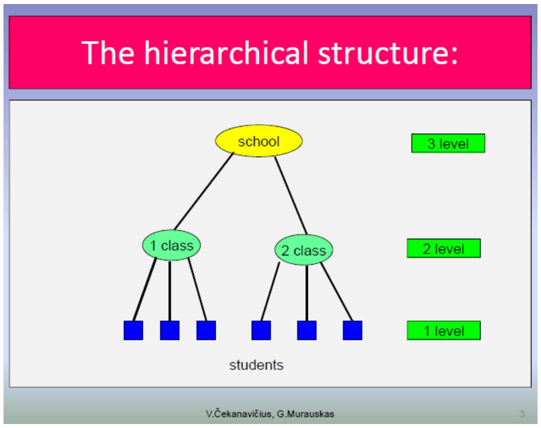
Hierarchical linear model

M. Pieškutė

What is HLM?

- Mathematically HLM is the generalization of a linear regression to the case, when regression coefficients are random variables.
- Practically HLM means that we take into account the hierarchical structure of data.

The hierarchical structure



The hierarchical structure:

V.Čekanavičius, G.Murauskas 3

- The second-level characteristics (teacher‘s experience, number of students) apply to the whole class.

Remarks

- We must have a sufficient number of observations for all levels of data. For example, it is not allowed to construct HLM just for two schools.
- The majority of observed variables are normal.
- It is allowed to include a few categorical variables (dummy variables).

Typical sequence of HLM constructing

- Unconditional Model.
- A few more complex models.
- Comparison of models.
- The choice of acceptable model or conclusion that such a model can not be obtained.

HLM with R

Unconditional HLM model

Unconditional model allows to check if the hierarchical structure should be taken into account. It also allows us to numerically estimate the importance of the second level variables.

Data and model:

- 8‘th grade students from Vilnius.
- Number of students: 559, number of schools: 27.
- Dependent variable: test score for mathematics.
- Regression equations are written for both – student‘s and school‘s – levels.

The idea:

- For the first (student‘s) level: **Student‘s achievement = school‘s mean + Individual differences from school‘s mean**
- For the second (school‘s) level: **School‘s score = mean score for all school‘s + differences from that mean**
- Student‘s level: $MAT = \beta_0 + e$
- School‘s level: $\beta_0 = \gamma_{00} + u_0$
- Pooled model: $MAT = \gamma_{00} + u_0 + e$
- Parameter γ_{00} is equal to the common mean score for all schools.
- The differences of schools is reflected by the magnitude of the variance τ_{00} of the variable u_0.
- The individual differences among students are reflected by the magnitude of the variance σ^2 of the residual e
- Larger variances mean bigger differences.

Some notation

- γ_{00}- is called the fixed effect paramete,
- τ_{00} and σ^2 - are the random effect parameters.

Unconditional(zero) model with R

Data

- Data has the traditional structure: one respondent - one case.
- Moreover we have variable *IDMOK*, which contains school‘s code .

```r
library(dplyr)
library(foreign) ## for reading data
library(knitr)
library(afex)
library(nlme) ## for lme(...)
library(msm) ## for deltamethod(...)
library(car) ## Anova(...)

data.hlm <- read.spss("stat3hlm1.sav",
                  to.data.frame = TRUE,
                  use.value.labels = TRUE,
                 strings.as.factors = TRUE)

#head(data.hlm) ## see first 6 rows of data
#lapply(data.hlm, FUN= class) ## check the type of data in every column of dataframe
data.hlm$IDKLA<-factor(data.hlm$IDKLA)
```

Model syntax and summary

```r
## with nlme package
model.0<-lme(MAT ~1, data=data.hlm, random= ~1|IDMOK)
summary(model.0)
```

```
## Linear mixed-effects model fit by REML
##  Data: data.hlm
##        AIC      BIC   logLik
##   6421.181 6434.154 -3207.59
##
## 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
```

```r
gamma_00 <- data.frame(fixef(model.0))[1,1] ## extract the fixed effect coefficient
```

Estimates

Estimate for parameter γ_{00}: $\hat{\gamma_{00}} = 499.4717141$

Estimates for the two random parameters are also given and they are large : $\hat{\tau_{00}} \approx 48.54^2 \approx 2356$ and $\hat{\sigma^2} \approx 71.53^2 \approx 5117$. We are squaring the result, because we are gived the standard deviation and SPSS output contains dispersion. The model should be improved.

