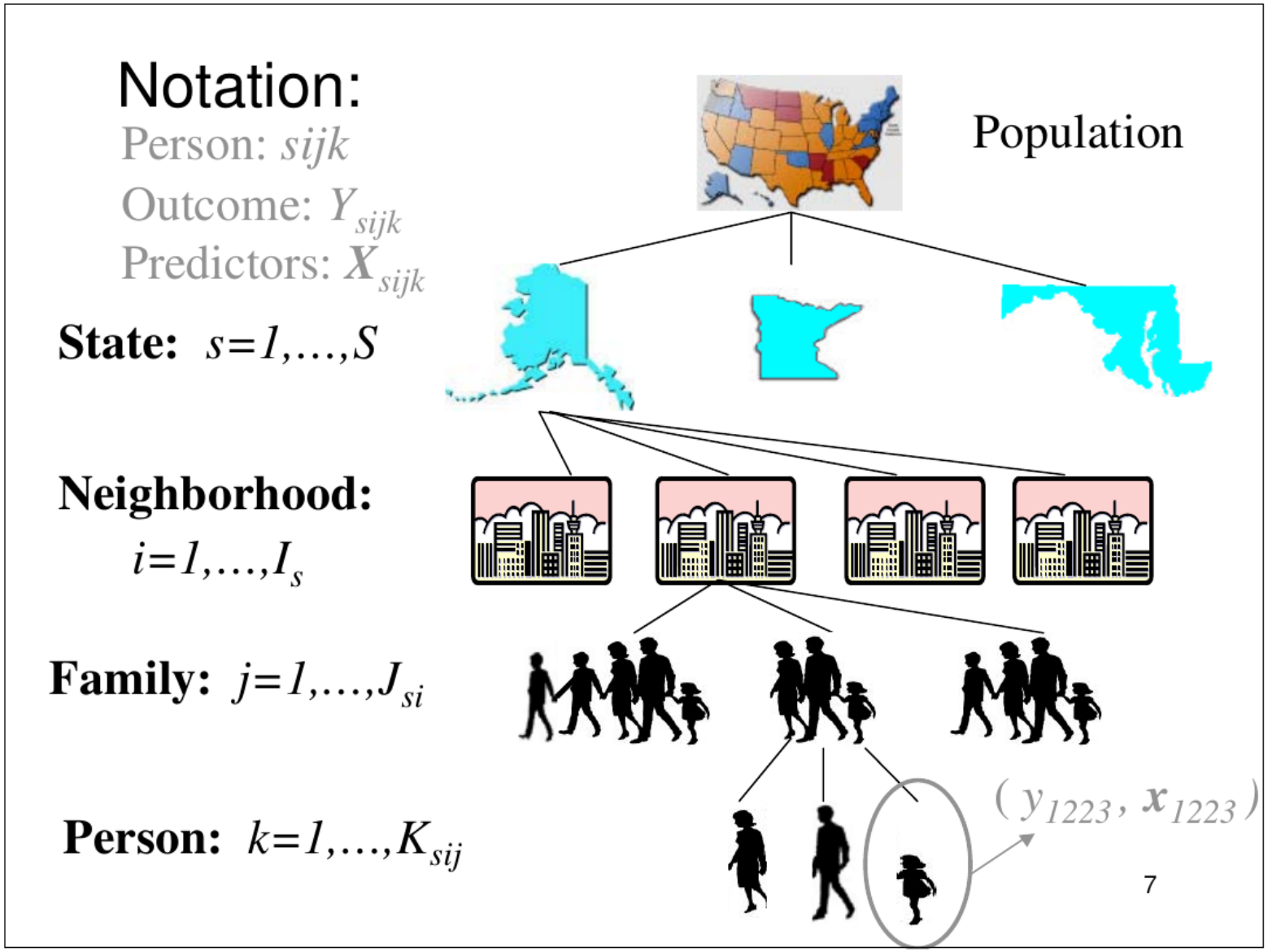
Matematičke metode u marketingu
Hijerarhijski linearni modeli
(Mixed models)

