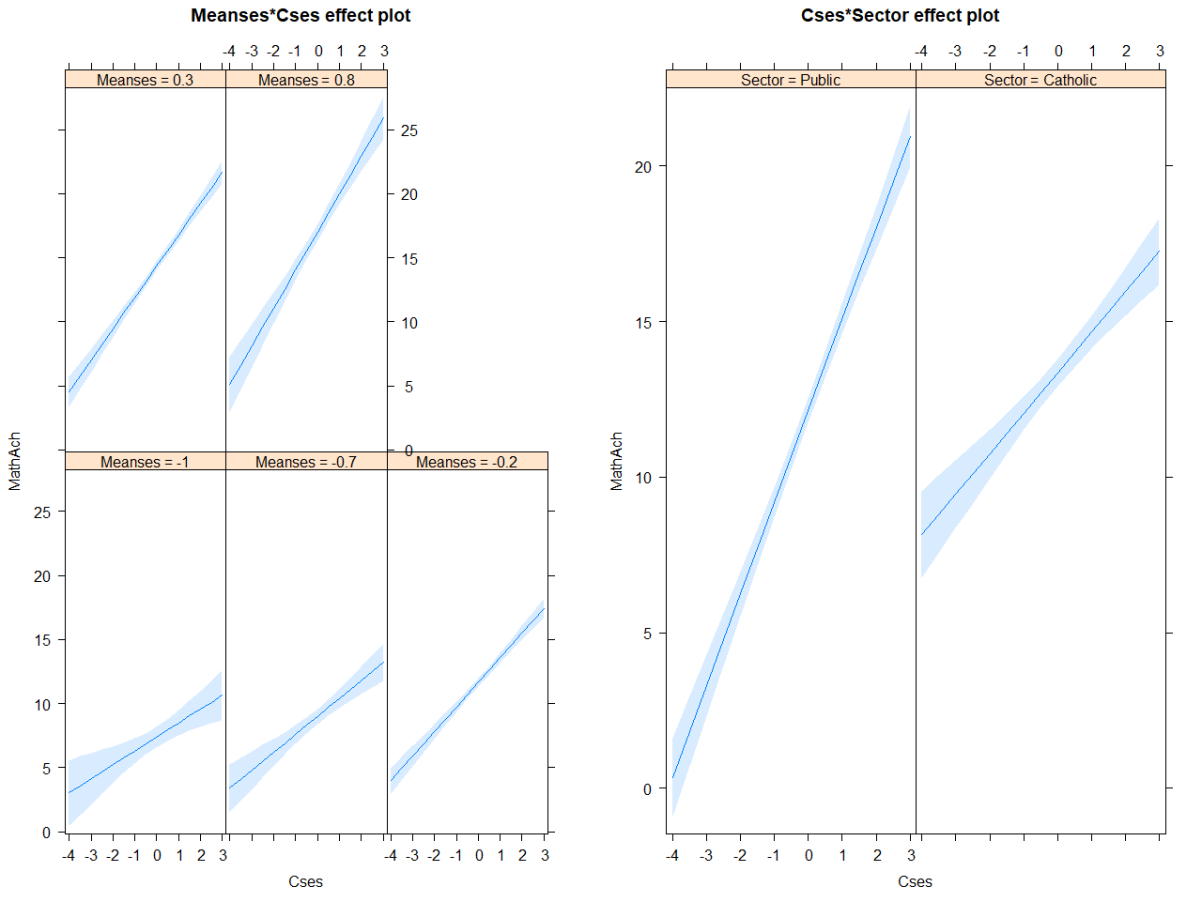
```
summary(Mathlme2)
```

```
## Linear mixed-effects model fit by REML
##  Data: Math
##        AIC      BIC    logLik
##   46520.79 46575.82 -23252.39
##
## Random effects:
##  Formula: ~1 | School
##         (Intercept) Residual
## StdDev:    1.541213 6.063503
##
## Fixed effects: MathAch ~ Meanses * Cses + Sector * Cses
##                         Value Std.Error   DF  t-value p-value
## (Intercept)         12.128207 0.1991998 7022 60.88462   0e+00
## Meanses              5.336670 0.3689849  157 14.46311   0e+00
## Cses                 2.942146 0.1512091 7022 19.45747   0e+00
```

```
## SectorCatholic       1.224529 0.3061193  157  4.00017   1e-04
## Meanses:Cses         1.044438 0.2910482 7022  3.58854   3e-04
## Cses:SectorCatholic -1.642156 0.2330937 7022 -7.04505   0e+00
##  Correlation:
##                     (Intr) Meanss Cses   SctrCt Mnss:C
## Meanses              0.256
## Cses                 0.000  0.000
## SectorCatholic      -0.699 -0.356  0.000
## Meanses:Cses         0.000  0.000  0.295  0.000
## Cses:SectorCatholic  0.000  0.000 -0.696  0.000 -0.351
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.17011509 -0.72487675  0.01484507  0.75424205  2.96551328
##
## Number of Observations: 7185
## Number of Groups: 160
```

