

PSYb2520

Statistická analýza dat v psychologii II

Víceúrovňový lineární model

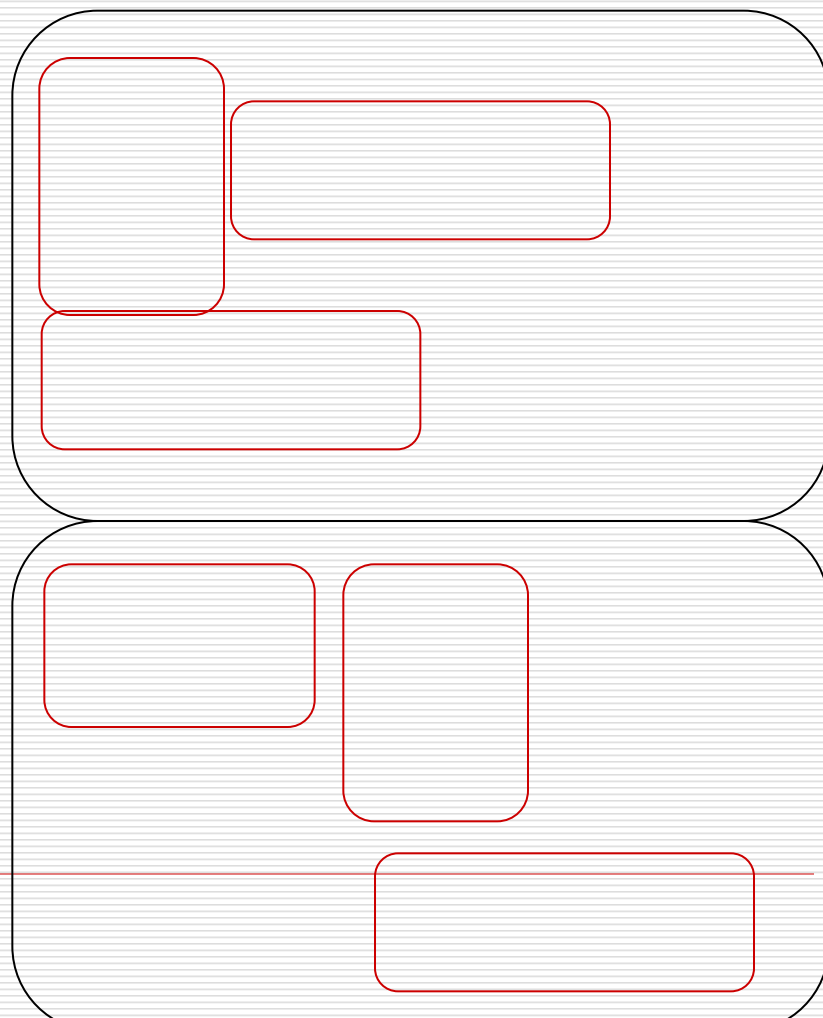
(multilevel, hierarchical, mixed,
random-coefficients model)

Víceúrovňová data

ID	Třída	Výkon
100	1	11
...	1	20
120	1	31
121	2	40
..	2	52
150	2	63
151	3	20
...	3	40
180	3	30
181	4	100

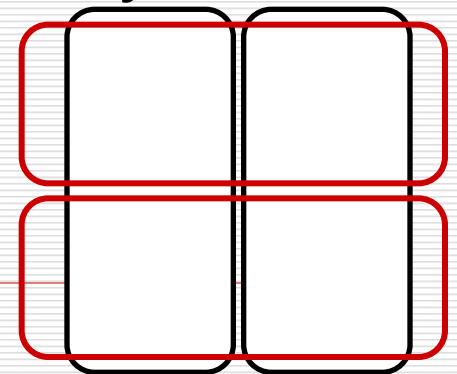
Víceúrovňová data

ID	Třída	Výkon
100	1	11
...	1	20
120	1	31
121	2	40
..	2	52
150	2	63
151	3	20
...	3	40
180	3	30
181	4	100



Víceúrovňová data – **vnořené (nested) faktory**

- Určité úrovně faktorů nižší úrovně se vyskytují pouze v jediné úrovni faktorů vyšší úrovně
 - Proto též hierarchická data - Multilevel linear model
 - Konkrétní třída je jen v jedné škole, žák je členem jen jedné třídy
- Protikladem pro vnořené faktory jsou **zkřížené (crossed) faktory** – vyskytují se všechny kombinace jejich hodnot
 - Mixed linear model

