

**Bi7491 Regresní modelování**

---

# **Smíšené modely**

# Co již znáte z minulých hodin?

- ➔ Užití lineárního regresního modelu – spojité výsledky  
(definice, předpoklady – analýza reziduí, praktické užití)
- ➔ Binární výsledky (např. onemocnění) – práce s logistickou regresí, specifika, interpretace výsledků – poměr šancí
- ➔ Analýza deviance, Poissonova regrese, nadměrný rozptyl
- ➔ Kauzální vztahy – zkreslující faktory, interakce, kauzální diagramy
- ➔ Příprava dat: kategoriální proměnné, spojité proměnné – transformace, centrování, škálování

# Co byste po dnešní hodině měli vědět a umět?

- popsat problém s klasickými statistickými metodami v případě úloh s opakovanými pozorováními u stejných subjektů (skupin)
- znát rozdíl mezi pevnými a náhodnými efekty
- znát definici smíšeného a longitudinálního modelu
- provést základní hodnocení dat se shluky prostřednictvím popsaných metod

# Motivace

- Předchozí příklady: **průřezový (cross-sectional) design**  
Měření jsou provedena v jediném časovém okamžiku  
Např. srovnání podílu tělesného tuku u 10letých a 15letých dívek – dvě kohorty, nepárový t-test, možná zkreslení
- **Longitudinální design**  
Stejně dívky, měřeny v 10 a 15 letech – párový t-test  
Nyní již všechna měření **nejsou nezávislá (předpoklad standardních statistických technik)**  
Dívky tvoří „shluky“ v datech – pozorování uvnitř shluku budou zřejmě podobnější, než v různých shlucích

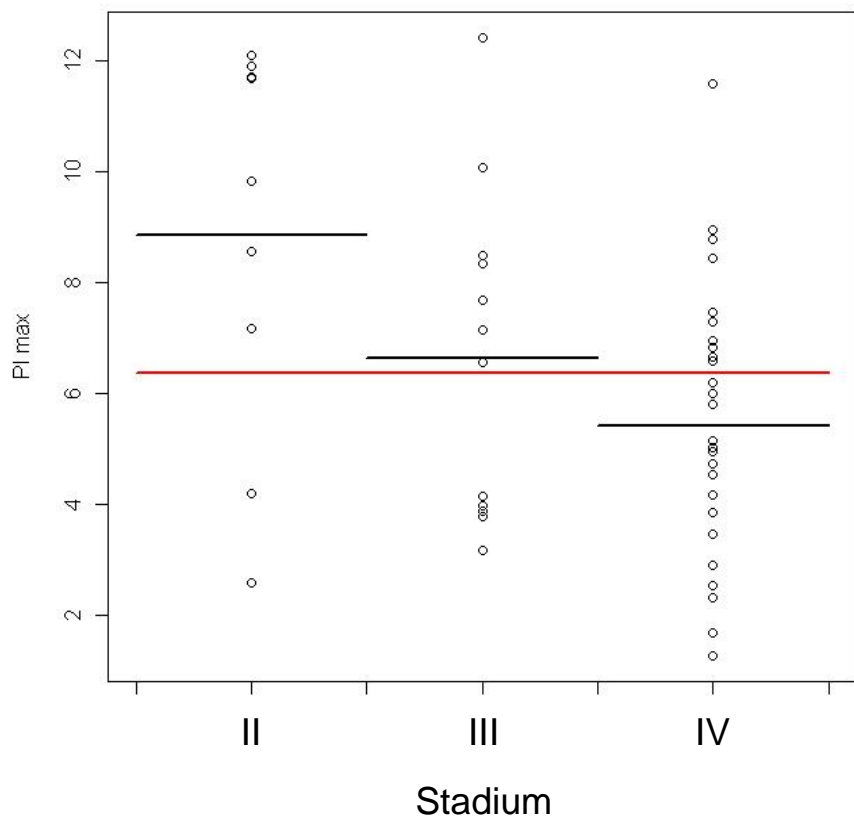
## **Smíšené modely**

---

**Opakování – analýza rozptylu  
(viz Biostatistika pro MB)**

# Příklad – CHOPN

## Maximální inspirační tlak dle stadií



$$n_1 = 9$$

$$\bar{y}_1 = 8,9 \text{ kPa}$$

$$s_1 = 3,5 \text{ kPa}$$

$$n_2 = 12$$

$$\bar{y}_2 = 6,6 \text{ kPa}$$

$$s_2 = 2,9 \text{ kPa}$$

$$n_3 = 27$$

$$\bar{y}_3 = 5,4 \text{ kPa}$$

$$s_3 = 2,5 \text{ kPa}$$

Celkový průměr („grand mean“)

$$n = 48$$

$$\bar{y} = 6,4 \text{ kPa}$$

$$s = 3,0 \text{ kPa}$$

# Značení

*k* skupin, v *i*-té skupině  $n_i$  pozorování

→ Součty:  $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$   $Y_{\cdot\cdot} = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$