Претходни модел (*Mathlme2*) је вероватно лакше визуализирати у ефектним парцелама , добијеним уз помоћ пакета *effects*.

```
library(effects)

## Loading required package: carData

## Use the command
##     lattice::trellis.par.set(effectsTheme())
##   to customize lattice options for effects plots.
## See ?effectsTheme for details.

plot(allEffects(Mathlme2,response="Cses",x.var="Cses"),rug=FALSE)
```



Позабавимо се и протумачимо детаљније податке о $fixed$ ефектима, добијене на излазу позива $summary$ (трећи одељак) нашег првобитног хијеархијског линеарног модела за предвиђање броја постигнутих поена на задатом тесту из математике ($Mathlme1$). Дакле, дати $fixed$ ефекти су сви значајно различити од нуле ($p < 0.001$). С обзиром да $Sector$ представља категоричку променљиву чија вредност означава припадност једној од неколико могућих категорија (вредност 0 означава припадност државном типу школе, а 1 католичком), уочимо следеће:

- **Public** : MathAch = 12.13 + 5.33 Meanses + 2.94 Cses + 1.03 Meanses* Cses ...
- **Catholic**: MathAch = 13.36 + 5.33 Meanses + 1.30 Cses + 1.03 Meanses* Cses ...

Такође, с обзиром да је променљива $Cses$ центрирана , а променљива $Meanses$ има средњу вредност нула, можемо приметимо да је просечан број поена на датом тесту из математике мањи у државним него у католичким школама, тачније просечан број поена на датом тесту у државним школама је 12.13, а католичким, пак , 13.36. Затим, просечни нагиб за $Cses$ у државним школама је већи него у католичким (приметимо исто тврђење и на претходном графичком приказу тј. ефектним парцелама). Исто тако , просечни нагиб за $Cses$ је већи у школама са бољим (већим) социоекономским статусом породица ученика ($Meanses$).

Такође, поред претходно коришћене функције *lme*, иста анализа може се извести и користећи функцију *lmer* из пакета *lme4*.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:nlme':
##
##     lmList
```

```
Mathlmer1 <- lmer(MathAch ~ Meanses*Cses + Sector*Cses + (Cses | School),data=Math)
summary(Mathlmer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: MathAch ~ Meanses * Cses + Sector * Cses + (Cses | School)
##    Data: Math
##
## REML criterion at convergence: 46503.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.15926 -0.72319  0.01704  0.75445  2.95822
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  School   (Intercept)  2.3796  1.5426
##           Cses         0.1012  0.3181   0.39
##  Residual             36.7212  6.0598
## Number of obs: 7185, groups:  School, 160
##
```