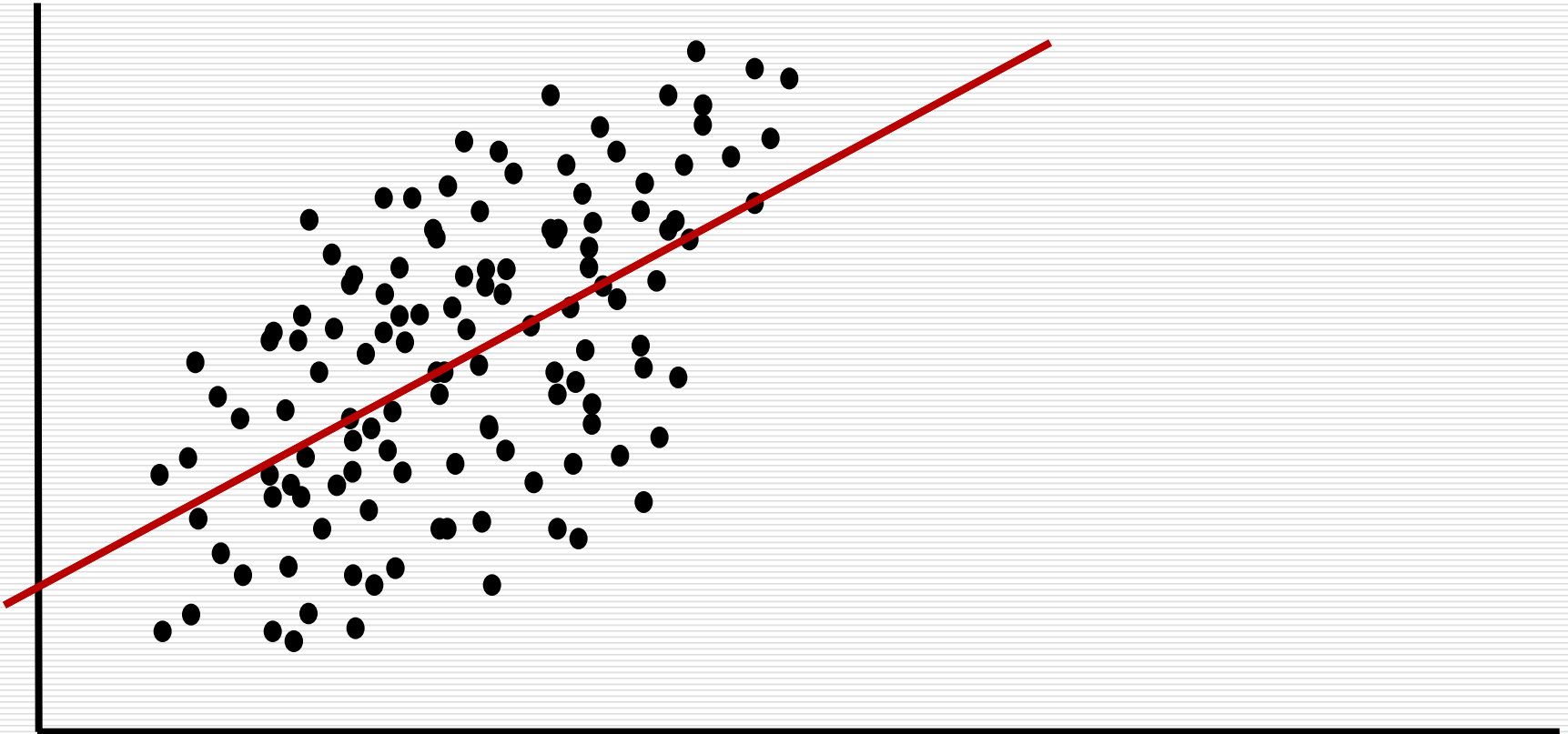


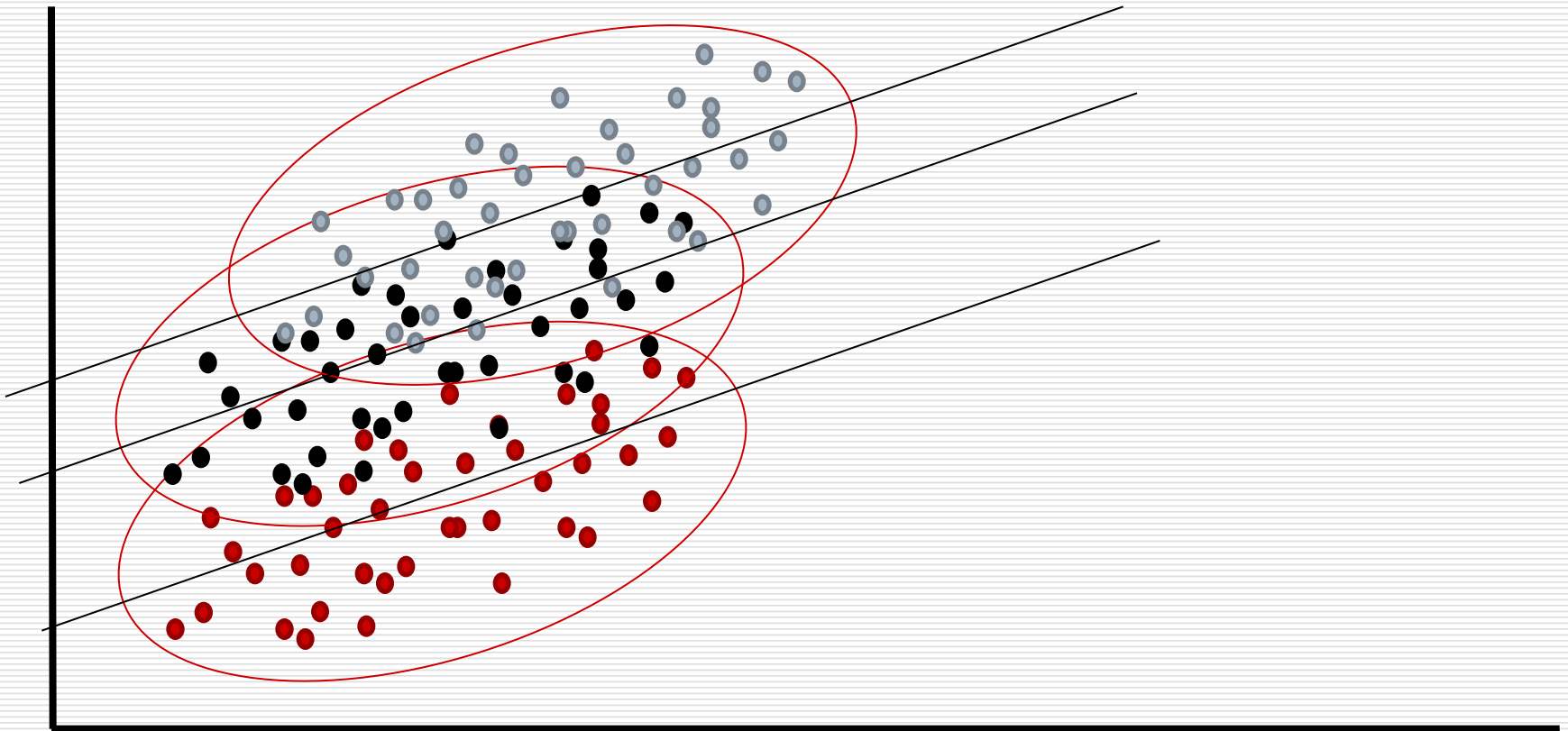
Příklady víceúrovňových dat

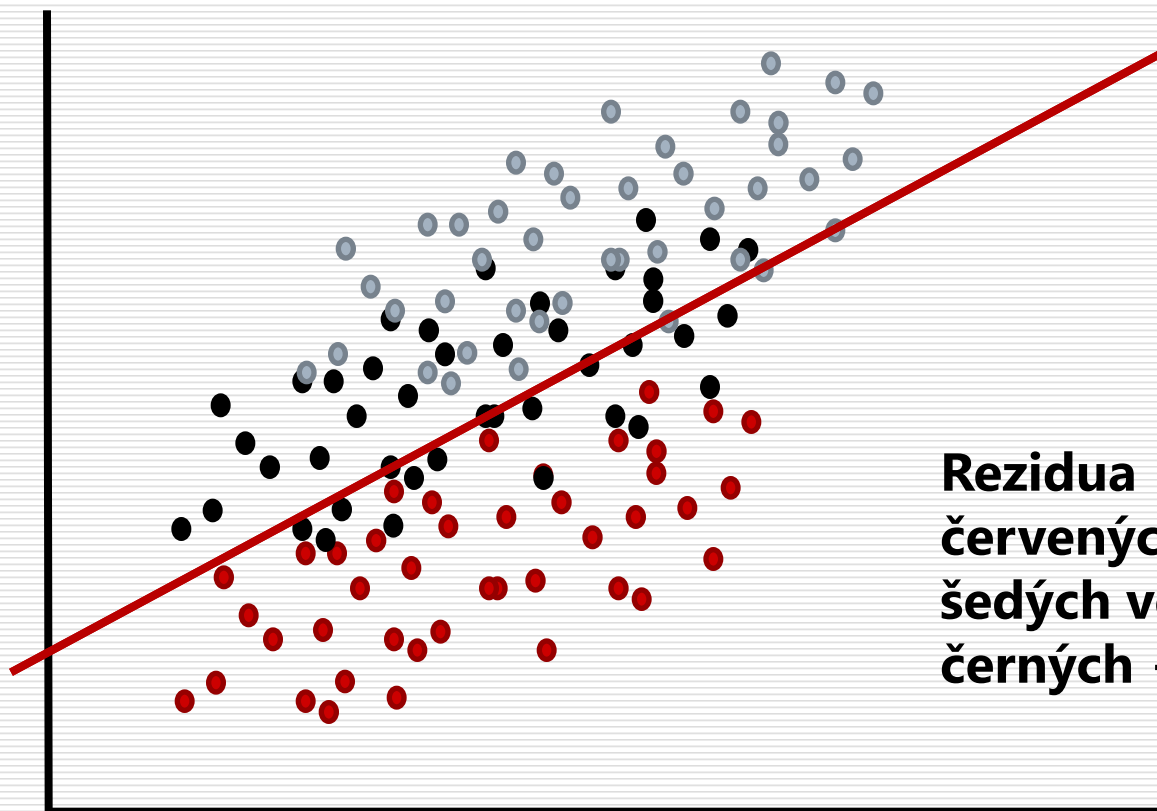
- Žáci(L1) ve třídách (L2) ve školách (L3) v okresech (L4) ...
 - Účastníci experimentu (L1) testování po skupinkách (L2), popř. na různých místech (L2 či L3)
 - ...
 - Opakovaná měření (L1) týchž lidí (L2)
-