Lavoslav Čaklović
PMF-MO

2016

Sadržaj


Notation:
Person: $sijk$
Outcome: Y_{sijk}
Predictors: X_{sijk}
State: $s=1,\ldots,S$
Neighborhood:
$i=1,\ldots,I_s$
Family: $j=1,\ldots,J_{si}$
Person: $k=1,\ldots,K_{sij}$
Population
(y_{1223}, x_{1223})
7

Hijerarhijska struktura podataka

1. Učenici grupirani po razredima (učiteljima) u školama.
2. Pacijenti grupirani po doktorima unutar klinika.
3. Ponovljena mjerenja pojedinaca grupiranih po tretmanima (medicina, sport).

Zanima nas utjecaj *drugog nivoa* (škola, klinika, tretman) na uspjeh (rezultat) pojedinca (random efekt).

Također se govori o *ugnježđenoj* strukturi podataka.

Primjeri varijabli:

Razina-1: spol (starost) učenika u razredu [ono po čemu se razlikuju učenici]

Razina-2: tip (veličina) škole [ono po čemu se razlikuju škole]

Obrada ugnježđenih podataka

Pristupi: (1) *razjedinjavanje*[1] *(R)*, (2) *ujedinjavanje*[2] *(U)*, (3) *Hijerarhijski Linearni Modeli (HLM)*

(R) Varijabla iz viših razina pridjeljuje objektima 1. razine vrijednosti pomoću usrednjavanja. Gubi se nezavisnost (varijabli).

(U) Ignoriraju se razlike unutar grupe. Individua postaje homogeni entitet. Odzivna varijabla se agregira (upitna interpretacija)

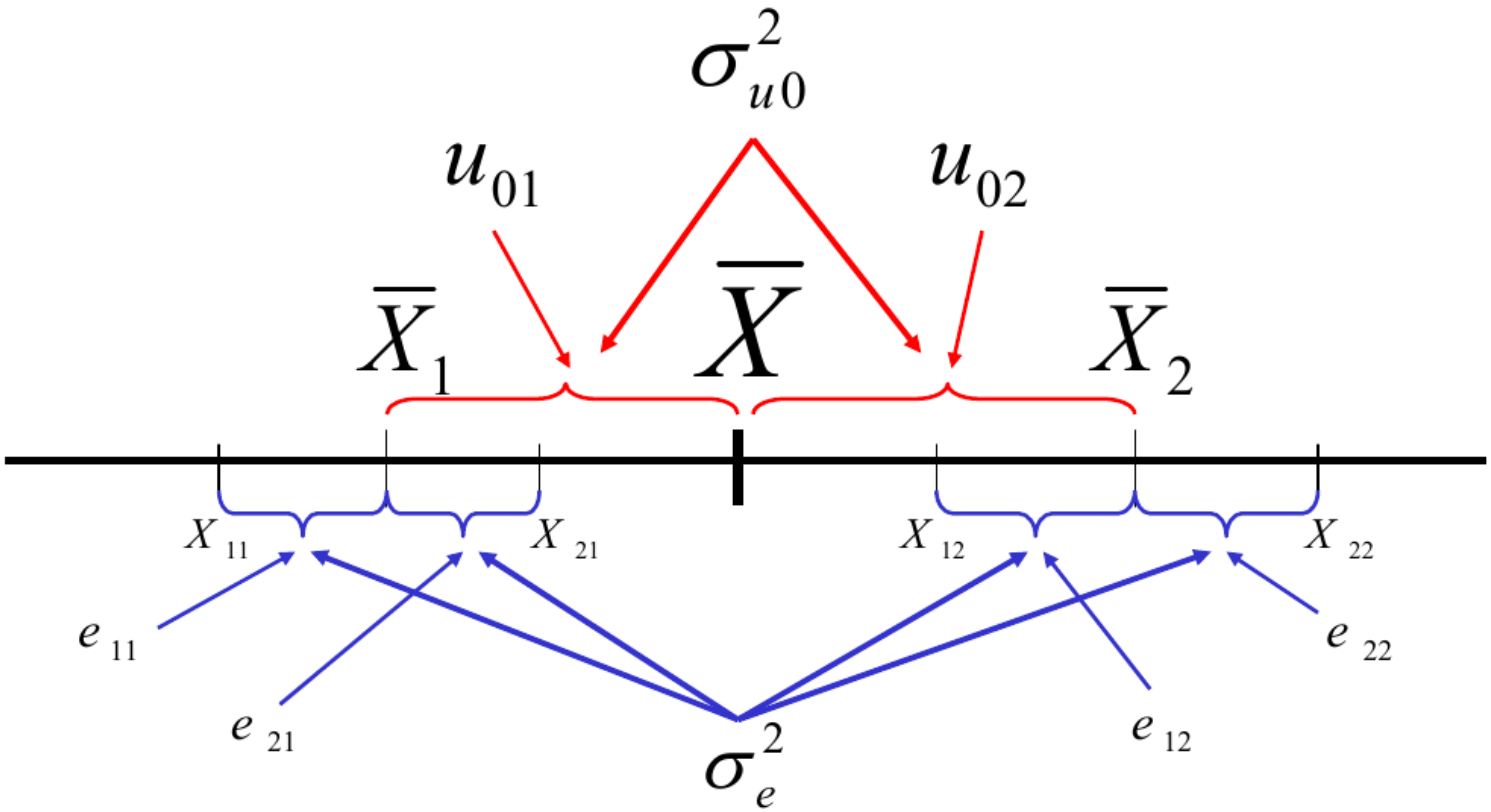
(HLM) uočava i individualne i grupne efekte u odzivnoj varijabli (varijabla 1. razine).

