PSY252
Statistická analýza dat v psychologii II

Víceúrovňový lineární model

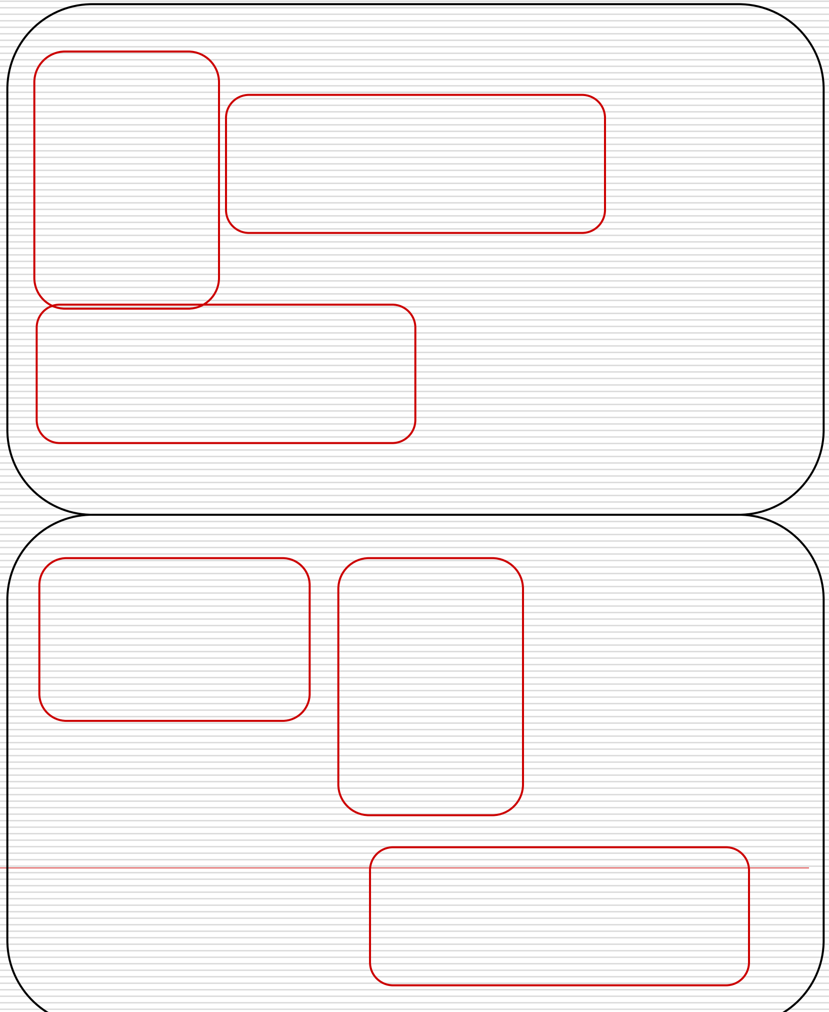
(multilevel, hierarchical, mixed, random-coefficients model)

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

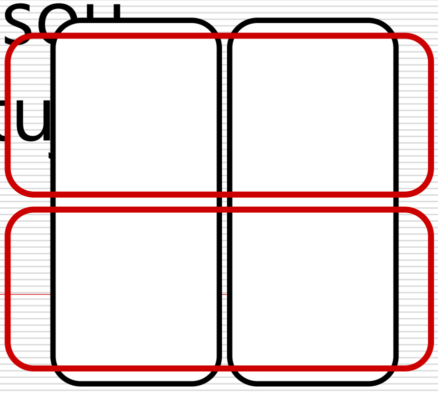
Víceúrovňová data

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
  - Konkrétní třída je jen v jedné škole, žák je členem jen jedné třídy
- Protikladem pro vnořené faktory jsou zkřížené (crossed) faktory – vyskytu všechny kombinace jejich hodnot