Víceúrovňovost způsobuje závislost reziduí

- Pokud proměnná definující skupiny na vyšší úrovni jakkoli souvisí s modelovanou charakteristikou, její ignorování způsobuje to, že rezidua lidí ve skupině si budou podobnější než rezidua lidí napříč skupinami.
 - Může mít podobu třeba rozdílných průměrů skupin nebo rozdílných efektů prediktoru na závislou v různých skupinách
-







Rezidua
červených jsou většinou záporná,
šedých většinou kladná,
černých +-

Odbočka

Autokorelace

- ❑ Jak vyjádříme to, že jsou si rezidua jednotlivců uvnitř skupin podobnější? = neplatnost nezávislosti reziduí
- ❑ Jedním způsobem je udělat na reziduích ANOVu se skupinou jako faktorem...
- ❑ Někdy se k tomu využívá **autokorelace** – korelace proměnné se sebou samotnou posunutou o jeden (lag 1) nebo více případů
- ❑ V SPSS funkce ACF (Analyze > Forecasting > Autocorrelations)
- ❑ Z této části statistiky přichází i test Durbin-Watson

X	X (lag 1)
1	2
2	3
3	45
45	6
6	8
8	7
7	4
4	5
5	21
21	

Chceme tedy zohlednit to, že vztahy, které zjišťujeme se mohou lišit napříč L2 skupinami

Lineární regrese, jak ji známe

□ $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$

- b_0, b_1, b_2 platí pro všechny lidi i
 - Pro predikci dosazujeme každému člověku i jeho hodnoty X_1 a X_2
 - b_0, b_1, b_2 jsou **fixované koeficienty/efekty**
-

Chceme tedy zohlednit to, že vztahy, které zjišťujeme se mohou lišit napříč L2 skupinami

Jak bychom mohli zajistit, aby se b_0 , b_1 nebo b_2 mohly lišit napříč skupinami?

$$\square Y_i = b_{0j} + b_1 X_{1i} + b_2 X_{2i} + e_i$$

$$b_{0j} = b_{00} + u_{0j}$$

- Pro predikci dosazujeme každému člověku i jeho hodnoty X_1 a X_2 , ale průsečík b_0 použijeme takový, které platí ve skupině, do které člověk i patří
 - b_1 , b_2 jsou **fixované koeficienty/efekty**
 - Průsečík b_0 je **náhodný koeficient/efekt**
-

Víceúrovňový model zohledňuje závislost reziduí danou členstvím ve skupinách

$$Y_i = b_0 + b_1 X_i + e_i$$

$$Y_{ij} = b_{0j} + b_{1j} X_{ij} + e_{ij}$$

$$b_{0j} = b_{00} + u_{0j}$$

<1. úroveň>

<2. úroveň>

Průsečík ve skupině j

Průměrný průsečík

Odchylka průsečíku skupiny j od průměrného průsečíku

Odchylky rozptyl

b_0 se stává náhodným koeficientem (random coefficient)

Víceúrovňový model zohledňuje závislost reziduí danou členstvím ve skupinách

$$Y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij} \quad <1. \text{ úroveň}>$$

$$b_{0j} = b_{00} + u_{0j} \quad <2. \text{ úroveň}>$$

Alternativně (dosazením sloučeno)

$$Y_{ij} = (b_{00} + u_{0j}) + b_{1j}X_{ij} + e_{ij}$$

$$Y_{ij} = b_{00} + b_{1j}X_{ij} + (e_{ij} + u_{0j})$$

Random-intercept model

$$Y'_{ij} = b_{00} + b_1 X_{ij} + (e_{ij} + u_{0j})$$

Y'_{ij} predikovaná hodnota Y člověka i (ze skupiny j)

Efekty (fixed effects)

b_{00} průměrný průsečík napříč skupinami

b_1 efekt pro všechny skupiny (*není random*)

Struktura reziduí (kovarianční parametry)

$\text{Var}(u_{0j})$ rozptyl průsečíků napříč skupinami, $u_{0j} \sim N(0, \sigma^2_{u0})$

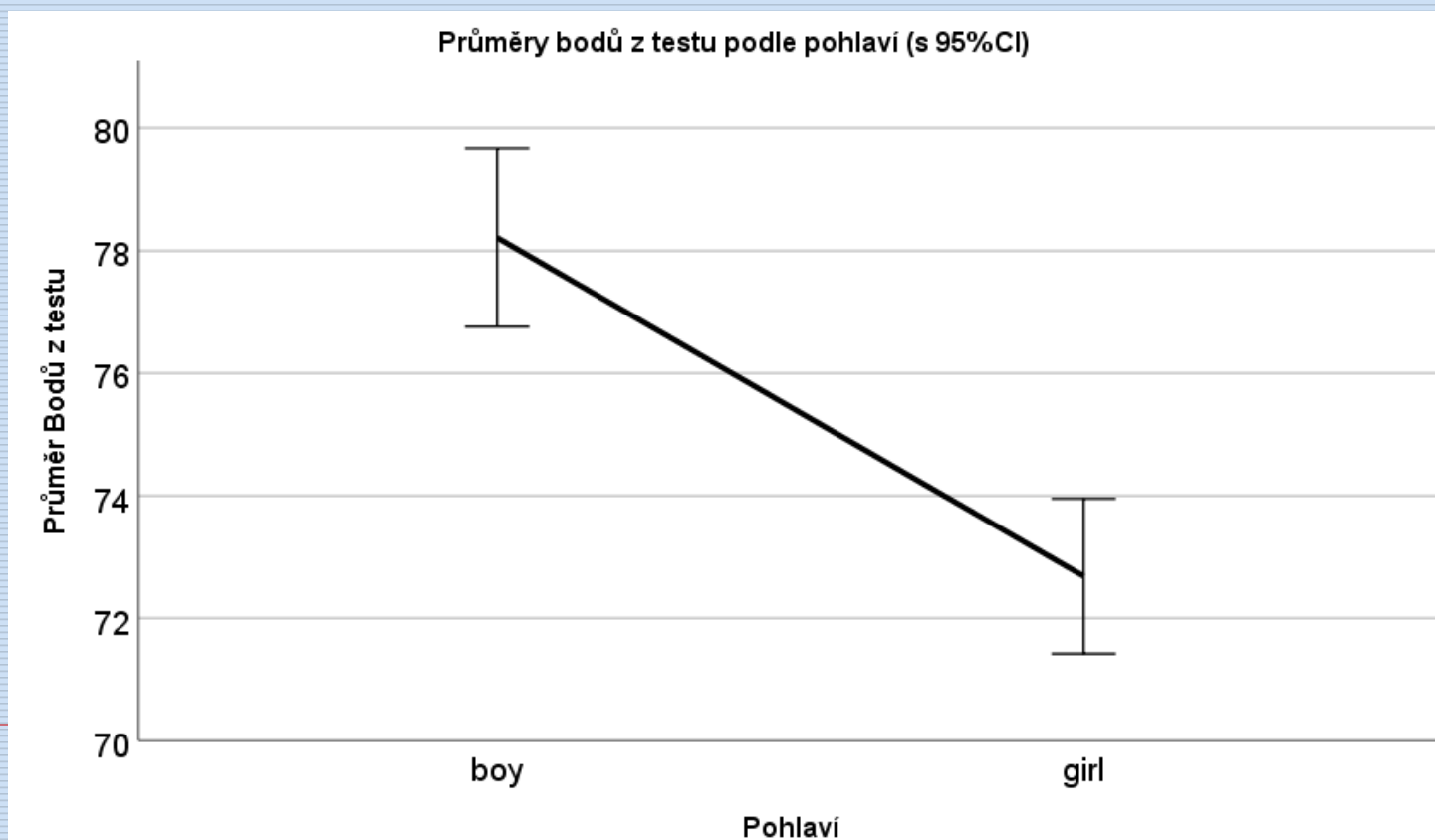
$\text{Var}(e_{ij})$ rozptyl reziduí, $e_{ij} \sim N(0, \sigma^2_e)$

Model má 4 odhadované parametry.

Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

□ Ano, $m_B - m_G = -5,5$ ($t(1903) = 5,57$, $d \approx 0,25$)



Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

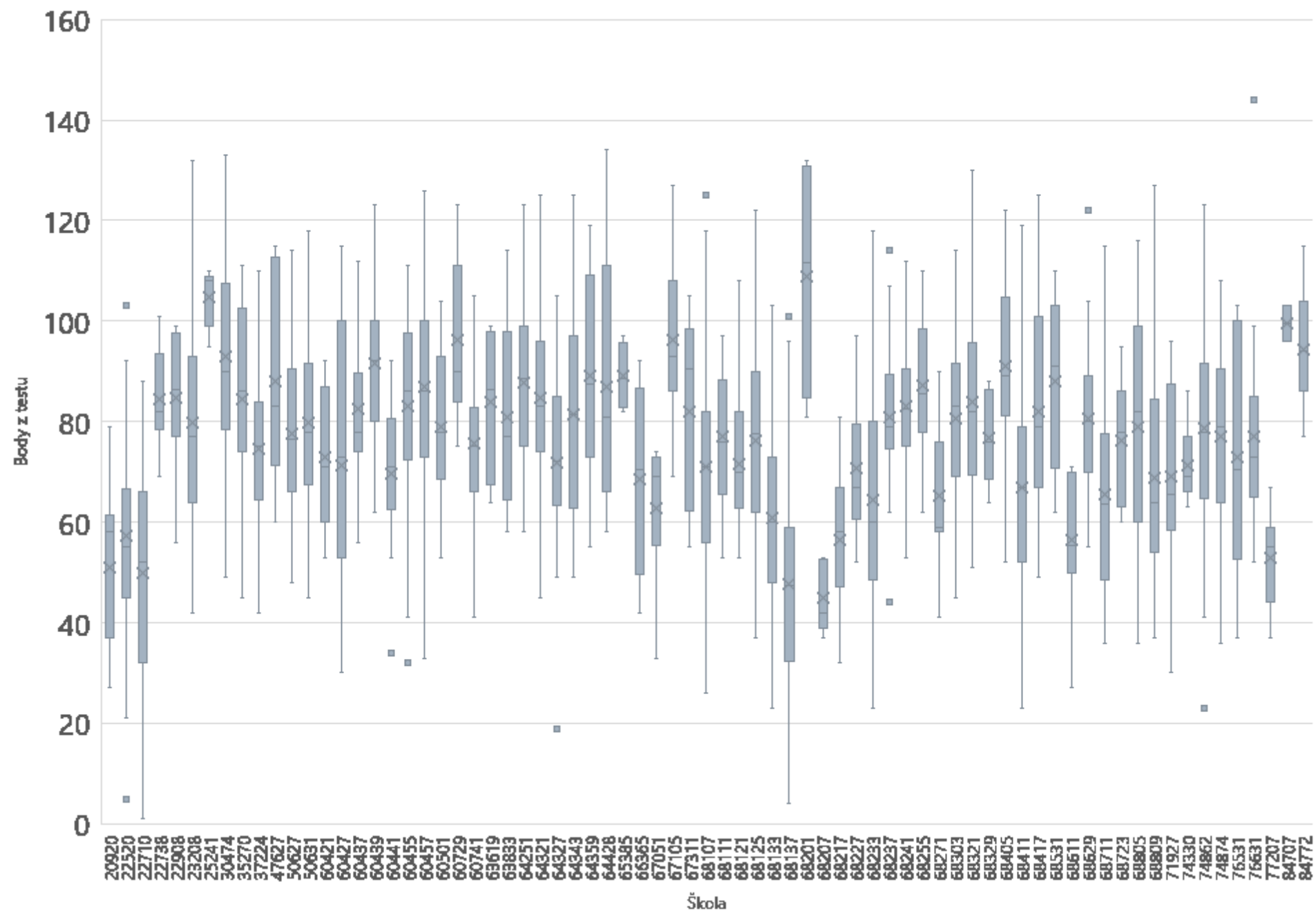
- Ano, $m_B - m_G = -5,5$ ($t(1903) = 5,57$, $d \approx 0,25$)
- $Test_i = b_0 + b_1 Gender_i + e_i$
- $Test_i = 78,2 - 5,5 Gender_i + e_i$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	78,214	0,763		102,548	0,000
	gender Pohlaví	-5,529	0,991	-0,127	-5,578	0,000

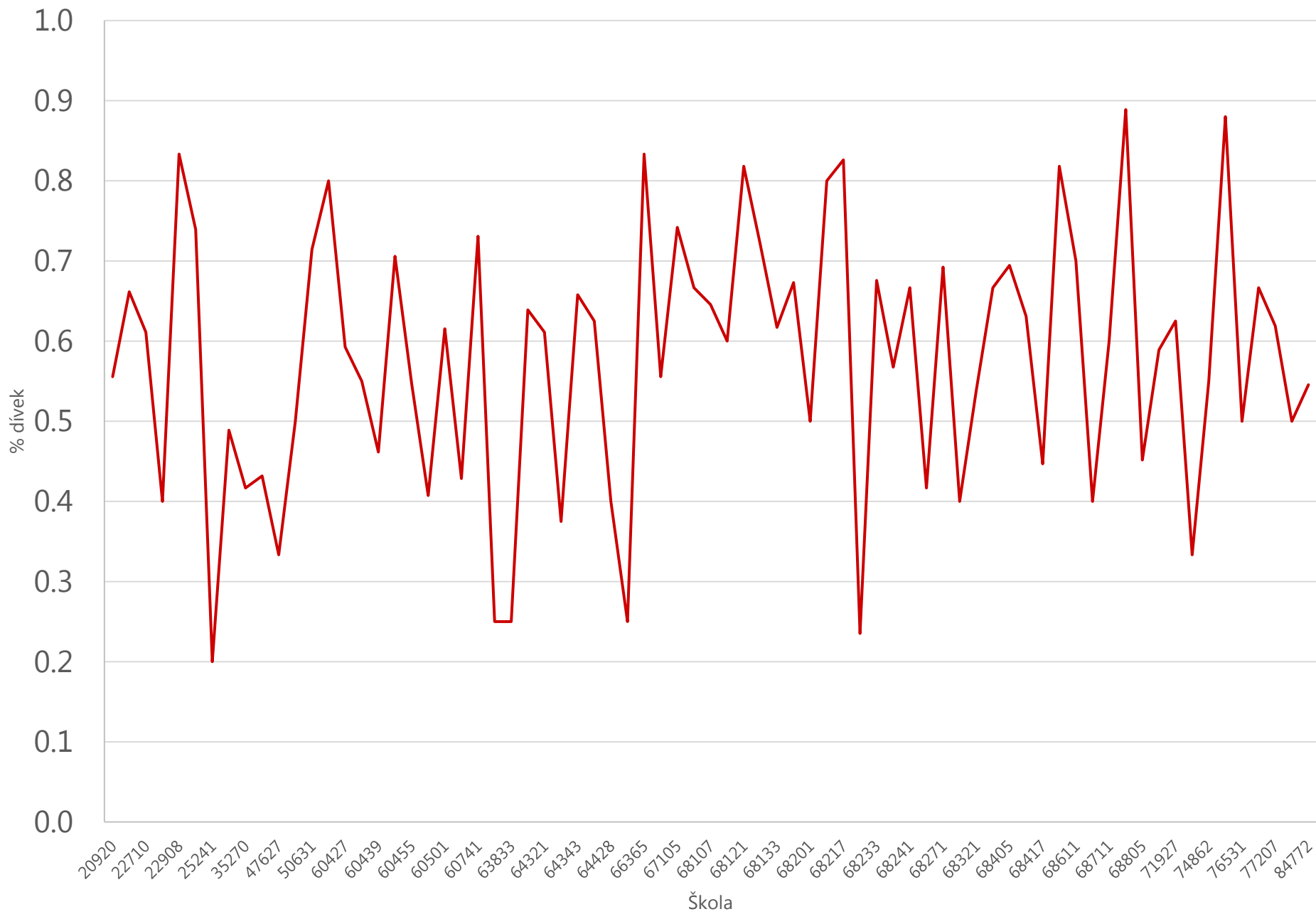
a. Dependent Variable: test Body z testu

- Jenže různé školy se liší průměrnou výkonností, ale i zastoupením pohlaví.

