Mixed-effects model

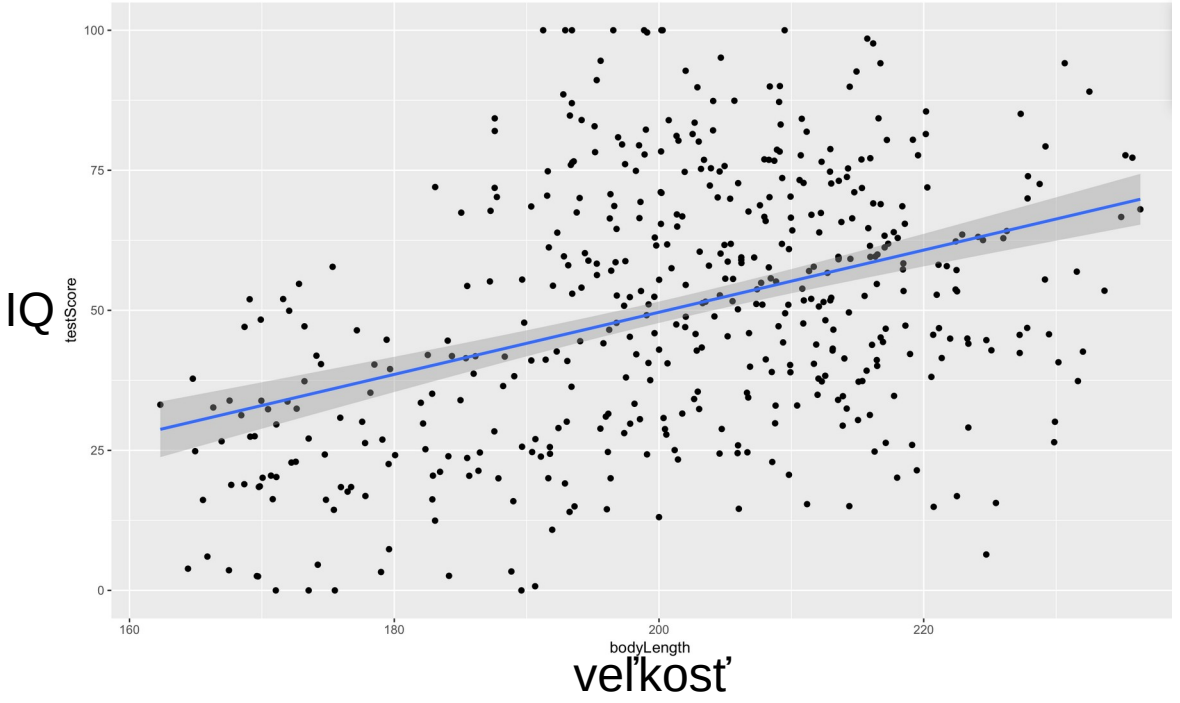
- Zdroje:
  - Tutorial Introduction to linear mixed models – biológia/životné prostredie, DRACI, dosť podrobné od základov, dobre pochopiteľné, kódy v R
    - https://ourcodingclub.github.io/tutorials/mixed-models/
  - Wikipedia - stručná
    - https://en.wikipedia.org/wiki/Mixed_model
  - Introduction to linear mixed models -lekárske
    - https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/
  - Chapter 17: Mixed Effects Modeling - pizza, mätúce, **kódy v R**
    - https://ademos.people.uic.edu/Chapter17.html

Mixed-effects model/Zmiešaný model (lineárny)

- Obsahuje
  - Fixné efekty / Fixed effects
  - Náhodné efekty /Random effects
- longitudálne štúdie (biológia,lekárske vedy) – opakované merania na jedincoch
- Napr. meranie nejakej charakteristiky s vekom jedinca (výška…) - bude charakteristické pre jedinca (náhodný efekt), ale predpokladajme že trend rastu s vekom bude podobný (fixný efekt)