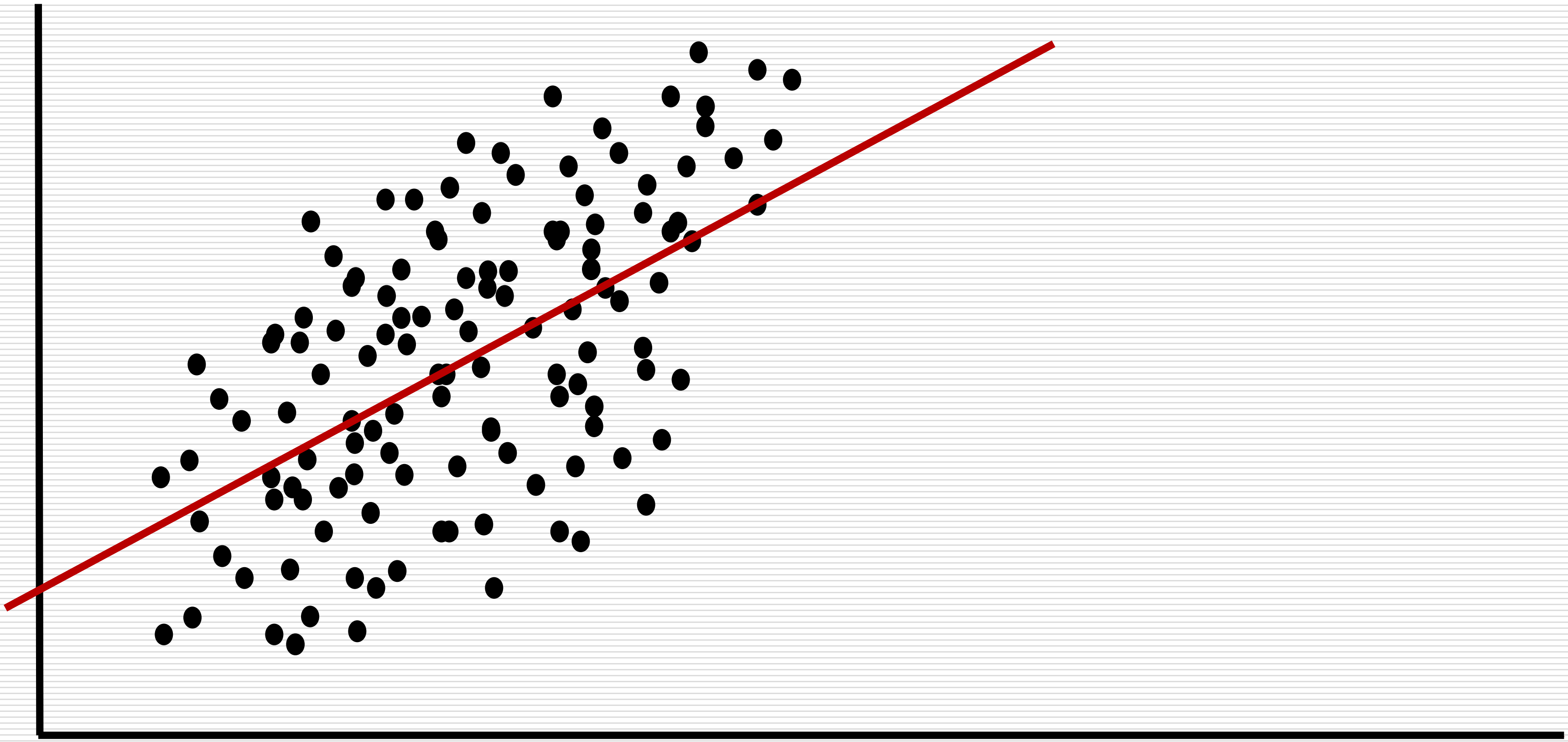
Příklady víceúrovňových dat

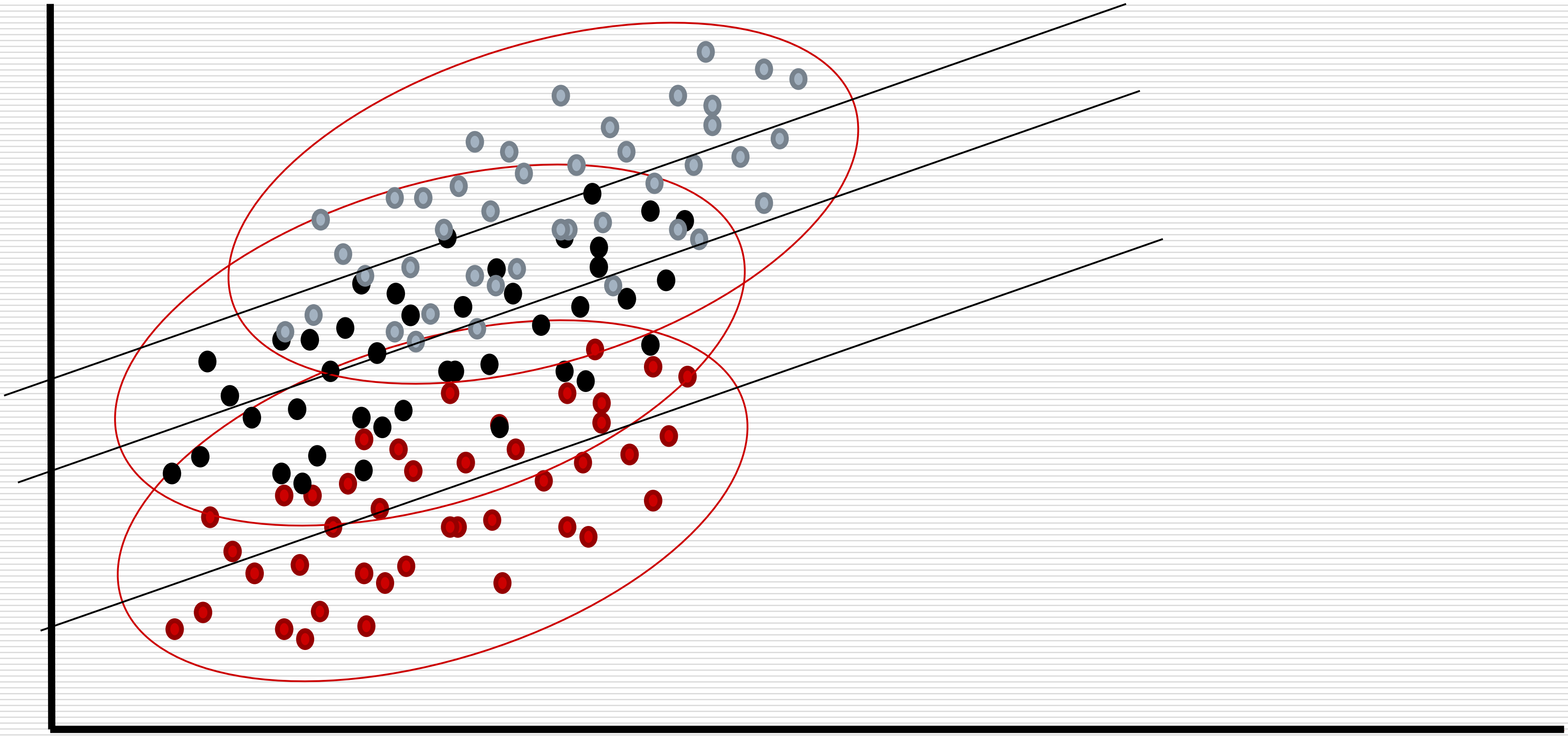
- Žáci(L1) ve třídách (L2) ve školách (L3) v okresech (L4) …
- Účastníci experimentu (L1) testovaní po skupinkách (L2), popř. na různých místech (L2 či L3)
- …
- Opakovaná měření (L1) týchž lidí (L2)