HLM je grupni naziv za više (sličnih) metoda koje u osnovi počivaju na metodi najmanjih kvadrata.

Sinonimi: *{multilevel, mixed level, mixed effects, random effects}* – modeling.

[1] eng. *desaggregation*
[2] eng. *aggregation*

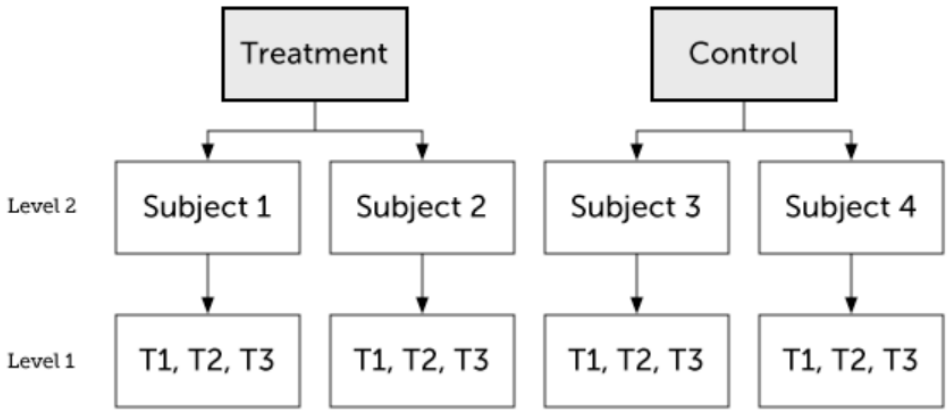
Organizacija podataka



Tipični ulazni podaci (*data.frame*):

y	*id*	*tretman*	*terapeut*	*t*
10	1	*A*	1	0
12	1	*A*	1	1
14	1	*A*	1	2
4	2	*A*	1	0
14	2	*A*	1	1
3	2	*A*	1	2
13	3	*A*	2	0
12	3	*A*	2	1
15	3	*A*	2	2