Simple Boxplot of Body z testu by Škola



Podíl dívek napříč školami

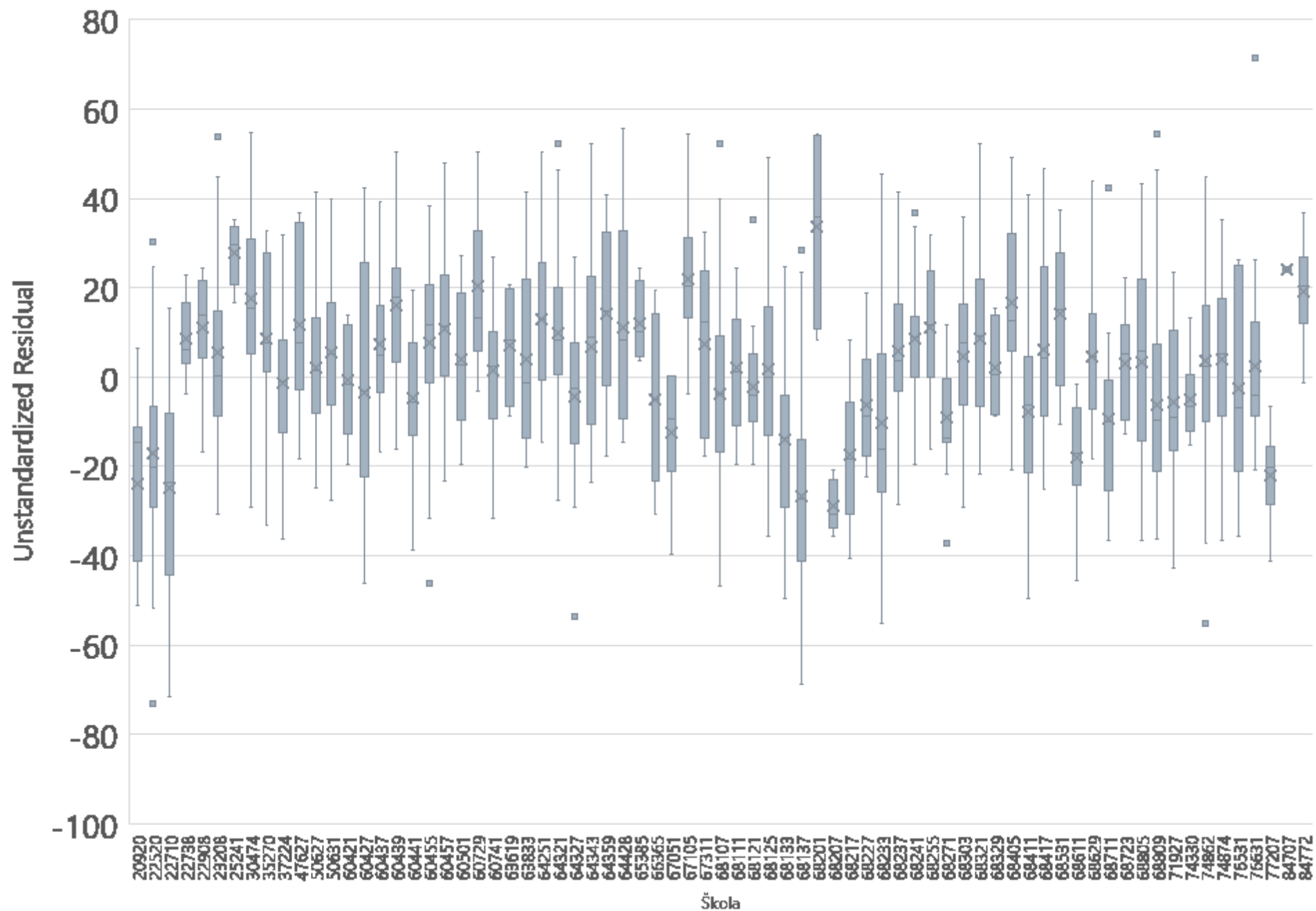


Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

- Ano, $m_B - m_G = -5,5$ ($t(1903) = 5,57$, $d \approx 0,25$)
 - Jenže různé školy se liší jednak průměrnou výkonností, tak zastoupením pohlaví.
 - Pokud by náhodou bylo ve školách s vysokou výkonností více kluků, mohli by kluci vyjít lépe jen díky tomu.
 - Navíc, Durbin-Watson = 1,4 (a lag-1 ACF = 0,31)
-

Boxploty reziduí podle školy



Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

- Multilevel model, kde je zohledněno to, jaké školy žáci pocházejí
- Random-intercept model = předpokládáme, že
 - školy se liší průměrnou výkonností v testu (random Intercept)
 - rozdíl mezi pohlavími je ve všech školách stejný (fixed Slope/effect)
 - ID jsou vnořena do škol – škola je L2 proměnná

$$Test_{i\check{s}} = b_{0\check{s}} + b_1 Gender_i + e_{i\check{s}} \quad <1. \text{ úroveň}>$$

$$b_{0\check{s}} = b_{00} + u_{0\check{s}} \quad <2. \text{ úroveň}>$$

Specifikace ML modelu v SPSS

- Analyze -> Mixed models -> Linear
 - 1. okno: L2 proměnnou do Subjects
 - School do Subjects
 - 2. okno
 - ZP do Dependent variable, kategorické do Factors, spojité so Covariates
 - Fixed:
 - Vložit **všechny** prediktory (a případné interakce), zaškrtnout Include intercept
 - Random:
 - Covariance type: VC, nebo UN
 - Zaškrtnutím „Include intercept“, má-li být průsečík random
 - Ty efekty, které mají být random, vložíme do Model
 - L2 proměnnou dáme do Combinations
 - Estimation: ML
-

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	79,184309	1,504421	86,628	52,634	0,000	76,193929	82,174689
gender	-3,990176	0,857784	1 859,095	-4,652	0,000	-5,672497	-2,307854

a. Dependent Variable: test Body z testu.

Estimates of Covariance Parameters^a

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		318,168776	10,518350	30,249	0,000	298,206884	339,466913
Intercept [subject = School]	Variance	125,116268	24,691785	5,067	0,000	84,982313	184,203981

a. Dependent Variable: test Body z testu.

Empirical Best Linear Unbiased Predictions^a

School Škola	Parameter	Prediction	Std. Error	df	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
20 920	Intercept	-20,160	5,366	612,460	0,000	-30,697	-9,623
22 520	Intercept	-18,439	2,565	524,384	2,281E-12	-23,478	-13,399
22 710	Intercept	-23,484	4,127	841,318	1,756E-8	-31,585	-15,383
22 738	Intercept	5,510	5,164	658,376	0,286	-4,629	15,649
22 908	Intercept	6,303	6,185	436,148	0,309	-5,854	18,460
22 922	Intercept	2,455	2,752	252,522	0,124	-2,912	7,822

Pro každou školu je regresní rovnice
trochu jiná

Průměr výkonu kluků v průměrné škole je 79,18
Rozdíl mezi pohlavími korigovaný na průměrnou
úroveň škol je -3,99 (-5,5 před korekcí).