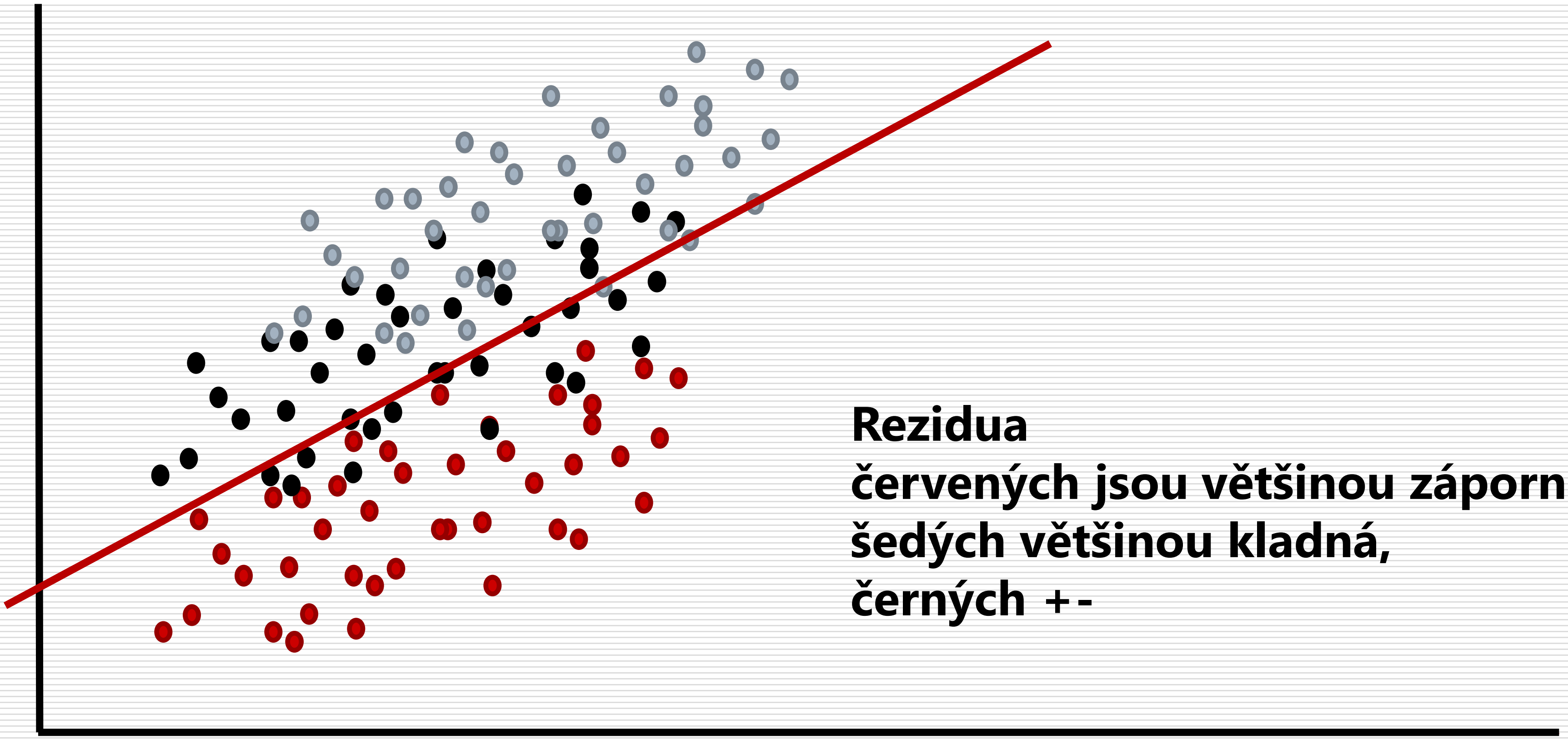
Víceúrovňovost způsobuje závislost reziduí

- 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 ba závislou v různých skupinách







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

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
  - 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?

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

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

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

Víceúrovňový model zohledňuje závislost reziduí danou členstvím ve skupinách

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

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

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

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

$b_{0j} = b_{00} + u_{0j}$ <*2. úroveň*>

Alternativně (dosazením sloučeno)

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

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

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

Y predikovaná proměnná

Efekty (fixed effects)

b_{00} průměrný průsečík napříč skupinami

b_{1j} efekt pro všechny skupiny (*není random*)

Struktura reziduí (kovarianční parametry)