id – osoba, *t* – pon. mjerenje

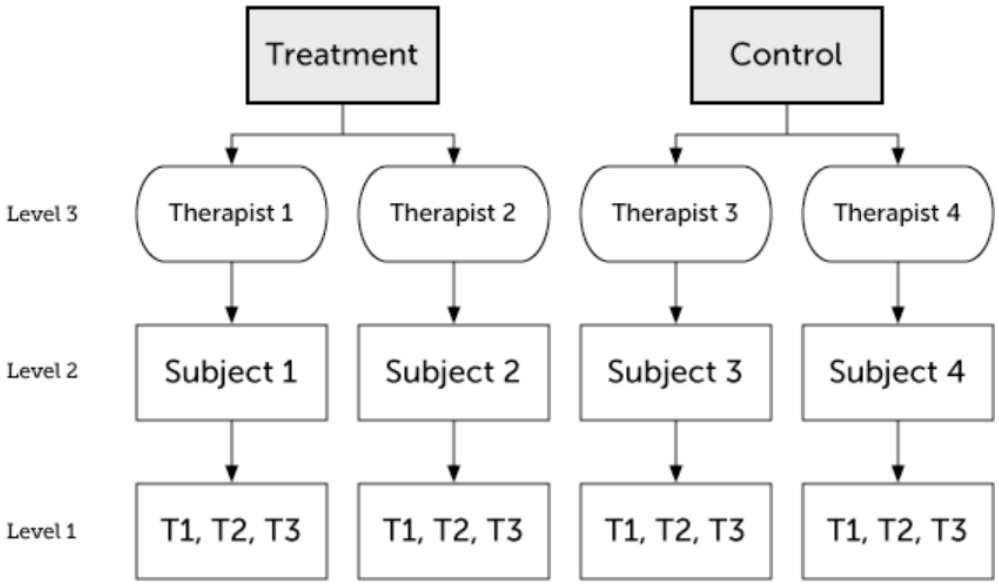


Dva nivoa

Tipični ulazni podaci (*data.frame*):

y	*id*	*tretman*	*terapeut*	*t*
10	1	*A*	1	0
12	1	*A*	1	1
14	1	*A*	1	2
4	2	*A*	1	0
14	2	*A*	1	1
3	2	*A*	1	2
13	3	*A*	2	0
12	3	*A*	2	1
15	3	*A*	2	2

id – osoba, *t* – pon. mjerenje



Tri nivoa

Dvije razine. Mjerenje 'unutar pojedinca'.

Bazni model (bez uvjeta):

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}, \quad \gamma_{00} - \text{grand mean}$$

$j-$indeks grupe, $i-$ indeks podatka u grupi

razina 1:

$$y_{ij} = \beta_{0j} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2)$$

$\beta_{0j} -$ sredina grupe

razina 2:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim \mathcal{N}(0, \tau_{00}^2)$$

```
# lme4
lmer(y ~ 1 + (1 | id), data=data)

# nlme
lme(y ~ 1, random = ~ 1 | id, data=data)
```

Zabavni park

Podaci u tablici rezultati ankete provedene među posjetiteljima zabavnog parka. Varijable su: `id` – posjetitelj, `rating` – ocjena profila, `speed`, `height`, `const`, `theme` – profil usluga.

ix	id	rating	speed	height	const	theme
1	1	4	40	300	Steel	Dragon
2	1	3	50	200	Wood	Dragon
3	1	5	50	300	Wood	Dragon
4	1	6	60	300	Wood	Eagle
5	1	4	60	400	Wood	Eagle

Tablica: Rezultati ankete profila usluga.

Uprava parka želi reorganizirati sadašnju ponudu kako bi privukla što više posjetitelja.

Zabavni park – nastavak

Odzivna varijabla je `rating` definirana na skupu indeksa. Varijabla 2. razine je `id`, ostale varijable su prediktori. Model je:

```
fit <- lmer(rating ~ speed + height + const + theme +
          (speed + height + const + theme | id), data=df)
```

Bazni model: interpretacija lmer() outputa

```
Linear mixed model fit by REML ['lmerMod']
Formula: rating ~ 1 + (1 | id)
   Data: df

REML criterion at convergence: 15649.3

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.07422  0.2724
 Residual             7.71075  2.7768
Number of obs: 3200, groups:  id, 200

Fixed effects:
            Estimate Std. Error t value
(Intercept)  5.26750    0.05273   99.89
```

Bazni model: interpretacija `lmer()` outputa – nastavak

```
Linear mixed model fit by REML ['lmerMod']
Formula: rating ~ 1 + (1 | id)
   Data: df
```

Kratice `REML` stoji za *Restricted Maximul Likelihood*. `lmerMod` je jedan od modula u `lmer()` funkciji.