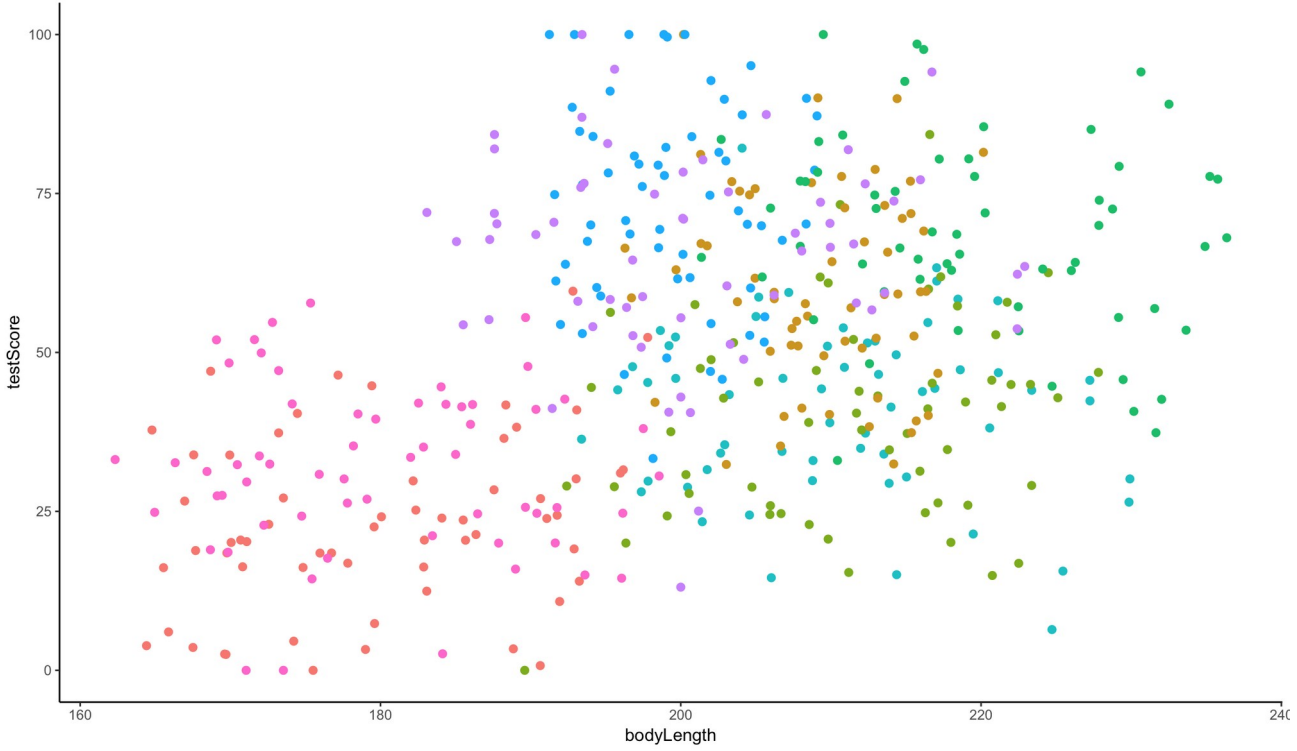
Prečo a kedy používať zmiešaný model?

Chceme si vychovať draka → merania IQ drakov v rôznych údoliach → Závisí ich IQ na veľkosti? Ak nie, menší drak toho menej zožerie…