Rezidua mají $M=0$ a $SD=17,84$

Školní průměrné výkony kluků mají normální
rozložení s průměrem 79,18 a $SD=11,19$

Školní průměrné výkony holek jsou o 3,99 nižší.

Nepodmíněný model průměrů

Unconditional means model, variance components

- Model bez prediktorů zohledňující strukturu dat
- Pouze dělí rozptyl na reziduální rozptyl a rozptyl průměrů skupin
 - ICC=rozptyl průměrů/(rozptyl průměrů+reziduální rozptyl)
 - ICC= jaká část rozptylu výkonů je vysvětlitelná pouze rozdíly mezi školami?
- 3 parametry – průměrný průměr škol (b_{00}), rozptyl průměrů škol, rozptyl reziduí (variabilita uvnitř škol)

$$Test_{ij} = b_{0\check{s}} + e_{i\check{s}} \quad \langle 1. \text{ úroveň} \rangle$$

$$b_{0\check{s}} = b_{00} + u_{0\check{s}} \quad \langle 2. \text{ úroveň} \rangle$$

Random-intercepts model

□ Předpokládá,

- že jednotky vyššího řádu se liší svým průměrem,
 - a že průměry mají normální rozložení,
 - že efekty prediktorů jsou stejné (fixed) napříč všemi jednotkami vyššího řádu
 - že rezidua jsou napříč jednotkami vyššího řádu stejná
-

Random-slopes model

- Předpokládá,
 - že všechny jednotky vyššího řádu mají stejný průměr,
 - že efekt prediktoru je v každé jednotce vyššího řádu jiný a
 - že tyto efekty mají nějakou průměrnou hodnotu a nějakou variabilitu
-

Random-slopes model

$$Y_{ij} = b_0 + b_{1j}X_{ij} + e_{ij} \quad \langle 1. \text{ úroveň} \rangle$$
$$b_{1j} = b_{10} + u_{1j} \quad \langle 2. \text{ úroveň} \rangle$$

Alternativně (dosazením sloučeno)

$$Y_{ij} = b_0 + (b_{10} + u_{1j})X_{ij} + e_{ij}$$
$$Y_{ij} = b_0 + b_{10}X_{ij} + (e_{ij} + u_{1j})$$

Jen zřídka má smysl předpokládat náhodné efekty při fixovaných průsečících!

Random intercept and slope model

Předpokládá,

- že jednotky vyššího řádu mají různé průměry(průsečíky),
 - že efekt prediktoru je v každé jednotce vyššího řádu jiný,
 - že tyto průsečíky i efekty mají nějakou průměrnou hodnotu a nějakou variabilitu napříč skupinami,
 - že reziduální rozptyl je napříč skupinami konstantní.
 - Lze uvažovat i to, že mezi hodnotou průsečíku a efektu je nějaká korelace.
-

Random intercept and slope model

$$Y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij} \quad \langle 1. \text{ úroveň} \rangle$$

$$b_{0j} = b_{00} + u_{0j} \quad \langle 2. \text{ úroveň} \rangle$$

$$b_{1j} = b_{10} + u_{1j} \quad \langle 2. \text{ úroveň} \rangle$$

Alternativně (dosazením sloučeno)

$$Y_{ij} = b_{00} + b_{10}X_{ij} + (e_{ij} + u_{0j} + X_{ij}u_{1j})$$

Random intercept and slope model

$$Y_{ij} = b_{00} + b_{10}X_{ij} + (e_{ij} + u_{0j} + X_{ij}u_{1j})$$

Efekty (fixed effects)

b_{00} průměrný průsečík napříč skupinami

b_{10} průměrný efekt pro všechny skupiny

Struktura reziduí (kovarianční parametry)

$\text{Var}(u_{0j})$ rozptyl průsečíků, $u_{0j} \sim N(0, \sigma^2_{u0})$

$\text{Var}(u_{1j})$ rozptyl efektů, $u_{1j} \sim N(0, \sigma^2_{u1})$

$\text{Var}(e_{ij})$ rozptyl reziduí, $e_{ij} \sim N(0, \sigma^2_e)$

Model má 5 odhadovaných parametrů.

Šestý $\text{Cov}(u_{0j}, u_{1j})$ kovariance průsečíků s efekty

Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

□ Zvažme, zda se mohou lišit i efekty napříč školami

$$Test_{i\check{s}} = b_{00} + b_{10}G_{i\check{s}} + (e_{ij} + u_{0\check{s}} + G_{i\check{s}}u_{1\check{s}})$$

Příklad – Skotské zkoušky

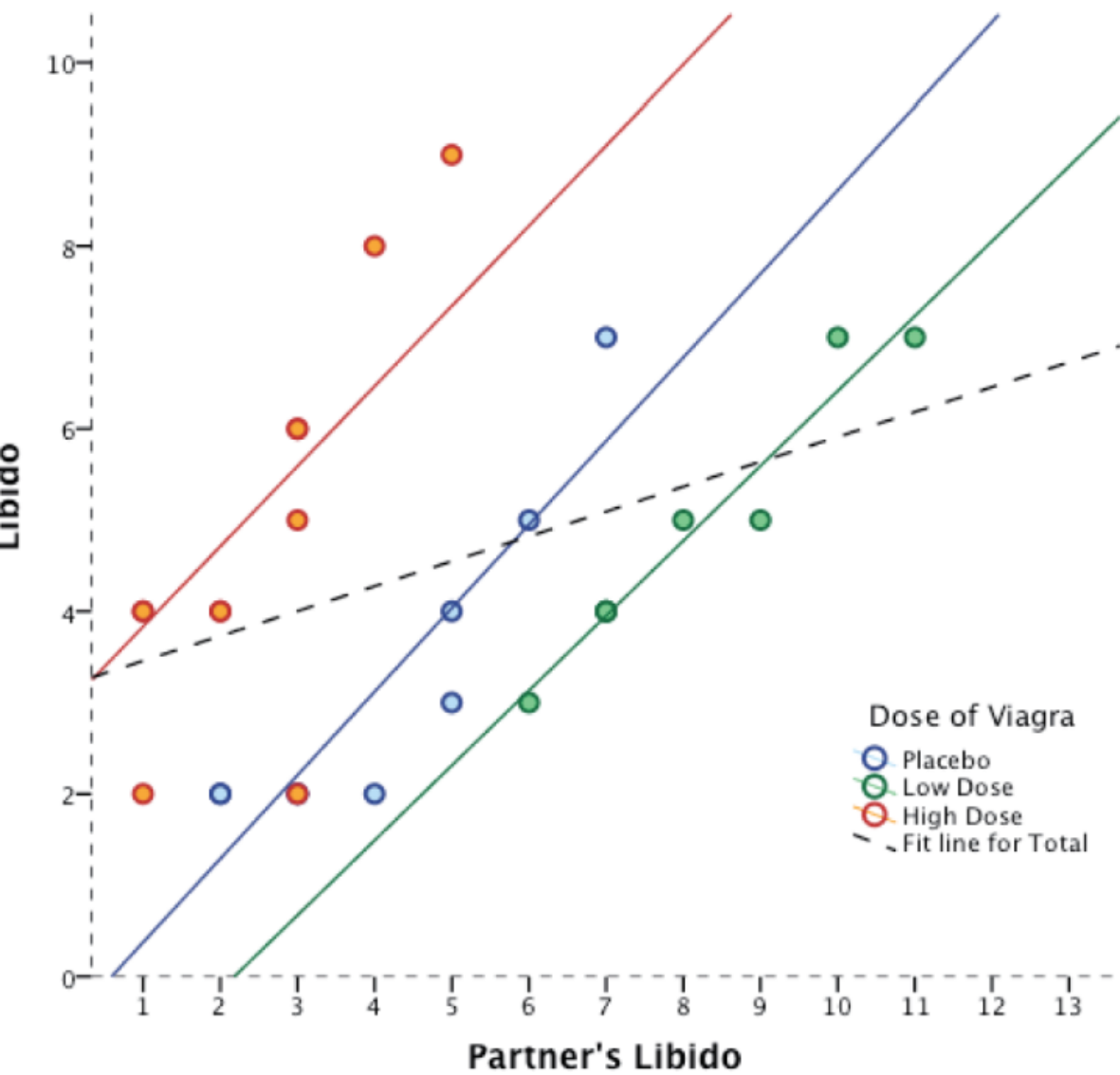
$$\square Test_i = (79 \pm 12) - (4,0 \pm 2,7)G_i \pm 17,8$$

Jsou-li kluci 0 a holky 1, pak...