→ Průměry:  $\bar{y}_{i\cdot} = Y_{i\cdot} / n_i$   $\bar{y}_{\cdot\cdot} = Y_{\cdot\cdot} / n$

Skupinový průměr  
(„population mean“)

Celkový průměr  
(„grand mean“)

→ Celková variabilita v souboru:

$$S_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{\cdot\cdot})^2$$

Stupně volnosti:  $df_T = n - 1$

→ Variabilita v rámci skupin (reziduální součet čtverců):

$$S_e = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i\cdot})^2$$

Stupně volnosti:  $df_e = n - k$

→ Variabilita mezi skupinami (příslušná sledovanému vlivu = proměnné):

$$S_A = \sum_{i=1}^k n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

Stupně volnosti:  $df_A = k - 1$

# Vztahy mezi odhady variability

→ Platí:

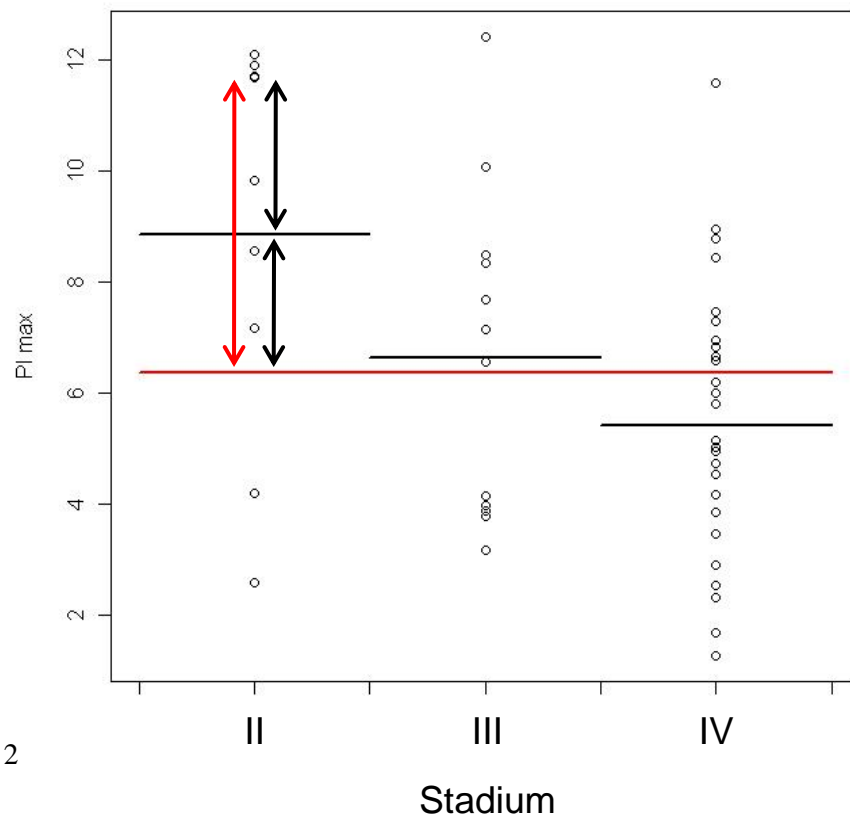
$$Y_{ij} - \bar{y}_{..} = (Y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$$

→ Dále se dá ukázat, že platí:

$$S_T = S_e + S_A$$

→ Tedy platí, že celková variabilita se dá rozložit na variabilitu v rámci skupin a variabilitu mezi skupinami:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^k n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$





# Umělý příklad

Léčba	Pozorovaná hodnota	Skupinový průměr	Skupinový průměr – celkový průměr	Pozorovaná hodnota – skupinový průměr	Pozorovaná hodnota – celkový průměr
1	10	12	-4	-2	-6
1	12	12	-4	0	-4
1	14	12	-4	2	-2
2	19	20	4	-1	3
2	20	20	4	0	4
2	21	20	4	1	5
3	14	16	0	-2	-2
3	16	16	0	0	0
3	18	16	0	2	2
	<b>Celkový průměr = 16</b>		<b>Součet čtverců = 96</b>	<b>Součet čtverců = 18</b>	<b>Součet čtverců = 114</b>
			<b>Stupně volnosti = 2</b>	<b>Stupně volnosti = 6</b>	<b>Stupně volnosti = 8</b>

# Princip analýzy rozptylu

→ Testová statistika analýzy rozptylu:

vysvětlená variabilita

$$F = \frac{\frac{\sum_{i=1}^k n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{k-1}}{\frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i\cdot})^2}{n-k}} = \frac{S_A / df_A}{S_e / df_e}$$

reziduální variabilita

Za platnosti  $H_0$  platí:

$$F \sim F(k-1, n-k)$$

(pokud faktor A nic nevysvětluje – výběrové průměry ve skupinách jsou blízké – testové kritérium blízké 0 – nezamítáme nulovou hypotézu)