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

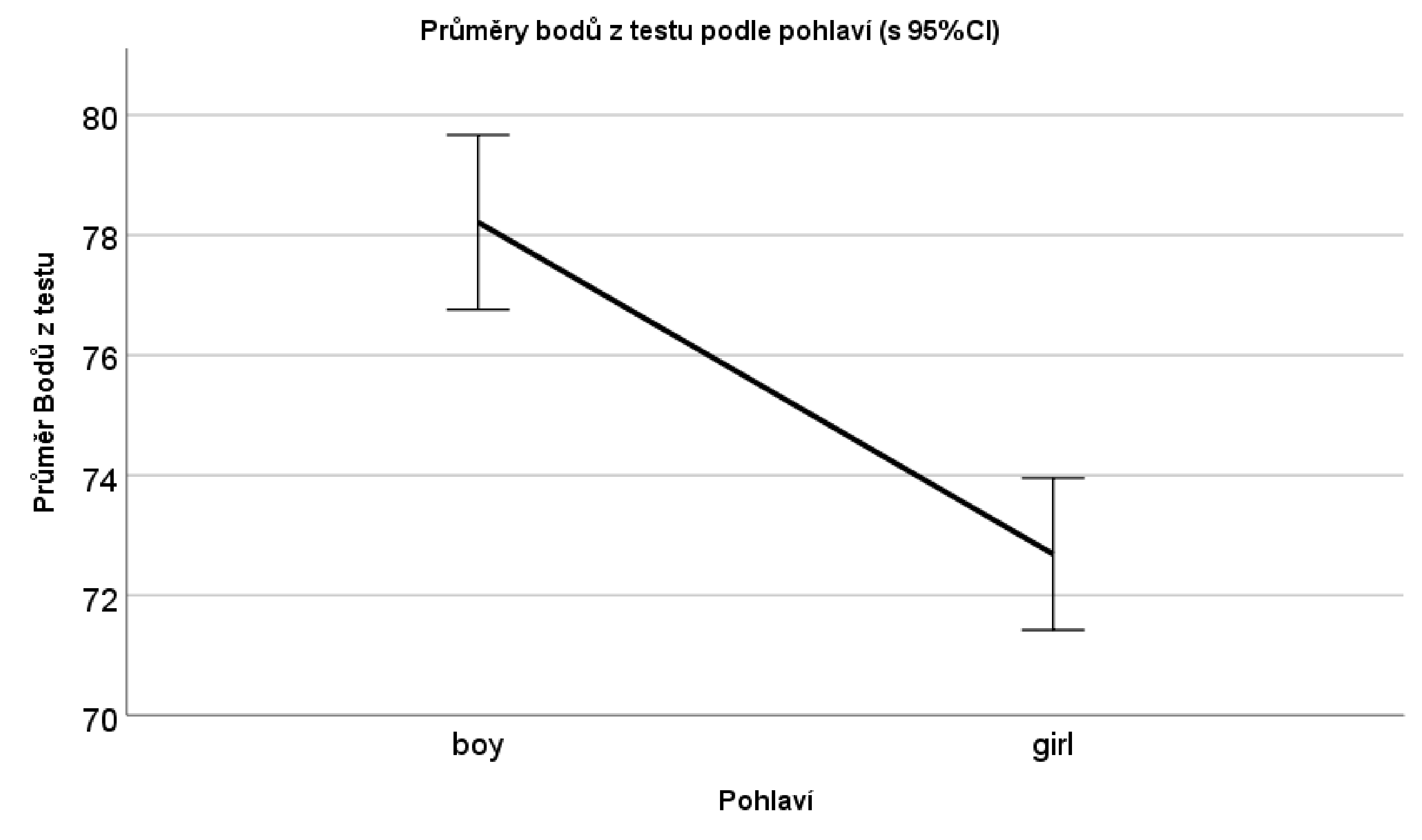
Var(e_{ij}) rozptyl reziduí, $e_{ij} \sim N(0, \sigma^2_{e})$

Model má 4 odhadované parametry.

Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

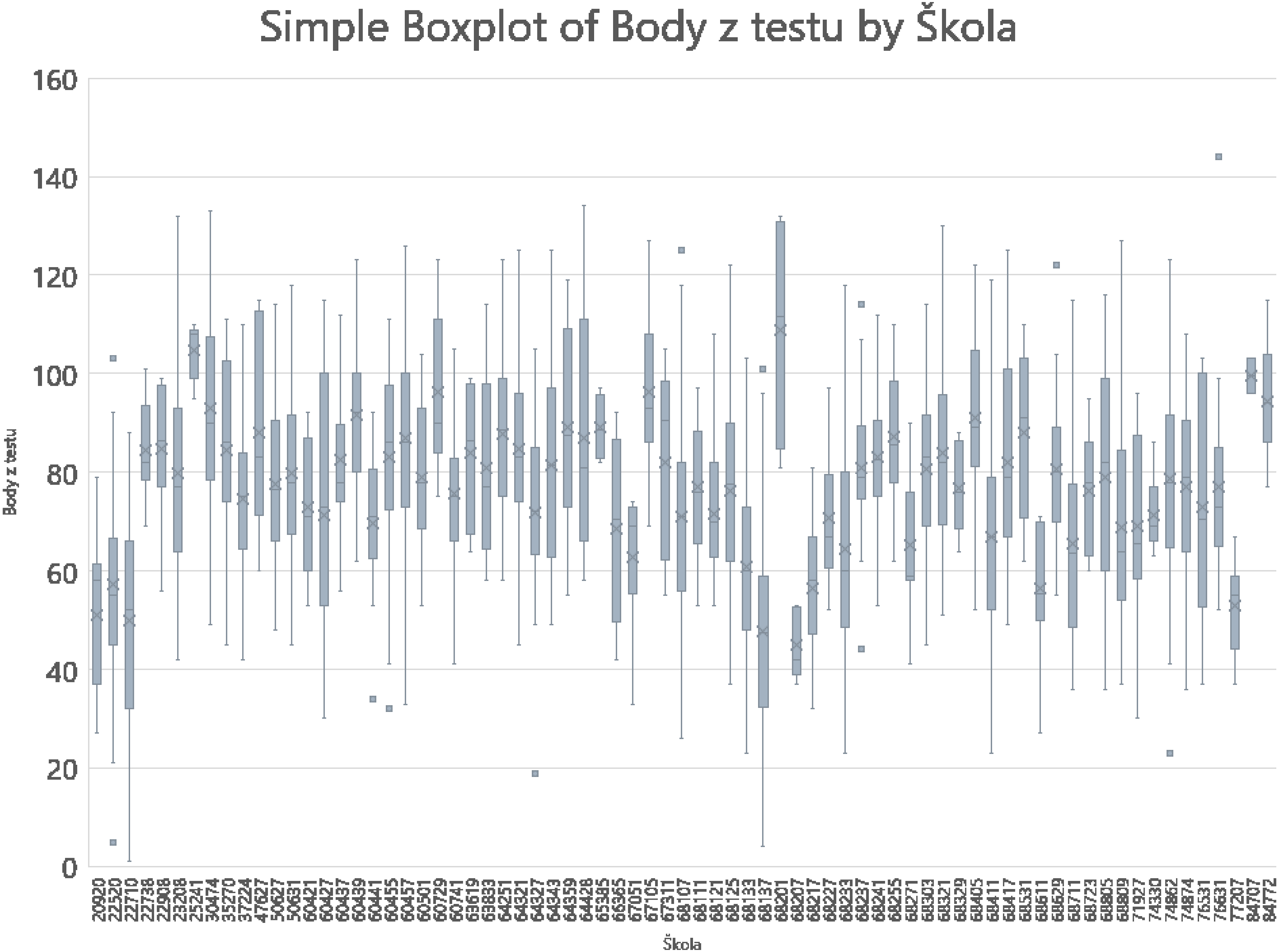
- Ano, m_B-m_G=-5,5 (t(1903)=5,57, d≈0,25)