```
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          12.1279     0.1993  60.856
## Meanses               5.3329     0.3692  14.446
## Cses                  2.9450     0.1556  18.927
## SectorCatholic        1.2266     0.3063   4.005
## Meanses:Cses          1.0392     0.2989   3.477
## Cses:SectorCatholic  -1.6427     0.2398  -6.851
##
## Correlation of Fixed Effects:
##             (Intr) Meanss Cses   SctrCt Mnss:C
## Meanses      0.256
## Cses         0.075  0.019
## SectorCthlc -0.699 -0.356 -0.053
## Meanses:Css  0.019  0.074  0.293 -0.026
## Css:SctrCth -0.052 -0.027 -0.696  0.077 -0.351
```

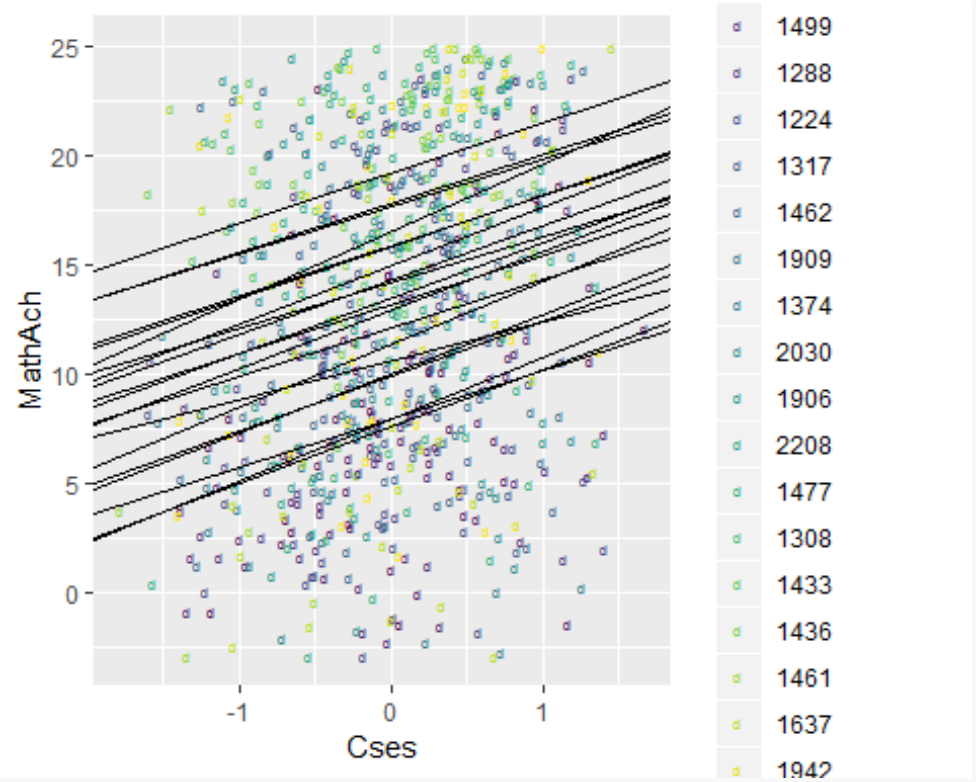
Процењене вредности *fixed* ефеката, као и варијансе и коваријансе *random* ефеката поклапају се са одговарајућим вредностима добијеним функцијом *lme*.

Међутим, сама спецификација датог модела се мало разликује. Тачније, уместо коришћена аргумента *random* за дефинисање *random* ефеката, као што је случај са *lme, random* ефекати дати су нам директно у моделној формули затворени заградом.

У наставку, приказаћемо још одређене графичке приказе и истраживања која су нам била од користи приликом доношења одлуке о коришћењу хијерархијског линеарног модела за предвиђање броја постигнутих поена на датом тесту из математике.

✓

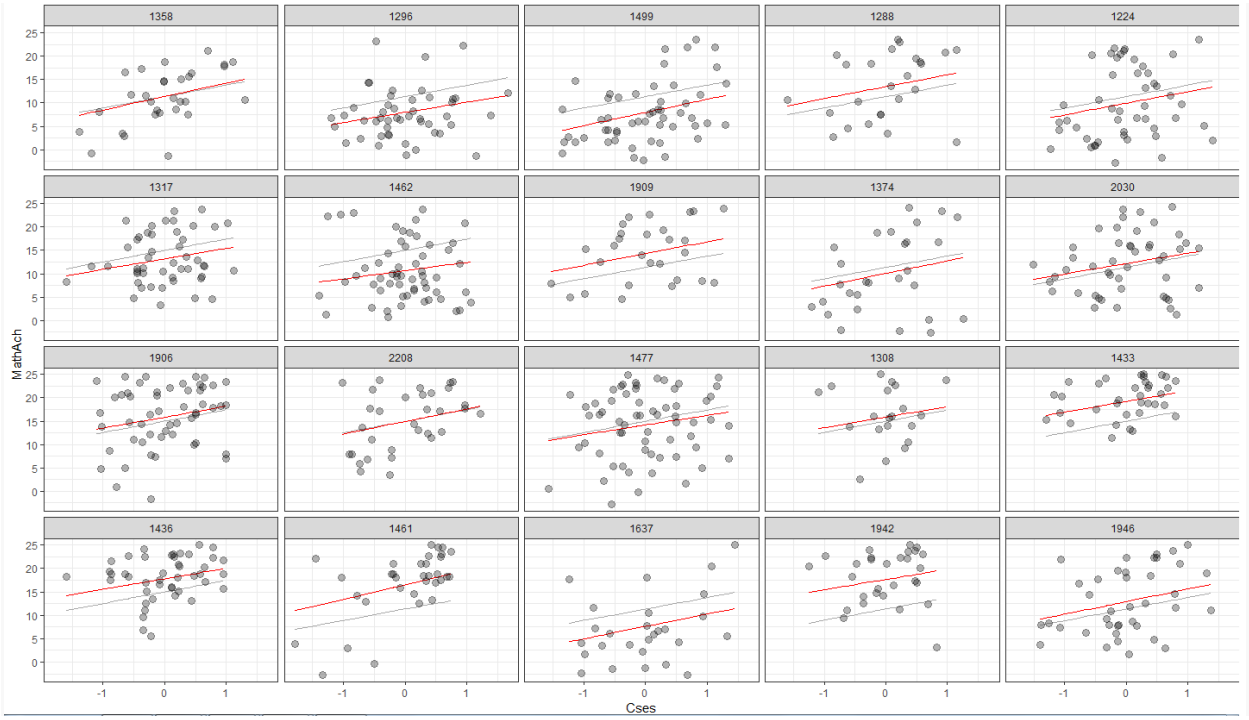
```
library(ggplot2)
Math2<-Math[1:785,]
Math2lmer<-lmer(MathAch ~ Cses + (1 + Cses | School), data=Math2)
intercepts <- coef(Math2lmer)$School[,1]
slopes <- coef(Math2lmer)$School[,2]
ggplot(Math2, aes(x=Cses, y=MathAch, color=as.factor(School))) +
  geom_point(shape=100) +
  geom_abline(slope=slopes, intercept=intercepts)
```



На графичком приказу дате су регресионе праве за првих двадесет школа. Приметно је да се коефицијент пресретања разликује за сваку њих, што је, управо, знак да је, у овом случају, паметнији избор хијерархијског линеарног модела у односу на обични линеарни модел.

✓