Odzivna varijabla *rating* definirana je na 1. razini koja se 'spominje' u formuli. Varijabla *id* je faktor i njeni nivoi predstavljaju osobe na kojima se vrši 'ponovljeno mjerenje'[3]. Broj `1` u formuli predstavlja *intercept*.

Unutar zagrade $(1 \mid id)$ je slučajni (grupni) prediktor koji je ovdje konstantan unutar grupe (oznaka 1). Izvor podataka je *data.frame* objekt `df`.

[3] Inače, to su vrijednosti odzivne varijable na entitetima 1. razine.

Bazni model: interpretacija `lmer()` outputa – nastavak

```
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.07422  0.2724
 Residual             7.71075  2.7768
Number of obs: 3200, groups:  id, 200
```

Stupac *Variance* pokazuje varijabilnost odziva po grupama i neobjašnjeni dio koji iznosi 7.71075.

Broj observacija je 3200, a broj grupa 200, što se vidi iz posljednjeg retka.

Intra Class Correlation faktor (ICC) je postotak objašnjene varijance i iznosi

$$ICC = \frac{0.07422}{0.07422 + 7.71075} = 0.009534323.$$

Bazni model: interpretacija `lmer()` outputa – nastavak

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  5.26750    0.05273   99.89
```

Tablica fiksnih efekata daje procjenu srednje vrijednosti odzivne varijable u retku *Intercept*: 5.26750 i srednju vrijednost prediktora koji u ovom slučaju nisu prisutni.

Bazni model: interpretacija `lmer()` outputa – nastavak

Poziv funkcije `lmer()` s formulom:

```
lmer(rating ~ speed + height + (1 | id), data=df)
```

daje tablicu fiksnih efekata u obliku:

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.77289    0.08111   34.19
speed50      0.69602    0.09486    7.34
speed60      1.05231    0.10287   10.23
speed70      3.75039    0.11843   31.67
height300    3.27343    0.07878   41.55
height400    1.46232    0.10332   14.15
```

U stupcu *Estimate* nalaze se srednje vrijednosti prediktora po vrijednostima faktora *speed*: {50, 60 70} i faktora *height*: {300, 400}. Nivoi *speed40* i *height200* su referentni nivoi.

Model rasta (bez uvjeta):

razina 1:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2)$$

razina 2:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \qquad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \mathcal{N} \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{10}^2 \end{pmatrix}$$

Ovdje prediktor varira unutar grupa (id).

```
# lme4
lmer(y ~ t + (t | id), data=data)

# nlme
lme(y ~ t, random = ~ t | id, data=data)
```

Model rasta bez uvjeta: interpretacija `lmer()` outputa

```
lmer(rating ~ speed + (speed | id), data=df)
```

daje tablicu random efekata:

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.373932 0.61150
          speed50     0.004145 0.06438   1.00
          speed60     0.087970 0.29660  -1.00 -1.00
          speed70     0.470725 0.68609  -1.00 -1.00  1.00
 Residual             5.655618 2.37815
Number of obs: 3200, groups:  id, 200
```

$$ICC = \frac{Variance(Intercept)}{sum(Variance) + Residual} = 0.05672177$$

Korelacije slučajnih efekata

U `Random effects` tablici na prethodnom slajdu dane su i korelacije slučajnih koeficijenata za svaki nivo faktora `id` (pojedinac). Matrica korelacije ispisuje se naredbom:

```
attr(VarCorr(hm)$id, "correlation")
# što daje
            (Intercept) speed50 speed60 speed70
(Intercept)           1       1      -1      -1
speed50               1       1      -1      -1
speed60              -1      -1       1       1
speed70              -1      -1       1       1
```

```
cor(ranef(hm)$id$speed60,ranef(hm)$id$speed50) = -1
```

Visoka koreliranost sl. koef. sugerira odsutnost varijabiliteta uzrokovanog grupiranjem na višem nivou. To isto sugerira i ICC.

Model rasta (s uvjetom):

razina 1:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2)$$

razina 2:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(tretman)_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(tretman)_j + u_{1j} \end{aligned} \qquad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{10}^2 \end{pmatrix} \right)$$