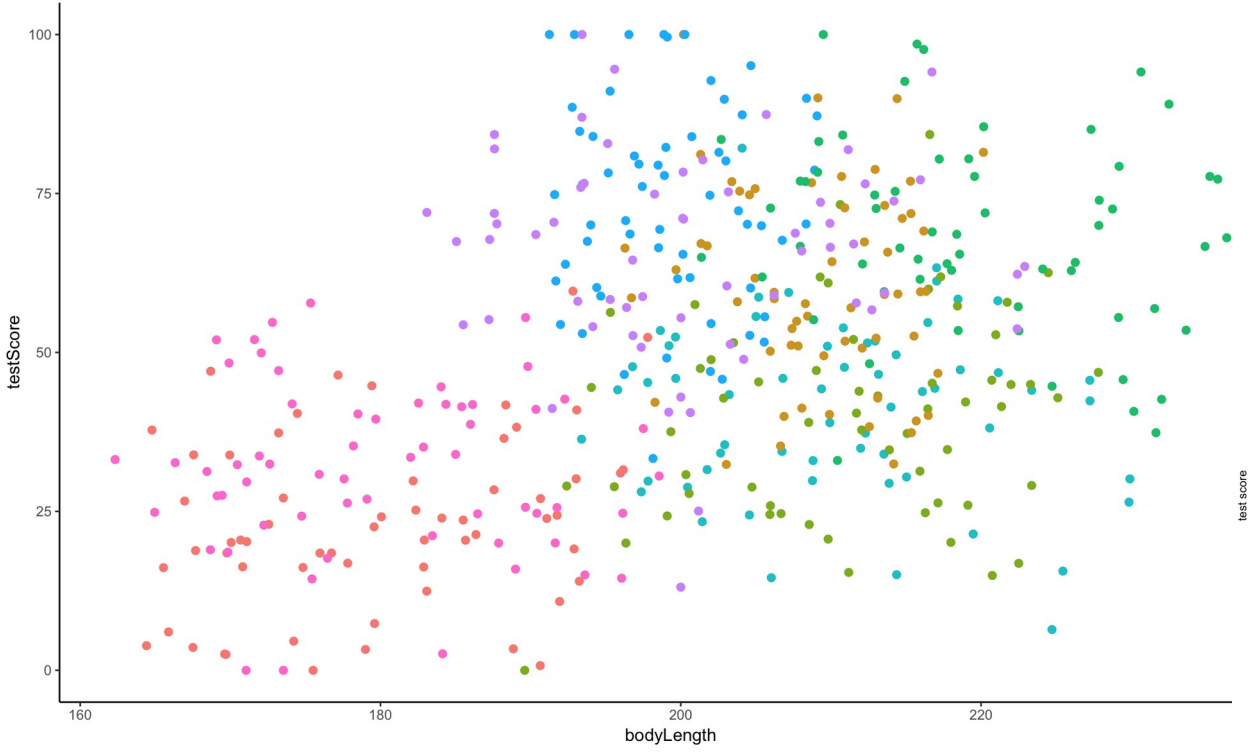
Prečo a kedy používať zmiešaný model?

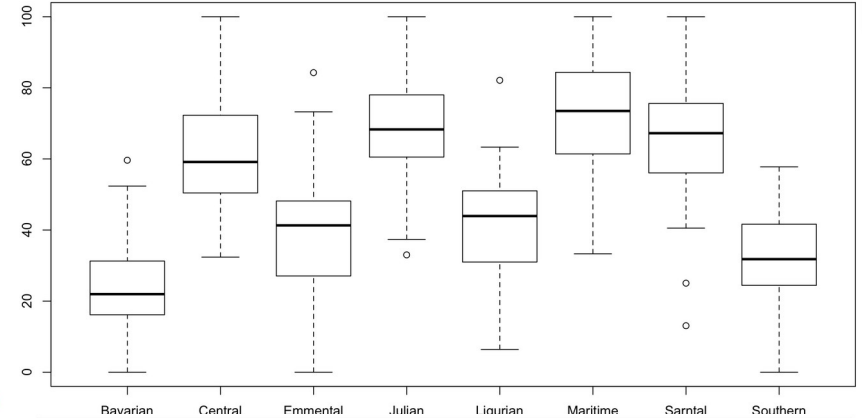
- Závislé kde sme merali?