Hypothesis testing

Standard hypothesis is checked:

$$\begin{cases} H_0: & \gamma_{00} = 0. \\ H_1: & \gamma_{00} \neq 0. \end{cases}$$

```r
Anova(model.0, type = "III", test = "Chisq")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: MAT
##              Chisq Df Pr(>Chisq)
## (Intercept) 2557.4  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since $p < 0.05$, we conclude that $\gamma_{00} \neq 0$ statistically significantly .

- Recalling that γ_{00} is the mean math score for all schools, we see that our statistical hypothesis is not very important – nobody expected that mean to be zero.
- The estimate $\hat{\gamma_{00}} = 499.4717141$ is more informative (especially if we know the maximum possible score of the test).

Two hypotheses about the variances for error terms of both levels are tested:

$$\begin{cases} H_0: & \tau_{00} = 0. \\ H_1: & \tau_{00} > 0. \end{cases}$$

$$\begin{cases} H_0: & \sigma^2 = 0. \\ H_1: & \sigma^2 > 0. \end{cases}$$

```r
## how to get  std. error estimates that SPSS produces
var <-model.0$apVar
par<-attr(var, "Pars")
vc<-exp(par)^2

## we get the std. error estimates that SPSS produces
int.var.st.err <- deltamethod (~ exp(x1)^2, par, var)
resid.var.st.err <- deltamethod (~ exp(x2)^2, par, var)

wald.z.r <- vc[2]/resid.var.st.err
wald.z.int <- vc[1]/int.var.st.err

## p-values

p.r <- 1 - pnorm(wald.z.r)
p.int <- 1 - pnorm(wald.z.int)
p.r
```

```
## lSigma
##      0
```

```
p.int
```

```
## reStruct.IDMOK
##   0.0006930289
```

Both zero hypotheses are rejected, because p-value0<0.05 and $0 < 0.05$ and we conclude that τ_{00} abd σ^2 are >0.

We can also get confidence intervals for the estimates with the function *intervals(model.0)*

ICC

Intraclass correlation coefficient shows the impact of the second level variables on the results:

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

The large ICC indicates that the hierarchical structure of data should be taken into account.

In our example $ICC = \frac{2356}{2356+5117} = 0.315...$

- ICC can be intrepretated as the proof that $31,5$ percent of student's math. score can be attributed to the differences among schools.
- Conclusion: hierarchical structure must be taken into account and more complex HLM should be constructed.
- We seek to reduce estimates $\hat{\tau_{00}} = 2356$ and $\hat{\sigma^2} = 5117$ since these estimates reflect everything that can have impact on the scores but is not included in the model.

Information indices

In R we can easily get two information indeces - **AIC**(Akaike) and **AICC**(Hurvich and Tsai).

These indices are used for the comparison of two models (so far we have unconditional model only). The better model has **smaller** indices. Which index to use decides the researcher himself.

So far we have just one unconditional model. The information indices will be used later for comparisons.

In R we can get information indices this way:

```
library(AICcmodavg)
AIC(model.0)
```

```
## [1] 6421.181
```

```
AICc(model.0) ## corrected for small samples
```

```
## [1] 6421.224
```

HLM model with the second level interval variable

Suppose that:

- Student‘s math. test score depends on his/her social-economic status (CSES);
- School‘s impact is the same for all students attending that school;
- School‘s impact depends on the mean value of the social-economic status of its students (MSES).

Model assumptions:

- Dependent variable (MAT) is interval (normal)
- Variables *CSES*, *MSES* are interval (normal).
- Error terms (residuals) – are normal and, for the first level, independent.

Next we assign variables to levels:

- The first (student‘s) level variable: *CSES*.
- The second (school‘s) level variable: *MSES*.
- Student‘s level model equation: $MAT = \beta_0 + \beta_1 CES + e$
- Here e is a random error, which is normal with mean zero and variance σ^2. This variance reflects individual differences among students, which **are not** caused by the differences in the social-economic status or by the impact of MSES for schools.
- The second level variable MSES can be included into the model for β_0, if we assume that MSES has the same (constant) impact for each student of the corresponding school. β_1, if we assume that MSES has interraction with CSES.
- Let us assume that there is no interaction with CSES: $\beta_0 = \gamma_{00} + \gamma_{01} MES + u_0$, $\beta_1 = \gamma_{10} + u_1$. Here γ_{ij} are unknown constants and residual error terms u_0, u_1 have unknown variances τ_{00}, τ_{11} and covariance τ_{01}
- Large values of τ_{00}, τ_{11} mean that there are other than MSES characteristics of school which can have important impact on students results (that is our model is not very good).
- Small variances mean that all important characteristics of school are included in the model.
- It is recommended to write the pooled model equation:

$$MAT = \gamma_{00} + \gamma_{01} MSES + \gamma_{10} CSES + e + u_0 + u_1 CSES$$

- In the model we distinguish between the fixed effect component and the random effect component.
- Fixed effect parameters γ_{00}, γ_{01}, γ_{10}
- Random effect parameters are variances and covariances of e, u_0, u_1, that is τ_{00}, τ_{11}, τ_{01}, σ^2
- Variables with γ coefficients are called fixed effects variables (fixed effects). Variables multiplied by u are called random effects variables. The same variable can belong to both classes of variables. In our model:
- Fixed effect variables: CSES ir MSES.
- Random effect variable CSES. $MAT = \gamma_{00} + \gamma_{01} MSES + \gamma_{10} CSES + e + u_0 + u_1 CSES$

R