```
# lme4
lmer(y ~ t * tretman + (t | id), data=data)

# nlme
lme(y ~ t * tretman, random = ~ t | id, data=data)
```

Model rasta s uvjetom ali bez *slučajnog* nagiba:

razina 1:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2)$$

razina 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(tretman)_j + u_{0j}, \quad u_{0j} \sim \mathcal{N}(0, \tau_{00}^2)$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(tretman)_j$$

```
# lme4
lmer(y ~ t * tretman + (1 | id), data=data)

# nlme
lme(y ~ t * tretman, random = ~ 1 | id, data=data)
```

Model rasta s uvjetom bez *slučajnog* pomaka[4]:

razina 1:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2)$$

razina 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(tretman)_j$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(tretman)_j + u_{1j}, \quad u_{1j} \sim \mathcal{N}(0, \tau_{10}^2)$$

```
# lme4
lmer(y ~ t * tretman + (0 + t | id), data=data)

# nlme
lme(y ~ t * tretman, random = ~ 0 + t | id, data=data)
```

[4] eng. *intercept*

Model rasta s uvjetom bez *intercept-slope* kovarijance:

razina 1:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma^2)$$

razina 2:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(tretman)_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(tretman)_j + u_{1j} \end{aligned} \qquad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \mathcal{N}\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & 0 \\ 0 & \tau_{10}^2 \end{pmatrix}$$

```
# lme4
lmer(y ~ t * tretman + (t || id) , data=data)