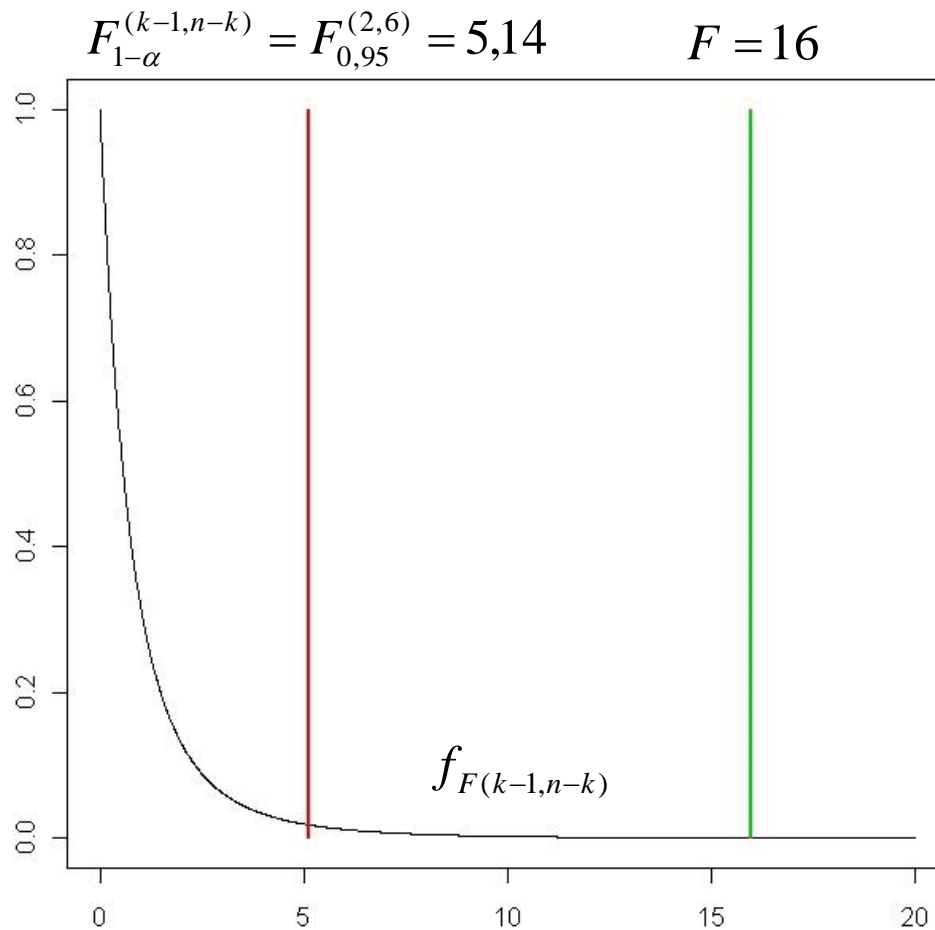
# Výsledek analýzy rozptylu

➔ Výsledné počty se standardně zaznamenávají do tzv. tabulky analýzy rozptylu:

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Mezi skupinami	$S_A = 96$	$df_A = k - 1 = 2$	$MS_A = 48$	<b>F = 16</b>	<b>0,004</b>
Uvnitř skupin	$S_e = 18$	$df_e = n - k = 6$	$MS_e = 3$		
Celkem	$S_T = 114$	$df_T = n - 1 = 8$			

➔ Nulovou hypotézu zamítneme/nezamítneme buď na základě srovnání výsledné p-hodnoty se zvolenou hladinou významnosti testu  $\alpha$ , nebo srovnáním výsledné F statistiky s kritickou hodnotou (kvantilem) rozdělení  $F(k - 1, n - k)$  příslušnou zvolené hladině významnosti testu  $\alpha$ .

# Výsledek umělého příkladu



Na hladině významnosti  $\alpha = 0,05$  zamítáme  $H_0$  o rovnosti středních hodnot.

# Kontrolní dotaz:

Jak vypadá matice plánu pro uvedený model?

# Kontrolní dotaz:

Jak vypadá matice plánu pro uvedený model?

- jedná se o lineární model s jedním kategoriálním prediktorem s  $k$  hodnotami

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ \hline 1 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 0 \\ \hline \vdots & \vdots & \dots & \vdots \\ \hline 1 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 1 \end{pmatrix}$$

$EY_i = \mu$

$EY_i = \mu + \alpha_1$

$EY_i = \mu + \alpha_{k-1}$

$H_0: \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{k-1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

$H_1: \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{k-1} \end{pmatrix} \neq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

# **Smíšené modely**

---

## **Pevné a náhodné efekty**

# Pevné a náhodné efekty

- **pevný efekt**  
neznámá konstanta, kterou se snažíme odhadnout konkrétní parametr – těmi jsme se zabývali doted'
- **náhodný efekt**
  - představuje náhodnou veličinu
  - jeho přidání do lineárního prediktoru umožňuje zavést korelaci mezi pozorováními – např. v podobě „náhodného interceptu“, který reprezentuje nepozorovatelnou individuální charakteristiku
  - neodhadujeme náhodný efekt (což ani nelze), ale parametry popisující jeho rozdělení – nezajímají nás efekty konkrétních jedinců, ale **informace o cílové populaci jako celku**