```
Mathlm<- lm(MathAch ~ Cses + Sector, Math2)
Math2$Predictions <- fitted(Mathlm)
Math2$MLPredictions <- fitted(Math2lmer)

ggplot(Math2, aes(x=Cses, y=MathAch, group=School)) +
  geom_line(aes(y = Math2$Predictions), color = "darkgrey") +
  geom_line(aes(y = Math2$MLPredictions), color = "red") +
  geom_point(alpha = 0.3, size = 3) +
  facet_wrap(~School) +
  theme_bw()
```



На графичком приказу дате су регресионе праве из хијерархијског (црвене) и обичног (сиве) линеарног модела. Уочимо да код неких школа (на пример: 1433, 1461 ...) праве обичне линеарне регресије јако лоше описују модел, док се, пак, оне из хијерархијског, много боље понашају и описују дати модел.

✓

```
test<- Math[c(1:9, 48:56, 73:81,121:129,141:149,189:197,219:227,247:255,282:29
0,326:334),]
train<- Math[ -c(1:9, 48:56, 73:81,121:129,141:149,189:197,219:227,247:255,282
:290,326:334,358:7185),]
Mlmer<-lmer(MathAch ~ Meanses*Cses + Sector*Cses + (Cses | School), data=train
)

## singular fit

Mlm<- lm(MathAch ~ Cses + Sector, train)
global_pred<- predict(Mlm, newdata=test)
global_MSE <- mean((test$MathAch - global_pred)^2)
global_MSE

## [1] 38.38926

mlm_pred <-predict(Mlmer, newdata=test, allow.new.levels = TRUE)
mlm_MSE<- mean((test$MathAch - mlm_pred)^2)
mlm_MSE

## [1] 33.72613
```

Девет ученика из сваке од првих десет школа узето је за тестирање, док је преостало узето у обзир за тренирање. Приметимо да је *mean squared error* мања за хијерархијски линеарни модел. Дакле, можемо да рећи да хијерархијски линеарни модел у просеку боље предвиђа број постигнутих поена на задатом тесту из математике.

Литература

- Raudensbush, S.W. and Bryk, A.S. (2002). Hierarchical Linear Models, Sage Publications

- http://www.statstutor.ac.uk/resources/uploaded/multilevelmodelling.pdf?fbclid=IwAR0K71v4tUoIvoQygjzPgCSt7jSPFTZ5v3L9x-QT3tHFkx3Dgi8y0DzGT_8

- http://www.stats.ox.ac.uk/~snijders/MLB_new_S.pdf?fbclid=IwAR209Mmc6XmdtSbaZsK7JSqm7mtDaPWa60u6LAM6HTvdz7MuPMi6OuclGK0

- https://www.ida.liu.se/~732G34/info/singer.pdf?fbclid=IwAR22AeZNLPGr1A_-xIxMCPW2TivQquy5sa7J9ugcj8Xrfxl_ncyi3C4VD-I

- https://www.researchgate.net/publication/265226054_Hierarchical_Linear_Models_in_Education_Sciences_an_Application

- http://www.tqmp.org/RegularArticles/vol08-1/p052/p052.pdf?fbclid=IwAR0-6w7UOJU793npz6srx4olH1VNluDR5d8yan7F8mEHcp3AREvM7s4Ukrs