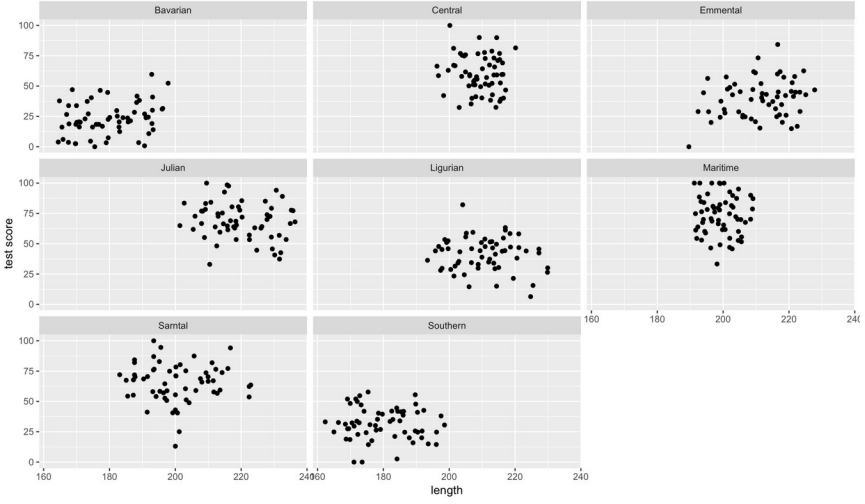


Prečo a kedy používať zmiešaný model?

- Závislé kde sme merali?



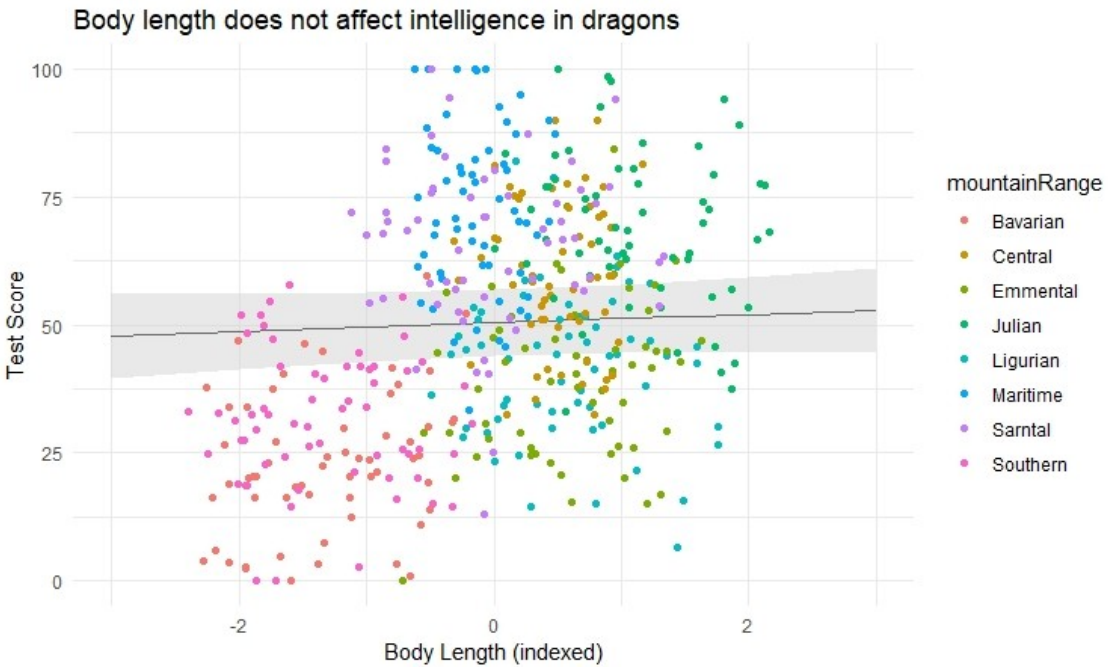




Prečo a kedy používať zmiešaný model?