# Smíšený model

- obsahuje pevné i náhodné efekty
- jednoduchý příklad: analýza rozptylu (two-way)

$$Y_{ijk} = \mu + \tau_i + \nu_j + \varepsilon_{ijk}$$

**pevné efekty**      **náhodné efekty**      **reziduum**

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$
$$\nu_j \sim N(0, \sigma_\nu^2)$$

# Smíšený model

$$Y_{ijk} = \mu + \tau_i + \nu_j + \varepsilon_{ijk}$$

$H_0 : \tau_i = 0, \forall i$   
**pevné efekty**      **několik parametrů**

**náhodné efekty**  
 $\nu_j \sim N(0, \sigma_\nu^2)$   
 $H_0 : \sigma_\nu^2 = 0$   
**jediný parametr**

# Smíšené modely

---

## Odhad parametrů

# Nejjednodušší model

- analýza rozptylu (one-way), vyvážený design

**náhodné efekty**

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

$i = 1, \dots, a$  skupiny

$j = 1, \dots, n$  pozorování  
ve skupině

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

**vzájemně  
nezávislé**

# Náhodné efekty vytvářejí korelaci

$$E(Y_{ij}) = \mu \quad D(Y_{ij}) = \sigma_{\alpha}^2 + \sigma_{\varepsilon}^2$$

$$\text{cov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma_{\alpha}^2 + \sigma_{\varepsilon}^2, & i = i', j = j' & \text{rozptyl} \\ \sigma_{\alpha}^2, & i = i', j \neq j' & \text{v jedné skupině} \\ 0, & i \neq i', j \neq j' & \text{v různých skupinách} \end{cases}$$

- koeficient vnitrotřídní korelace  
(*intraclass correlation coefficient*, ICC)

$$\rho = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2}$$

# Součet čtverců – odhad ANOVA

vyvážený design – n je počet ve skupině

$$\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^a \sum_{j=1}^n (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

celkový součet čtverců  
(SST)

reziduální součet  
čtverců (SSE)

součet čtverců  
efekt alfa (SSA)

stupně volnosti: an-1

an-a

a-1

průměrné čtverce MST

MSE

MSA

$$E(SSE) = a(n-1)\sigma_{\varepsilon}^2$$

$$E(SSA) = (a-1)(n\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

$$\hat{\sigma}_{\varepsilon}^2 = SSE / (a(n-1)) = MSE$$

$$\hat{\sigma}_{\alpha}^2 = \frac{SSA / (a-1) - \hat{\sigma}_{\varepsilon}^2}{n} = \frac{MSA - MSE}{n}$$

# Odhad metodou maximální věrohodnosti

- v praxi není ANOVA estimátor příliš vhodný
- pevné efekty s normálními chybami

$$Y = X\beta + \varepsilon$$

$X$  je matice  $n \times p$

**matice plánu pro pevné efekty**

- doplnění náhodných efektů – smíšený model

$$Y = X\beta + Z\gamma + \varepsilon$$

$Z$  je matice  $n \times q$

**matice plánu pro náhodné efekty**

# Odhad metodou maximální věrohodnosti

$$\gamma \sim N(0, \sigma^2 D)$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

$$Y = X\beta + Z\gamma + \varepsilon$$

$$Y \sim N(X\beta, \sigma^2 (I + ZDZ^T))$$

Máme věrohodnostní funkci  odhad parametrů  $\beta$ ,  $\sigma^2$ ,  $D$

- v praxi se využívá tzv. **REML** (*restricted maximum likelihood*)  
obecně dostáváme méně zkreslené odhady parametrů



# **Smíšené modely**

---

## **Užití smíšených modelů**

# Bloky jako náhodné efekty

- blok – experimentální jednotka
  - definovány podmínkami experimentu nebo úsudkem
  - např. konkrétní jedinec, laboratoř vyhodnocující vzorky, zdravotnické zařízení, vesnice, rodina, vrh, ...

# Příklad:

## Produkce penicilinu

- snažíme se porovnat čtyři výrobní procesy A, B, C, D
  - definovány a stanoveny, zajímají nás – PEVNÉ EFEKTY
- médium (kukuřičný výluh) lze vždy vytvořit v množství pro čtyři experimenty
  - náhodně utvořeny, nejsou předmětem výzkumu – NÁHODNÉ EFEKTY
- výsledkem je množství získaného penicilinu

# Příklad:

## Produkce penicilinu

- ANOVA s pevnými efekty:

```
> lmod <- aov(yield ~ blend + treat, penicillin)
```

```
> summary(lmod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
blend	4	264	66.00	3.504	0.0407	*
treat	3	70	23.33	1.239	0.3387	
Residuals	12	226	18.83			

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coef(lmod)
```

(Intercept)	blend1	blend2	blend3	blend4	treat1
86	6	-3	-1	2	-2
treat2	treat3				
-1	3				

- POZOR: Jsou využity odlišné typy kontrastů, které srovnávají výsledek s průměrem (interceptem) – součet všech je 0