```
data.hlm.n<- na.omit(data.hlm)
model.1<-lme(MAT~1+MSES+CSES, data=data.hlm.n, random= ~1+CSES|IDMOK)

summary(model.1)
```

```
## Linear mixed-effects model fit by REML
##  Data: data.hlm.n
##        AIC      BIC    logLik
##   6123.83 6153.871 -3054.915
##
## Random effects:
##  Formula: ~1 + CSES | IDMOK
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 23.653305 (Intr)
## CSES         3.031702 0.005
## Residual    65.505117
##
## Fixed effects: MAT ~ 1 + MSES + CSES
##                 Value Std.Error  DF  t-value p-value
## (Intercept) 175.94264  42.63200 515 4.127008       0
## MSES         16.56188   2.13507  25 7.757060       0
## CSES          8.04501   1.22706 515 6.556340       0
##  Correlation:
##      (Intr) MSES
## MSES -0.992
## CSES  0.000  0.000
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.8198281 -0.6403279 -0.0157316  0.6064878  2.9770521
##
## Number of Observations: 543
## Number of Groups: 27
```

We will

- compare AIC and other information indices with their counterparts from unconditional (zero) model;
- check which fixed effect parameters are statistically significant;
- check which random effect parameters are statistically significant;
- check for changes in variances.

The rule-of the thumb: the change in Akaike index exceeding 10 points is assumed to be important.

For the comparison of two models we need to use the **Maximum Likelihood** estimates, so we update the model.

```r
model.1ml<-update(model.1, method = "ML")
model.0ml<-update(model.0, method = "ML", data = data.hlm.n)
anova(model.1ml, model.0ml)
```

```
##           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model.1ml     1  7 6134.547 6164.627 -3060.273                        
## model.0ml     2  3 6217.295 6230.186 -3105.647 1 vs 2 90.74783  <.0001
```

And from the result above we can see that the drop in AIC is more than 10 points.

```r
anova(model.1)
```

```
##             numDF denDF  F-value p-value
## (Intercept)     1   515 8628.224  <.0001
## MSES            1    25   60.147  <.0001
## CSES            1   515   42.986  <.0001
```