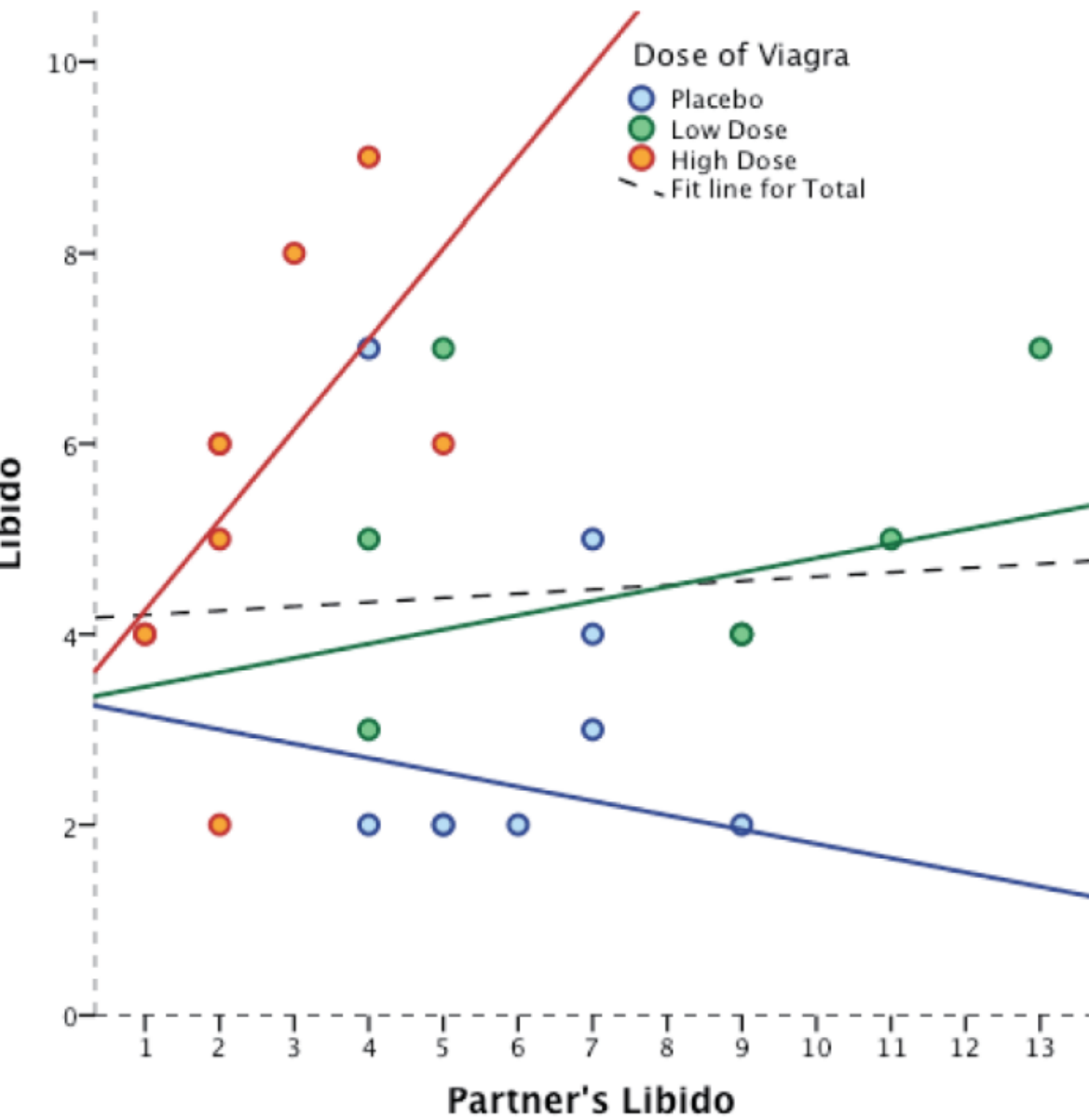
- \square Výkon průměrného kluka v průměrné škole je 79, přičemž školy se liší tak, že výkony průměrných kluků mají $SD=12$.
 - \square Průměrná holka má v průměrné škole o 4,0 bodu míň.
 - \square I když napříč školami mají rozdíly mezi průměrnou holkou a průměrným klukem $SD=2,7$, rozptyl efektů není signifikantně odlišný od 0.
 - \square Čím vyšší je průměr kluků ve škole, tím nižší (větší) je rozdíl jejich průměru od průměru holek, $r=-0,3$
-

Shrnutí

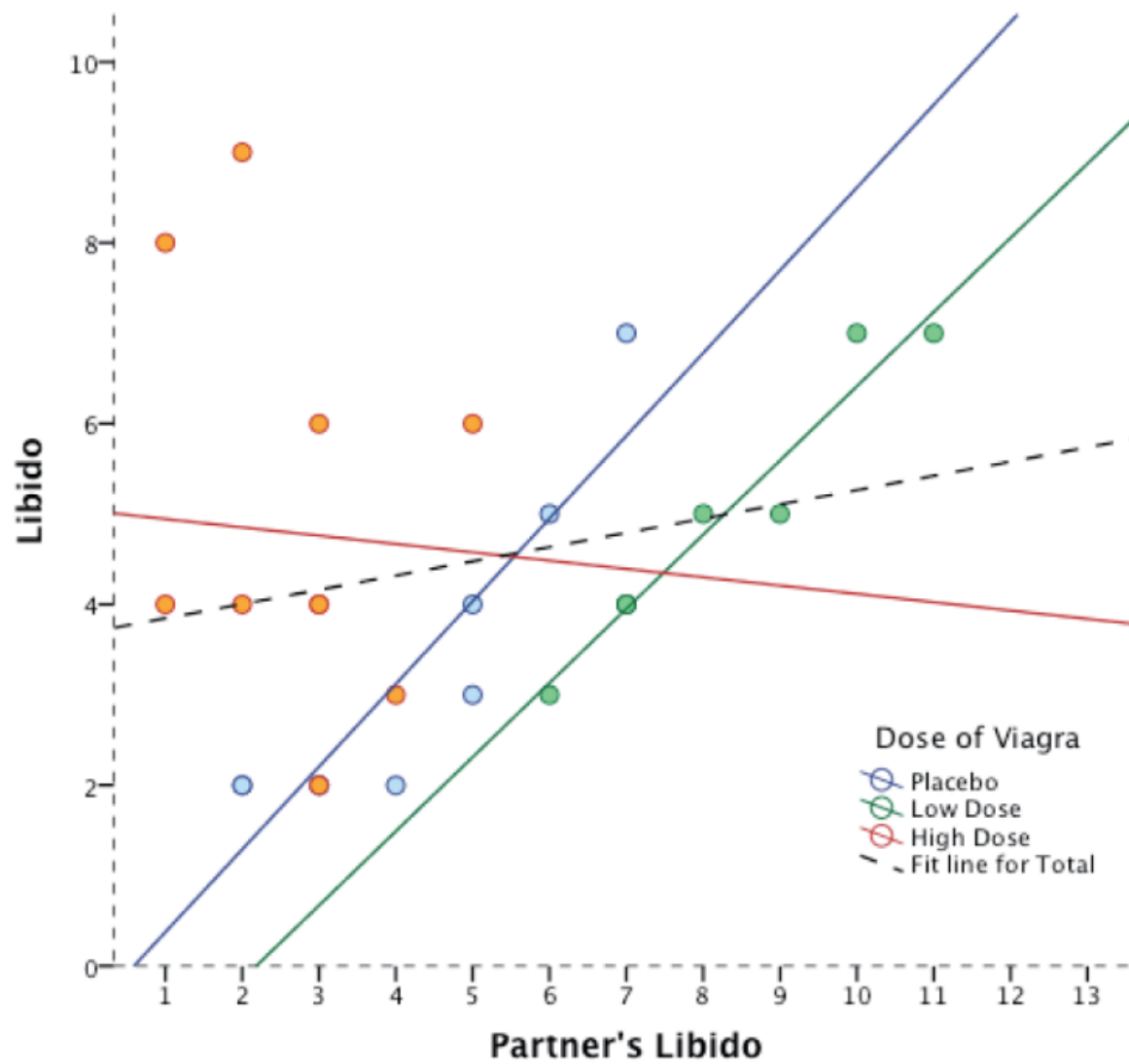
- ❑ Multilevel modely nám umožňují modelovat to, že některé parametry regresního modelu se mohou pro různé skupiny lišit.
- ❑ Od moderace se to liší tím, že různost parametrů má podobu normálního rozložení. Nezajímáme se o hodnoty pro jednotlivé skupiny – ze vzorku skupin usuzujeme na populaci skupin
- ❑ S tím je spojen předpoklad, že vzorek jednotek druhé úrovně (skupin) je reprezentativním vzorkem populace skupin