# isto kao gore
lmer(y ~ t * tretman + (1 | id) +  (0 + t | id), data=data)

# nlme
lme(y ~ t * tretman, random = list(subjects = pdDiag(~t)),
          data=data)
```

Organizacija viših razina hijerarhije

Tipični ulazni podaci (*data.frame*):

y	*id*	*tretman*	*terapeut*	*t*
10	1	*A*	1	0
12	1	*A*	1	1
14	1	*A*	1	2
4	2	*B*	1	0
14	2	*B*	1	1
3	2	*B*	1	2
13	3	*A*	2	0
12	3	*A*	2	1
15	3	*A*	2	2

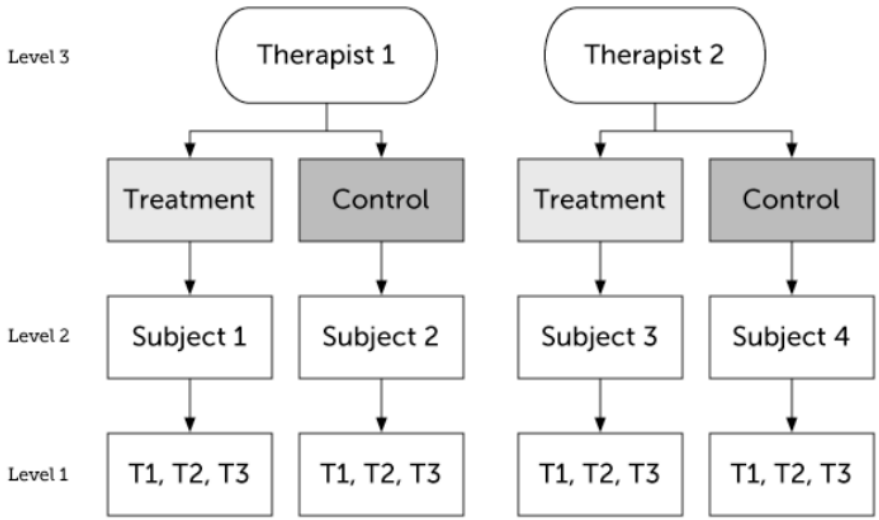
id – osoba, t – ponavljanje

$tretman : \{A - stari, B - novi\}$

Drugi nivo je osoba (id).

Što može biti treći nivo?

(1) Treći nivo može biti *tretman* kojeg izvodi *terapeut*. Svaki terapeut vrši sve tretmane (crossed effect)



Organizacija viših razina hijerarhije

Tipični ulazni podaci (*data.frame*):

y	*id*	*tretman*	*terapeut*	*t*
10	1	*A*	1	0
12	1	*A*	1	1
14	1	*A*	1	2
4	2	*B*	1	0
14	2	*B*	1	1
3	2	*B*	1	2
13	3	*A*	2	0
12	3	*A*	2	1
15	3	*A*	2	2

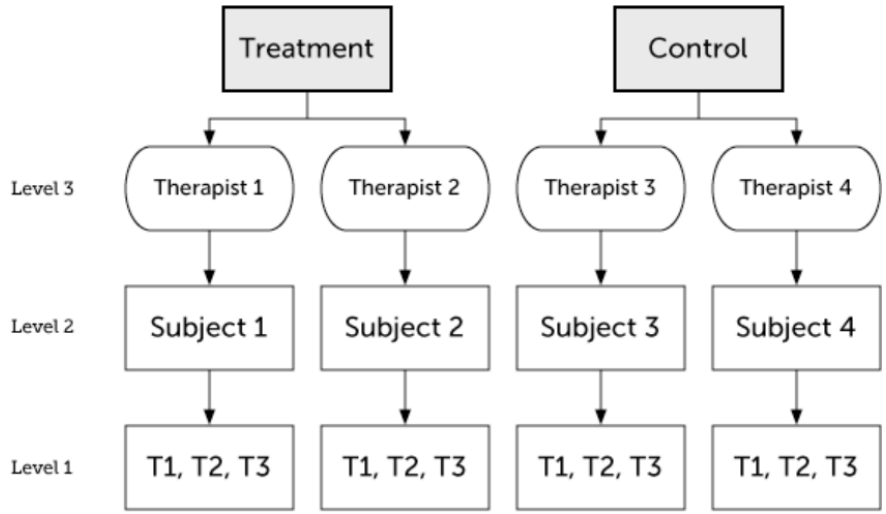
id – osoba, t – ponavljanje

$tretman : \{A - stari, B - novi\}$

Drugi nivo je osoba (id).

Što može biti treći nivo?

(2) Treći nivo može biti *terapeut* koji se razlikuju prema *tretmanu*. Svaki terapeut radi svoj tretman (ugnježđeni model)



Kvaliteta modela – ICC

Ispitat ćemo *Intra Class Correlation* (ICC) i R^2-devijancu.

– ICC se odnosi na prediktor iz višeg nivoa i mjeri 'postotak slučajnosti' koju taj prediktor unosi u model. Mala vrijednost ICC sugerira da se hijerarhijski model ne razlikuje mnogo od običnog linearnog modela (što se tiče analize 1. nivoa).

```
hm <- lme4::lmer(rating ~ speed + (speed | id), data=df)
resRnd <- attr(VarCorr(hm),'sc')^2 # random residual
varRnd <- diag(VarCorr(hm)$id) # var u Random effect:
totRnd <- sum(c(varRnd,resRnd))
(icc <- as.numeric(varRnd[1]/totRnd))
```

Kvaliteta modela – R^2

R^2 se u kontekstu LM-modela koristi u meta-analizi (upoređivanju modela). U hijerarhijskom modelu, posebno u *random slope* varijanti, svaki nivo faktora iz višeg nivoa ima svoju distribuciju reziduala, sa svojom varijancom, a time i svoj R^2. Postoji nekoliko ekstenzija R^2 u HLM kontekstu što je van interesa ove prezentacije.

Upute za daljnje čitanje:
– Raudenbush,Bryk: Hierarchical Linear Models: Applications and Data Analysis Methods, Sage Pub., 2002
– Snijders, Bosker: Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, Sage Pub., 1999
– Nakagawa and Schielzeth (2013, Methods in Ecology and Evolution).

Literatura

Kristoffer Magnusson (2015-04-21)
Kreft, I., De Leeuw, J. (1998). Introducing Multilevel Modeling. London: Sage Publications.
Raudenbush,Bryk: Hierarchical Linear Models: Applications and Data Analysis Methods, Sage Pub., 2002
Snijders, Bosker: Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, Sage Pub., 1999
Nakagawa and Schielzeth (2013, Methods in Ecology and Evolution).