# Kategoriální prediktory

## Součtové kontrasty

- Stanovena dodatečná podmínka:  $\sum \beta_i = 0$

Původní	Nové proměnné		
treat	treat1	treat2	treat3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

$$EY_i = \mu + \beta_1$$

$$EY_i = \mu + \beta_2$$

$$EY_i = \mu + \beta_3$$

$$EY_i = \mu - (\beta_1 + \beta_2 + \beta_3)$$

# Příklad:

## Produkce penicilinu

- Smíšený model:

```
> mmod <- lmer(yield ~ treat + (1|blend), penicillin)
```

**pevný efekt**



**náhodný efekt**



data seskupena podle blend

1 ... jen intercept

# Příklad:

## Produkce penicilinu

> **summary(mmod)**

```
Linear mixed model fit by REML
Formula: yield ~ treat + (1 | blend)
Data: penicillin
AIC      BIC logLik deviance REMLdev
118.6 124.6 -53.3  117.3  106.6
```

Random effects:

Groups	Name	Variance	Std.Dev.
blend	(Intercept)	11.792	3.4339
Residual		18.833	4.3397

Number of obs: 20, groups: blend, 5

**náhodný efekt**

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	86.000	1.817	47.34
treat1	-2.000	1.681	-1.19
treat2	-1.000	1.681	-0.59
treat3	3.000	1.681	1.78

**pevný efekt**

**pro orientační test použijeme normální  
aproximaci t-statistiky**

Correlation of Fixed Effects:

	(Intr)	treat1	treat2
treat1	0.000		
treat2	0.000	-0.333	
treat3	0.000	-0.333	-0.333

**obdobně i test vnořených modelů pomocí  
ANOVA (nutno použít ML místo REML)**

# Příklad:

## Produkce penicilinu

- **Odhad náhodných efektů??**  
Už to nejsou parametry, ale náhodné veličiny (s nulovou střední hodnotou)
- Lze však spočítat tzv. posteriorní střední hodnotu

```
> ranef(mmod) $blend
```

```
(Intercept)
```

```
Blend1  4.2878788  
Blend2 -2.1439394  
Blend3 -0.7146465  
Blend4  1.4292929  
Blend5 -2.8585859
```

Z modelu s pevnými efekty:

```
Blend1  6  
Blend2 -3  
Blend3 -1  
Blend4  2  
Blend5 -4
```

**Odhady se ve srovnání s  
původním modelem „scvrkly“:  
SHRINKAGE ESTIMATOR**

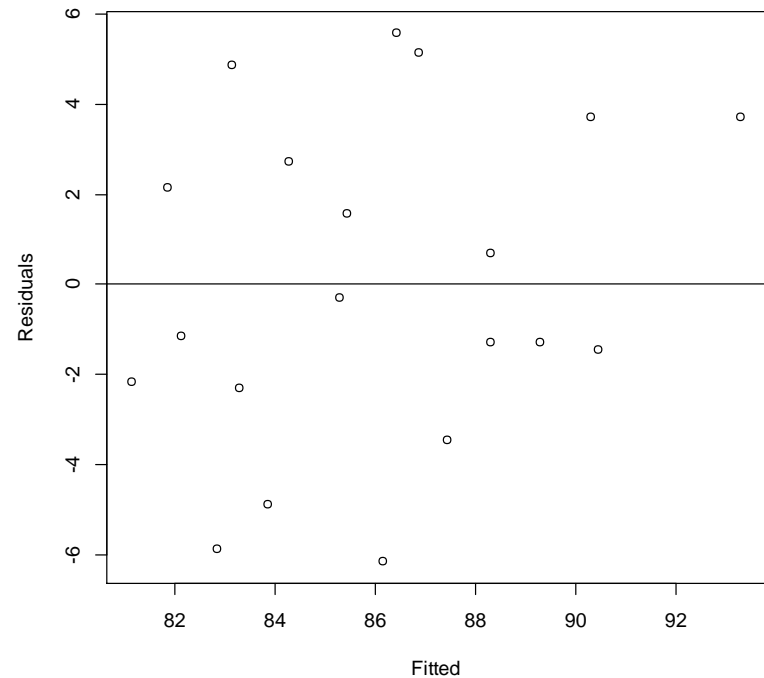
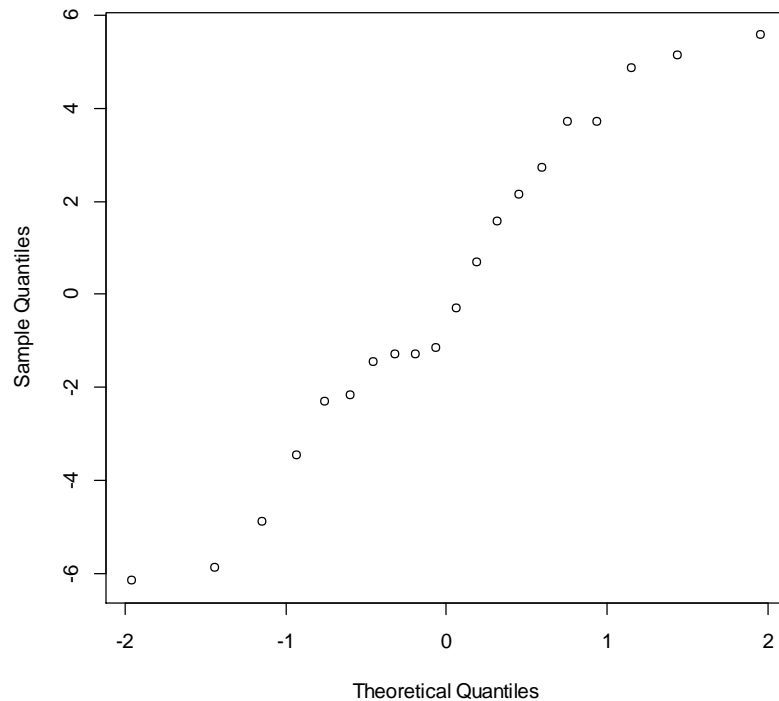
- Odhad hodnoty pro některý výluh  
Kombinace pevných efektů a posteriorní střední hodnoty:  
**BEST LINEAR UNBIASED PREDICTOR (BLUP)**