**Random Intercept,
Fixed Slope**



**Fixed Intercept,
Random Slope**



**Random Intercept,
Random Slope**

Prediktor na úrovni skupin

- Zatím jsme měli prediktor na L1 - pohlaví
 - Do modelu lze vložit i prediktor, který vysvětluje rozdíly mezi skupinami.
 - Například „nóblóznost“ spádové oblasti školy – ***nbrhd***
-

Random intercept and slope model s prediktorem na úrovni skupin S

$$Y_{ij} = b_{0j} + b_{1j}X_{ij} + e_{ij} \quad <1. \text{ úroveň}>$$

$$b_{0j} = b_{00} + b_{01}S_{ij} + u_{0j} \quad <2. \text{ úroveň}>$$

$$b_{1j} = b_{10} + b_{11}S_{ij} + u_{1j} \quad <2. \text{ úroveň}>$$

- S může být prediktorem náhodného průsečíku, směrnice, nebo obojího
 - Jeho zařazení pak vysvětluje rozptyl daného náhodného parametru
-

Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

A liší se i efekt školy, pokud je v chudém sousedství?

$$Test_{ij} = b_{0\check{s}} + b_1 G_i + e_{i\check{s}} \quad \langle 1. \text{ úroveň} \rangle$$

$$b_{0\check{s}} = b_{00} + b_{01} N_{\check{s}} + u_{0\check{s}} \quad \langle 2. \text{ úroveň} \rangle$$

$$Test_{i\check{s}} = b_{00} + b_{01} N_{\check{s}} + b_{10} G_{i\check{s}} + (e_{ij} + u_{0\check{s}} + u_{1\check{s}})$$



Shoda modelu s daty

- Podobně jako u logistické regrese vyjadřují celkový fit modelu informační kritéria založená na $-2LL$
 - AIC, AICC, CAIC, BIC
 - Vnořené modely lze srovnávat LRT – rozdíl $-2LL$ dvou vnořených modelů má chí-kvadrát rozložení s df rovným rozdílu v počtu parametrů mezi srovnávanými modely (*nefunguje s REML*)
-

REML vs ML

- Dva způsoby **odhadu parametrů** multilevel modelu
 - ML – Maximum likelihood
 - Podhodnocuje odhady rozptylů – reziduálního i random parametrů
 - Produkuje -2LL, které mají chíkvadrát rozložení umožňující srovnávání modelů pomocí LRT
 - REML – Restricted Maximum Likelihood
 - Poskytuje nezkreslené odhady rozptylů
 - Produkuje -2LL, který se neá použít pro LRT
 - Výchozí možnost v SPSS
 - Reportujeme REML parametry, modely srovnáváme mezi sebou prostřednictvím -2LL hodnot získaných ML odhadem
-

Velikost účinku

Multilevel alternativy R^2

ICC – vnitrotřídní korelační koeficient

- Random means model dělí rozptyl na reziduální rozptyl a rozptyl způsobený rozdílnými průměry skupin
 - $ICC = \text{rozptyl interceptů} / (\text{rozptyl interceptů} + \text{reziduální rozptyl})$
 - ICC = jaká část rozptylu výkonů je vysvětlitelná pouze rozdíly mezi L2 skupinami (př. školami)
 - Když přidáme L1 prediktor, měl by klesnout reziduální rozptyl $\rightarrow R^2_{\text{within}} = 1 - (\sigma^2_{e(s \text{ prediktorem})} / \sigma^2_{e(\text{bez prediktoru})})$
 - Interpretujeme jako R^2 v běžné regresi
 - L2 prediktor by měl snížit rozptyl náhodného efektu $\rightarrow R^2_{\text{between}} = 1 - (\sigma^2_{u(s \text{ prediktorem})} / \sigma^2_{u(\text{bez prediktoru})})$
 - Interpretujeme: prediktor vysvětlil x% rozptylu průsečíků
-

Typy kovariančních struktur náhodných koeficientů

- Ve výše popsaných modelech jsou smysluplné jen 2 volby a hraje to roli, jen, když máme v modelu více než 1 náhodný koeficient
- VC – Variance components – náhodné koeficienty nekorelují
- UN – Unstructured – náhodné koeficienty mohou korelovat

Předpoklady

- Jako lineární regrese
 - Je-li závislost reziduí modelovatelná (=je to skupinami), vyléčí se tím problémem
 - Dostatečný počet jednotek i na druhé a vyšší úrovni (přibližně >20) pro dobrý odhad σ^2_u
-

Benefity Multilevel/Mixed modelu

- ❑ V mnoha situacích vyšší síla testu
 - ❑ Vyšší tolerance k chybějícím datům
 - ❑ Jednotné uvažování o spojitých a diskretních proměnných
 - ❑ Možnost modelovat heteroskedasticitu
-



Longitudinální, repeated data

1. úroveň: měření

2. úroveň: jednotlivec

- Čas, či pořadí měření je proměnnou na 1. úrovni.
 - Čas může nabývat různé hodnoty pro různé lidi v různé časy měření
- Charakteristiky jednotlivců jsou proměnnými na 2. úrovni.

LATENT GROWTH-CURVE MODELING

ŠIROKÁ

VS.

DLOUHÁ DATA

ID	EDA klid	EDA stres1	EDA stres2
101A	1	2	3
102A	4	5	6
...			
199A	5	3	5

ID	Stres	EDA
101A	Klid	1
101A	Stres1	2
101A	Stres2	3
102A	Klid	4
102A	Stres1	5
102A	Stres2	6
...		
199A	Klid	5
199A	Stres1	3
199A	Stres2	5

Převod širokých dat na dlouhá a zpět

- SPSS >> Data >> Restructure
(VARSTOCASES)
-