Chceme si vychovať draka → merania IQ drakov v rôznych údoliach → Závisí ich IQ na veľkosti? Ak nie, menší drak toho menej zožerie…

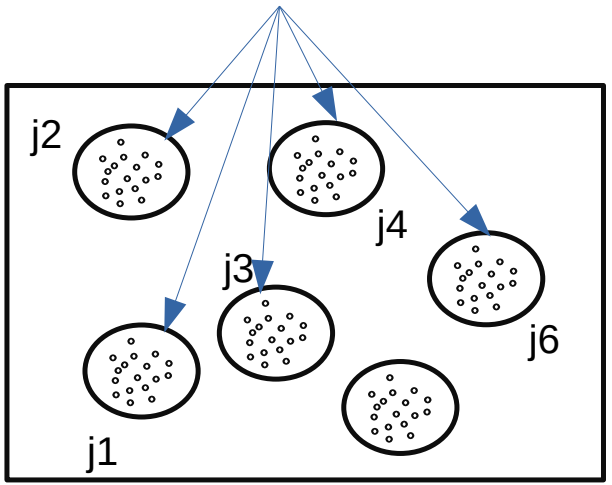
IQ



veľkosť

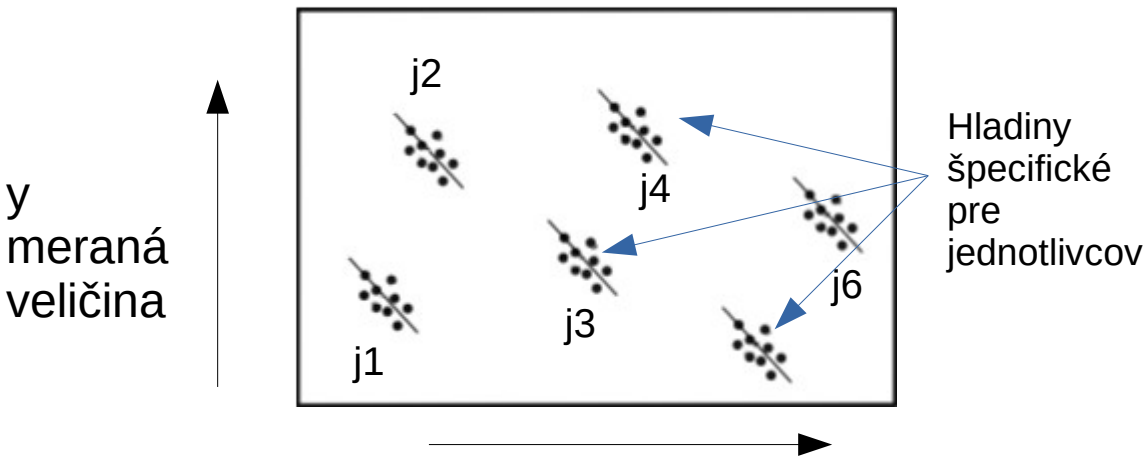
Prečo a kedy používať zmiešaný model?

- Dáta v skupinách:



uvažujme merania v rôznych časoch pre jednotlivcov

- Závislosti v dátach:



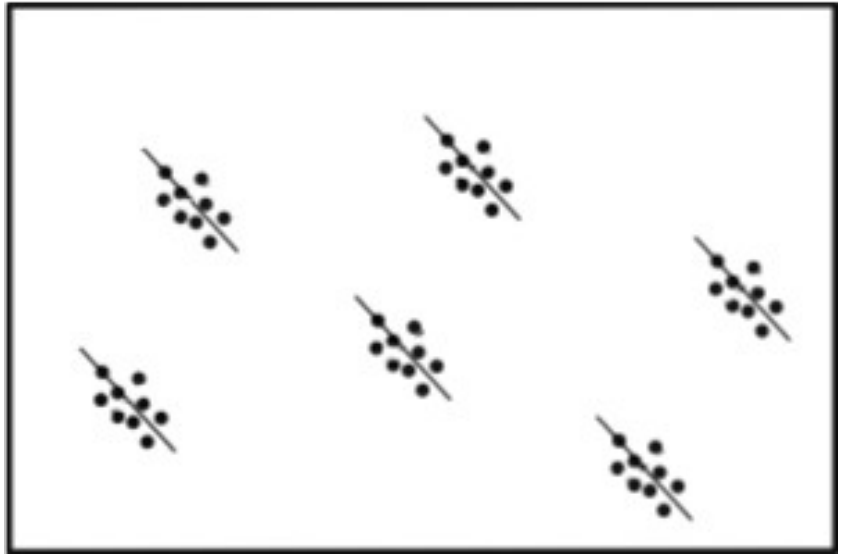
x závislá premenná
pre každého jednotlivca

Prečo a kedy používať zmiešaný model?

- Pre každého jedina máme niekoľko meraní,

  ale ak budeme prekladať lineárnu závislosť pre každého zvlášť, z veľkej sady dát sa dostaneme na oveľa menšiu

  ak budeme prekladať na celú sadu dát bez uváženia vplyvu jednotlivcov, vnášame tam chybu
- Zmiešané modely sú cesta:
  - nájdu variáciu jednotlivcov (random effect)
  - nájdu spoločný trend pre všetkých jednotlivcov (fixed effect)



Prečo a kedy používať zmiešaný model?

- Používať vždy keď dáta **nie sú nezávislé**
  - Hierarchické dáta (multilevel)
  - Ak sa dáta dajú rozdeliť do skupín
- Výsledné štandardné odchýlky opravené o závislosti v dátach
- Môžnosť skúmať vlastnosti medzi skupinami a vo vnútri skupiny

Zmiešaný model teoreticky

Klasická lineárna úloha $y = X\beta + \epsilon \longrightarrow y = X\beta + Zu + \epsilon$

- y … meraná veličina
- X …závislé veličiny pre fixné efekty (predictor/explanatory variables)
- β …regresné koeficienty pre fixné efekty
- Z ...závislé veličiny pre náhodné efekty
- u ...koeficienty pre náhodné efekty
- ϵ ...reziduá (nemodelované)

Zmiešaný model teoreticky

$$y = X\beta + Zu + \epsilon$$

- Predpoklady:
  - $E[y] = X\beta$
  - $E[u]=0$
  - $E[\epsilon]=0$
  - $cov(u,\epsilon)=0$
  - Normálne rozdelenia
- Ako sa to rieši? Hľadá sa maximálna pravdepodobnosť iteratívne (Expectation-maximization algorithm) a niekto nám to už nakódil

  pozn.: X a Z môžu obsahovať rovnaké veličiny

Zmiešaný model, čo je čo?

- Lineárna regresia klasicky: $y=a*x+b+\epsilon$
  - y_i namerané hodnoty v závislosti na x_i, hľadáme koeficienty a a b, ostane nám reziduum ϵ_i
- Lineárna regresia zmiešane: $y=a*x+b+u+\epsilon$
  - Namerané hodnoty y_i (pre dané x_i) nie sú nezávislé ale môžeme ich rozdeliť do skupín, kde každá skupina bude mať vlastný posun → $y_{jk}=a*x_{jk}+b+u_j+\epsilon_{jk}$. Hľadáme koeficienty a,b, ale aj u
- Pozn.: Náhodné efekty nás zase až tak nezaujímajú (asi tak ako nás nezaujíma každé ϵ), ale potrebujeme ich vyhodnotiť aby sme správne určili fixné efekty. Výsledok pre náhodné efekty býva len ich variancia (podobne ako výsledok pre jednotlivé ϵ býva zadaný formou štandardnej odchýlky).