# Příklad:

## Produkce penicilinu

- **Analýza reziduí**  
Srovnání s predikovanými hodnotami



# Vnořené efekty

- Příklad: Laboratorní testy  
Měření obsahu tuku ve vaječném prášku
- 6 laboratoří
  - v každé dva laboranti
    - každý dva vzorky
      - u každého dvě měření

$$Y_{ijkl} = \mu + L_i + T_{ij} + S_{ijk} + \varepsilon_{ijkl}$$

```
cmod <- lmer(Fat ~ 1
+ (1|Lab)
+ (1|Lab:Technician)
+ (1|Lab:Technician:Sample) ,
data=eggs)
```

# **Smíšené modely**

---

## **Longitudinální data**

# Longitudinální data

- proměnná je u jedince měřena opakovaně
- longitudinální studie se zabývají změnou příslušného výsledku v čase
- **cílem je charakterizovat změnu a faktory které ji ovlivňují**
- měření u jedince jsou korelovaná!!!

# Pro každého jedince...

- u každého jedince ( $i$ ) je provedeno  $n_i$  měření ve vektoru  $Y_i$

## pevné efekty

vektor společný pro celou populaci

možná explicitní  
autokorelace

$$Y_i | \gamma_i \sim N(X_i \beta + Z_i \gamma_i, \sigma^2 \Lambda_i)$$

## náhodné efekty

každý jedinec si „vylosuje“ vektor  
rozdělení společné pro celou populaci

$$\gamma_i \sim N(0, \sigma^2 D)$$

$$Y_i \sim N(X_i \beta, \Sigma_i) \quad \Sigma_i = \sigma^2 (\Lambda_i + Z_i D Z_i^T)$$