```r
gamma_00 <- data.frame(fixef(model.1)[1])[1,1]
gamma_01 <- data.frame(fixef(model.1)[2])[1,1]
gamma_10 <- data.frame(fixef(model.1)[3])[1,1]
```

All fixed effect parameters are significant.

- $\hat{\gamma_{00}} = 175.9426402$,
- $\hat{\gamma_{01}} = 16.5618791$,
- $\hat{\gamma_{10}} = 8.0450104$.

Estimates for random effect parameters:

- $\hat{\tau_{00}} \approx 23.653^2 \approx 559.476$,
- $\hat{\tau_{11}} \approx 3.031^2 \approx 9.19$,
- $\hat{\sigma^2} \approx 65.505^2 \approx 4290.921$,

We also see that the first level residual's (which reflects individual differences among students) is not small: $\hat{\sigma^2} = 4290.92$. However, comparing it with analogous estimate for zero model $\hat{\sigma^2} = 5117.19$ we see that now it is much smaller.

More precisely $\frac{\hat{\sigma^2_{old}}-\hat{\sigma^2_{new}}}{\hat{\sigma^2_{old}}} = \frac{5117.19-4290.92}{5117.19} = 0.161...$

Interpretation: in comparison to zero model, the differences among students unexplained by the model are reduced by 16%.

We can state that the second model is much better than the unconditional (zero) model. Results are measured in hundreds of points. Therefore, estimate $\hat{\sigma} = 65.49$ is not very large and we conclude that the second model is acceptable. This does not mean that better models are unavailable.

Forecasting

We use pooled model with fixed effect parameter estimates $MAT = \gamma_{00} + \gamma_{01}MSES + \gamma_{10}CSES$.

For our example: $MAT = 175.94 + 16.56MSES + 8.04CSES$

This pooled model equation allows to estimate the impact of MSES. More precisely, each additional point for MSES score adds 16.56 to the student result.

Categorical regressor

We will add the second level categorical variable VK, which shows if the school is from Vilnius.

- The First level equation $MAT = \beta_0 + \beta_1 CSES + e$
- The Second level equation depends on school's social-economic status and location :$MAT = \gamma_{00} + \gamma_{01}MSES + \gamma_{12}VK + u_0$
- For Vilnius the social-economic status is different from other schools $\beta_1 = \gamma_{10} + \gamma_{11}VK + u_1$
- Here e has variance σ^2 .
- Residuals u_0, u_1 have variances τ_{00}, τ_{11} and can be correlated (their covariance is τ_{01}).
- Pooled model equation:

$$MAT = \gamma_{00} + \gamma_{01}MSES + \gamma_{10}CSES + \gamma_{02}VK + \gamma_{12}VK * CSES + [u_1CSES + u_0 + e]$$

- Fixed variables: CSES, MSES, VK, VK x CSES .
- Random variables: CSES.

We will

- check which fixed effect parameters are statistically significant;
- check for changes in variances.
- The rule-of the thumb: the change in Akaike index exceeding 10 points is assumed to be important.

```
model.2<-lme(MAT ~1+CSES+MSES+VK+VK*CSES,data=data.hlm.n, random = ~1+CSES|IDMOK)
summary(model.2)
```

```
## Linear mixed-effects model fit by REML
##  Data: data.hlm.n
##        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) 19.524824 (Intr)
## CSES         3.205892 0.146
## Residual    65.528515
##
```

```
## Fixed effects: MAT ~ 1 + CSES + MSES + VK + VK * CSES
##                    Value Std.Error  DF   t-value p-value
## (Intercept)     47.23318  60.83742 514  0.776384  0.4379
## CSES             7.27030   1.55839 514  4.665277  0.0000
## MSES            23.93195   3.31679  24  7.215397  0.0000
## VKVilnius      -47.92946  17.50487  24 -2.738064  0.0115
## CSES:VKVilnius   2.07155   2.59352 514  0.798740  0.4248
##  Correlation:
##                (Intr) CSES   MSES   VKVlns
## CSES           -0.006
## MSES           -0.995  0.012
## VKVilnius       0.787 -0.029 -0.825
## CSES:VKVilnius  0.006 -0.601 -0.009  0.042
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.86523256 -0.64336852 -0.01344608  0.60060364  3.06657433
##
## Number of Observations: 543
## Number of Groups: 27
```

```
Anova(model.2, type = "III", test = "Chisq") ## tests of fixed effects
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: MAT
##               Chisq Df Pr(>Chisq)
## (Intercept)  0.6028  1    0.43752
## CSES        21.7648  1  3.082e-06 ***
## MSES        52.0620  1  5.378e-13 ***
## VK           7.4970  1    0.00618 **
## CSES:VK      0.6380  1    0.42444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not all fixed effect parameters are statistically significant.

For the comparison of two models we need to use the **Maximum Likelihood** estimates, so we update the model.

```
model.1ml<-update(model.1, method = "ML")
model.2ml<-update(model.2, method = "ML")
anova(model.1ml, model.2ml)
```

```
##           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model.1ml     1  7 6134.547 6164.627 -3060.273
## model.2ml     2  9 6130.467 6169.141 -3056.233 1 vs 2 8.080262  0.0176
```

We see that this model is not very good. Probably one should remove VK*CSES from the model. Perhaps one should drop γ_{00}

Important note

If we have the first level regressor which is strongly correlated to the dependent variable, then frequently no HLM is available. For example, by including the score for physics in the model $MAT = \beta_0 + \beta_1 FIZ + e$ we will not be able to construct meaningful HLM.

Recomendations

- Begin from zero model.
- Decide which variables belong to the first level and which – to the second level.
- Write equations for both levels. Do not forget to add e to the first level equation and u to the second level equations.
- Substitute the second level equations into the first level equation obtaining the pooled model.
- Collect all summands in the pooled model with u and e.
- Variables without u are fixed effect variables. Variables with u are random effect variables.
- Note that the same variable can be fixed and random variable.