Zmiešaný model, čo je čo?

- Draci z údolí:
  - 8 údolí, v každom sme namerali (IQ, veľkosť) pre nejaký počet drakov (okolo 50)
  - y: 8*50 hodnôt IQ
  - X: matica, 2 stĺpce, 8*50 riadkov $\begin{bmatrix} 1 & v_1 \\ \vdots & \vdots \\ 1 & v_{400} \end{bmatrix}$
  - Z: matica, 8 stĺpcov, 8*50 riadkov, všade 0 okrem 1 kde drak patrí do údolia $\begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ & & \ddots & \\ 0 & \dots & 0 & 1 \end{bmatrix}$

Riešenie v pythone: statsmodels

- https://www.statsmodels.org

  “statistical models, hypothesis tests, and data exploration”
- Triedy a funkcie na určovanie mnohých rôznych štatistických modelov, štatistické testy a štatistické skúmanie datasetov
- Vie toho veľa: regresia a lineárne modely – zobecnené lineárne inverzie, **zmiešané modely**, anova, time series analysis a ďalšie
- statsmodels podporuje zadávanie modelov pomocou formuly ako v R (veľa príkladov zmiešaných modelov v R)

  a pandas DataFrames (rozšírenie numpy array, veľmi elegantný kód)

Statsmodels Inštalácia

- Python 3.6, 3.7 a 3.8
- Ďalšie závislosti: Numpy, Scipy, Pandas, Patsy
- Anaconda (otestované)

```
conda install -c conda-forge statsmodels
```

- pip

```
pip install statsmodels
```

Statsmodels - základ

- Importuj moduly:

  import statsmodels.api as sm

  import pandas
- Načítaj dáta (pandas.read_csv)
- Fituj v 3 krokoch (príklad na obyčajné najmenšie štvorce)
  - 1. definuj model: `mod=sm.OLS(y, X)`
  - 2. fituj: `res = mod.fit()`
  - 3. výsledok: `print(res.summary())`

Príklad zo seizmológie - GMPE

- GMPE – jednoduchá rovnica medzi mierou pohybu pôdy *Y* a ďalšími parametrami (Mw,R,...)

$$\ln Y = F_E(\mathbf{M}, mech) + F_P(R_{JB}, \mathbf{M}, region) + F_S(V_{S30}, R_{JB}, \mathbf{M}, region, z_1) + \varepsilon_n \sigma(\mathbf{M}, R_{JB}, V_{S30})$$

  - F_E … source/event term
  - F_P … path term
  - F_S … site term
  - σ … total standard deviation
- My máme syntetickú úlohu, bez site efektov, 1 mechanizmus, syntetické *Y*

$$\ln Y = F_E(\mathbf{M}, \quad) + F_P(R_{JB}, \mathbf{M}, \quad) + \varepsilon_n \sigma(\mathbf{M}, R_{JB}, \quad)$$

- Budeme skúmať reziduá R_{ij}=ln Y_{ij} - ln Y_{ij}^{pred}

$$R_{ij} = c_k + \eta_i + \varepsilon_{ij}$$

Boore et al. 2014

Code

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
import pandas as pd

fltf="flt2.0short.csv"
df=pd.read_csv(fltf)

#kuk data, stlpce - #model, mw, ruptdist,residat,stressdrop, vrupt

#1. DEFINUJ MODEL, chceme len najst offset – fixed effect, random effect (offset variacia)
mod=smf.mixedlm("residat ~ 1", data=df, groups=df['#model'])

#2. FITUJ
modfit=mod.fit()