# Příklad:

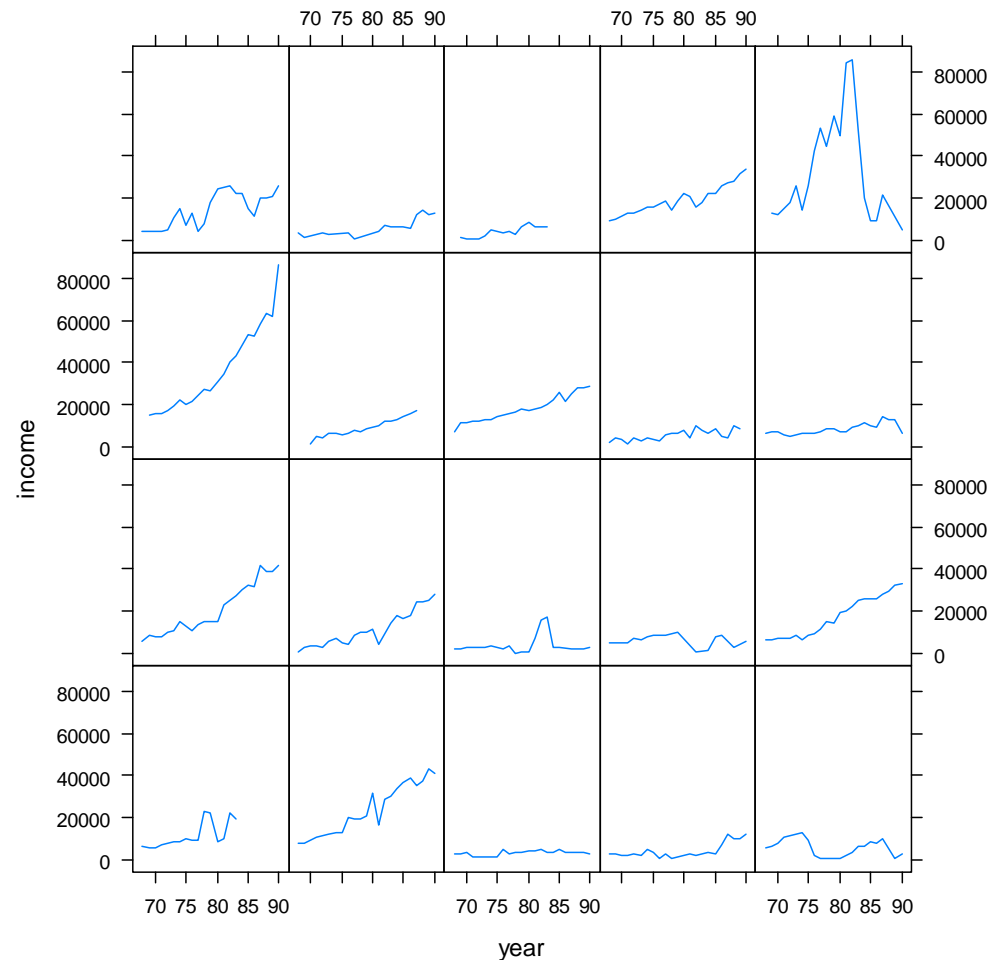
## Panelová studie příjmové dynamiky

- americká studie, vývoj příjmů
- 85 osob, alespoň 11 záznamů mezi 1968-1990

# Příklad:

## Panelová studie příjmové dynamiky

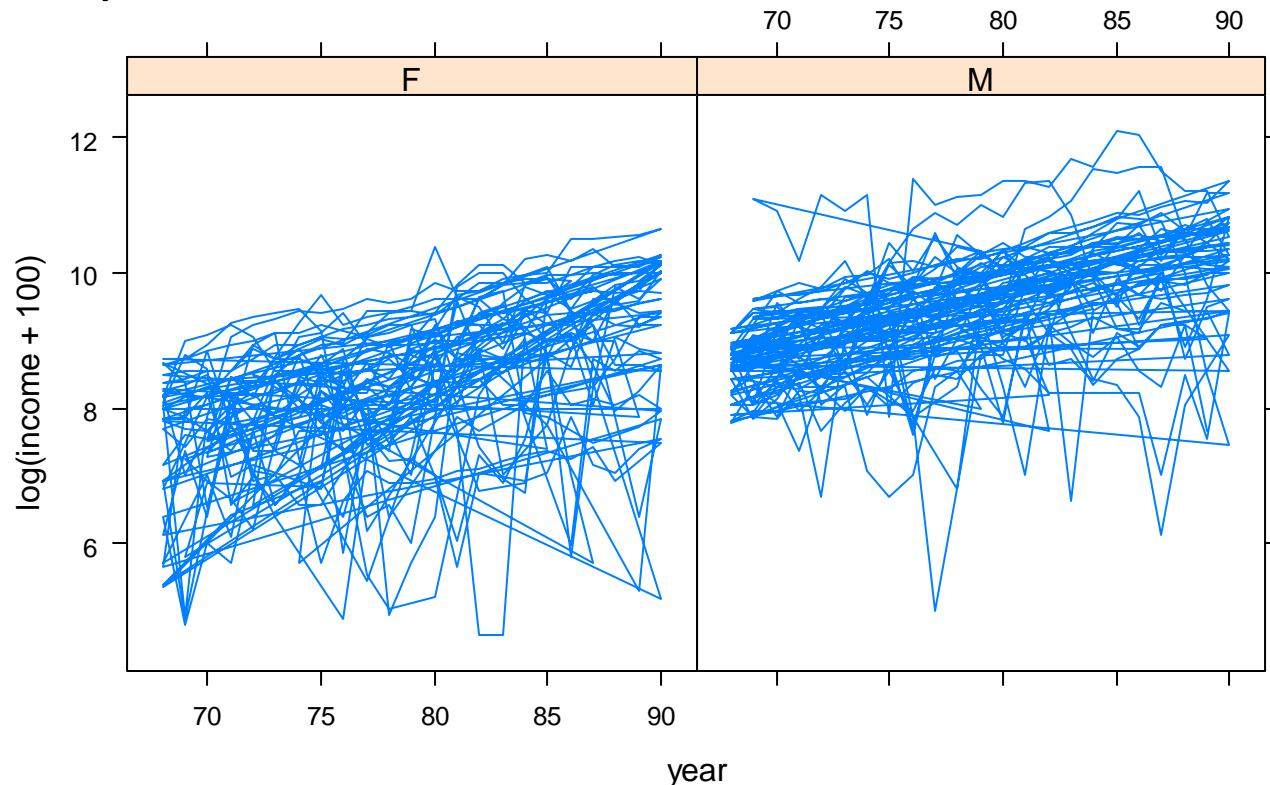
- vývoj u 20 osob:



# Příklad:

## Panelová studie příjmové dynamiky

- srovnání ženy x muži



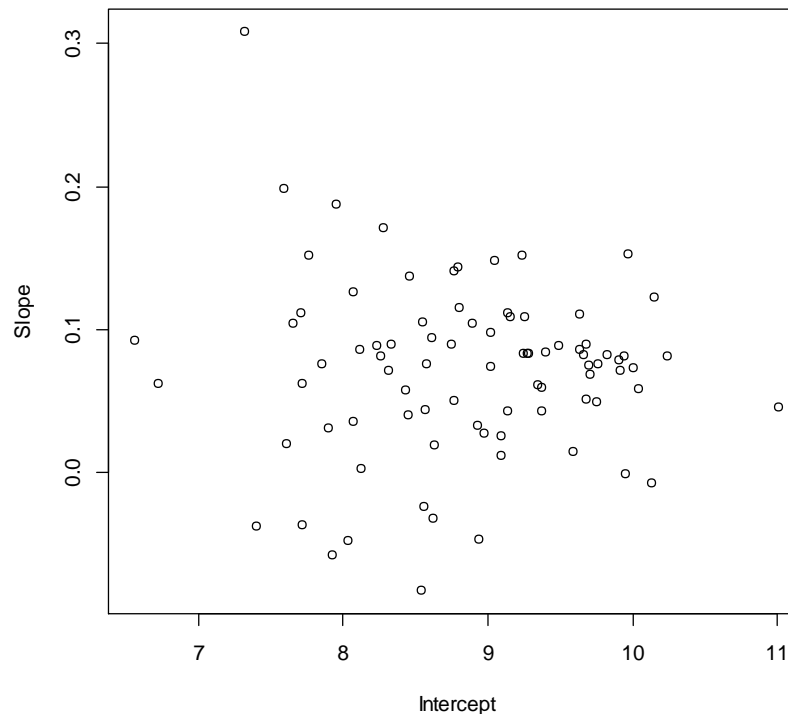
- zdá se, že muži mají vyšší plat, u žen však rychleji roste



# Příklad:

## Panelová studie příjmové dynamiky

- pro každého jedince lze sestavit vlastní model
  - vlastní absolutní člen (interpretovatelnost – v roce 1978) i sklon přímky



# Příklad:

## Panelová studie příjmové dynamiky

- Smíšený model:

```
> mmod <- lmer(log(income) ~ cyear*sex+age+educ+(cyear|person),psid)
```

**pevný efekt**

**náhodný efekt**

data seskupena podle osob

Jaké parametry model bude obsahovat?

# Příklad:

## Panelová studie příjmové dynamiky

- Smíšený model:

```
> mmod <- lmer(log(income) ~ cyear*sex+age+educ+(cyear|person),psid)
```

pevný efekt

náhodný efekt

data seskupena podle osob

Jaké parametry model bude obsahovat?

$$\log(\text{income})_{ij} = \mu + \beta_y \text{year}_i + \beta_s \text{sex}_j + \beta_{ys} \text{sex}_j \times \text{year}_i + \beta_e \text{educ}_j + \beta_a \text{age}_j \\ + \gamma_j^0 + \gamma_j^1 \text{year}_i + \varepsilon_{ij}$$

... i-tý rok u j-té osoby

# Příklad:

## Panelová studie příjmové dynamiky

```
Linear mixed model fit by REML
Formula: log(income) ~ cyear * sex + age + educ +
(cyear | person)
Data: psid
AIC BIC logLik deviance REMLdev
3840 3894 -1910 3786 3820
```

---

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
person	(Intercept)	0.28166	0.53071	
	cyear	0.00240	0.04899	0.187
	Residual	0.46727	0.68357	

Number of obs: 1661, groups: person, 85

---

**náhodný efekt**

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	6.674178	0.543334	12.284
cyear	0.085312	0.008999	9.480
sexM	1.150315	0.121293	9.484
age	0.010932	0.013524	0.808
educ	0.104212	0.021437	4.861
cyear:sexM	-0.026307	0.012238	-2.150

**pevný efekt**

**pro velký vzorek lze použít  
normální aproximaci t-statistiky**

**(cyear – centrovaný)**

# Smíšené modely

---

**Závěr**

# Co byste po dnešní hodině měli vědět a umět?

- popsat problém s klasickými statistickými metodami v případě úloh s opakovanými pozorováními u stejných subjektů (skupin)
- znát rozdíl mezi pevnými a náhodnými efekty
- znát definici smíšeného a longitudinálního modelu
- provést základní hodnocení dat se shluky prostřednictvím popsaných metod

# Praktický úkol

- MIT Growth and Development Study
  - Průvodní text a data k dispozici na adrese:  
<https://content.sph.harvard.edu/fitzmaur/ala2e/fat.txt>
1. Prostudujte si text ke studii
  2. Načtěte příslušná data do software R
  3. Proveďte základní popis longitudinálních dat
  4. Sestavte smíšený model pro vývoj podílu tělesného tuku
  5. Odpovězte na otázku, zda růst podílu tělesného tuku je stejný před a po menarche

# Použitá literatura

- ➔ Julian J. Faraway: Extending the Linear Model with R
- ➔ Garrett M. Fitzmaurice a kol.: Applied Longitudinal Analysis

➔ viz také

[http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug\\_mixed\\_sect022.htm](http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect022.htm)

➔ <https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html>