#3. VYSLEDKY
print(modfit.summary())
```

Možnosť zadávať pomocou R formuliek

Načítavanie dát do dataframov

Dáta sú nevyhnutné

R formula

Podľa čoho sú dáta zgrupované

Čo hľadať vo výsledku?

Mixed Linear Model Regression Results

==========

Model: MixedLM Dependent Variable: residat

No. Observations: 2200 Method: REML

No. Groups: 100 Scale: 0.1758

Min. group size: 22 Log-Likelihood: -1318.7462

Max. group size: 22 Converged: Yes

Mean group size: 22.0

Coef. Std.Err. z P>|z| [0.025 0.975]

Intercept -0.130 0.026 -4.965 0.000 -0.181 -0.078

Group Var 0.060 0.024

==========

Data info

Variancia ϵ (reziduí po odstránení náhodných efektov)

Parametre fixných efektov parameters + ich štandardná odchýlka

Variancia posunov pre jednotlivé grupy

Čo to znamená pre naše reziduá z GMPE?

Mixed Linear Model Regression Results

==========

Model: MixedLM Dependent Variable: residat

No. Observations: 2200 Method: REML

No. Groups: 100 Scale: 0.1758

Min. group size: 22 Log-Likelihood: -1318.7462

Max. group size: 22 Converged: Yes

Mean group size: 22.0

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.130	0.026	-4.965	0.000	-0.181	-0.078
Group Var	0.060	0.024				

==========

Variancia na jednotlivých staniciach (within event)

Offset celého datasetu od predikcie

Variancia posunov pre javy, charakteristika zemetrasenia/zdroja (between-event)

Zložitejší model – hľadáme závislé premenné

```
mod=smf.mixedlm("residat ~  ruptdist + vrupt",
data=df, groups=df['#model'])

modfit=mod.fit()

print(modfit.summary())
```

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	residat			
No. Observations:	2200	Method:	REML			
No. Groups:	100	Scale:	0.1120			
Min. group size:	22	Log-Likelihood:	-854.4796			
Max. group size:	22	Converged:	Yes			
Mean group size:	22.0					

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.770	0.175	-4.407	0.000	-1.113	-0.428
ruptdist	0.040	0.001	34.241	0.000	0.038	0.043
vrupt	0.138	0.079	1.742	0.082	-0.017	0.293
Group Var	0.067	0.032				

???Ako vybrať správny model? Ktoré parametre sú dôležité? Pozor na overfitting!

Pridanie náhodný efektu v smernici

```
modr=smf.mixedlm("residat ~  vrupt", data=df,
groups=df['#model'],re_formula='~ruptdist')
modfitr=modr.fit()
print(modfitr.summary())
```

```
modr2=smf.mixedlm("residat ~
vrupt+ruptdist", data=df,
groups=df['#model'],re_formula='~ruptdist')
modfitr2=modr2.fit()
print(modfitr2.summary())
```

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	residat
No. Observations:	2200	Method:	REML
No. Groups:	100	Scale:	0.1002
Min. group size:	22	Log-Likelihood:	-862.7192
Max. group size:	22	Converged:	Yes
Mean group size:	22.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.253	0.051	-4.987	0.000	-0.352	-0.153
Group Var	0.120	0.093				
Group x ruptdist Cov	-0.011	0.009				
ruptdist Var	0.002	0.001				

Kde nájdeme ďalšie výsledky?

- modfit.fe_params … koeficienty pre fixed effects
- modfit.random_effects … posuny pre random efekty
- modfit.fittedvalues … “predikcie fixed+ random efekt”
- modfit.resid … rezidua (nenamodelovane, data=resid+fittedvalues)

Otázky?

Je to rýchle? Áno, to teda je.

Pre mňa bolo najväčším orieškom definovať si správne úlohu, čo sa snažím nájsť? Čo má byť náhodným efektom a čo nie?

- náhodný efekt má mať nulovú strednú hodnotu, je to len variácia nie 1 dátového bodu ale skupiny dát
- chceme ten efekt presne modelovať? V zmysle chceme presne vedieť jeho hodnotu?
- náhodný efekt je v diskrétnych vstupoch

Ďakujem za pozornosť.