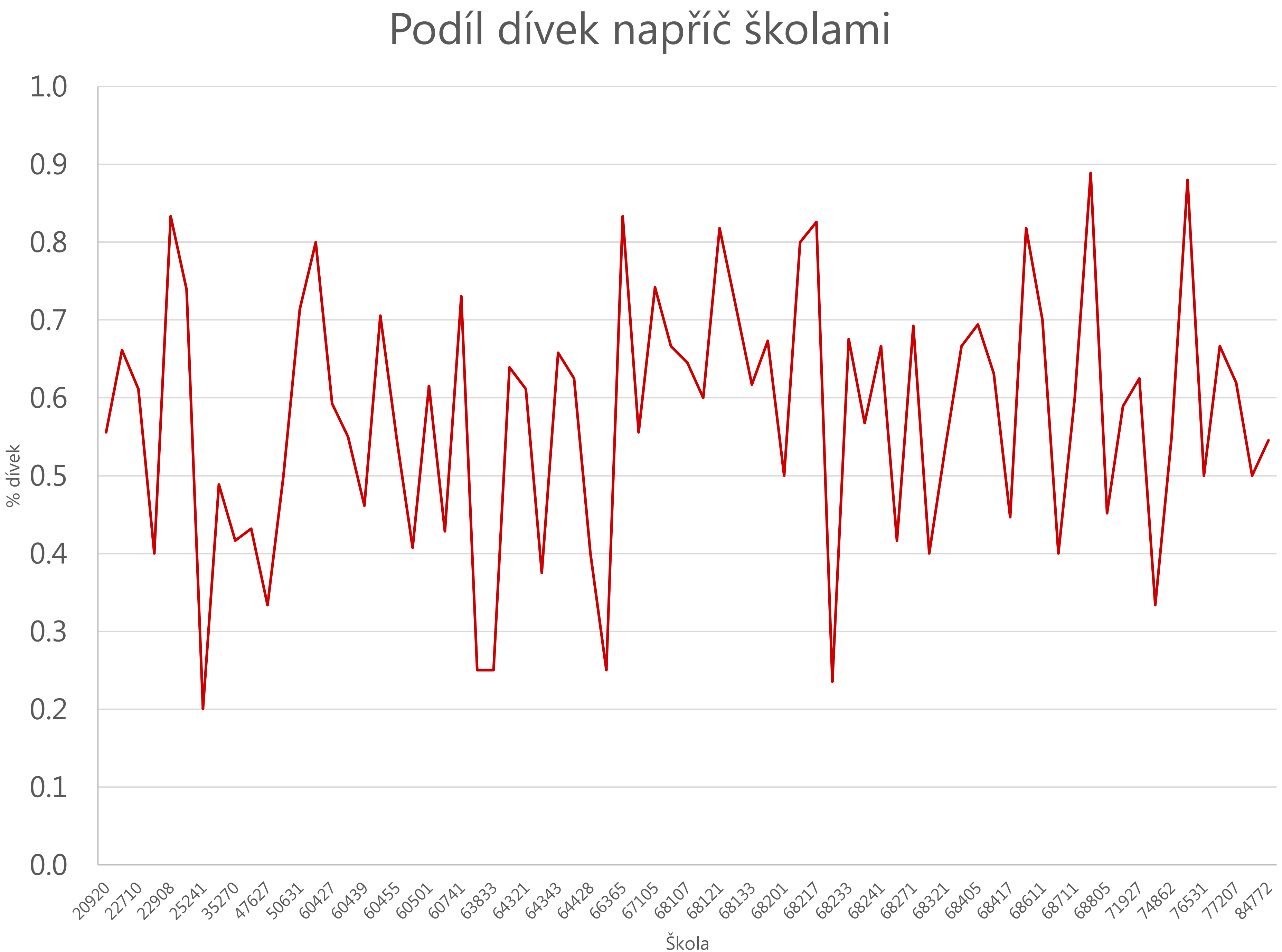


Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

- Ano, m_B-m_G=-5,5 (t(1903)=5,57, d≈0,25)
- Jenže různé školy se liší jednak průměrnou výkonností, tak zastoupením pohlaví.


Simple Boxplot of Body z testu by Škola
Body z testu
160
140
120
100
80
60
40
20
0
20920
22520
22710
22738
22908
23208
25241
30474
35270
37224
47627
50627
50631
60421
60427
60437
60439
60441
60455
60457
60501
60729
60741
63619
63833
64251
64321
64327
64343
64359
64428
65385
66365
67051
67105
67311
68107
68111
68121
68125
68133
68137
68201
68207
68217
68227
68233
68237
68241
68255
68271
68303
68321
68329
68405
68411
68417
68531
68611
68629
68711
68723
68805
68809
71927
74330
74862
74874
75531
76631
77207
84707
84772
Škola

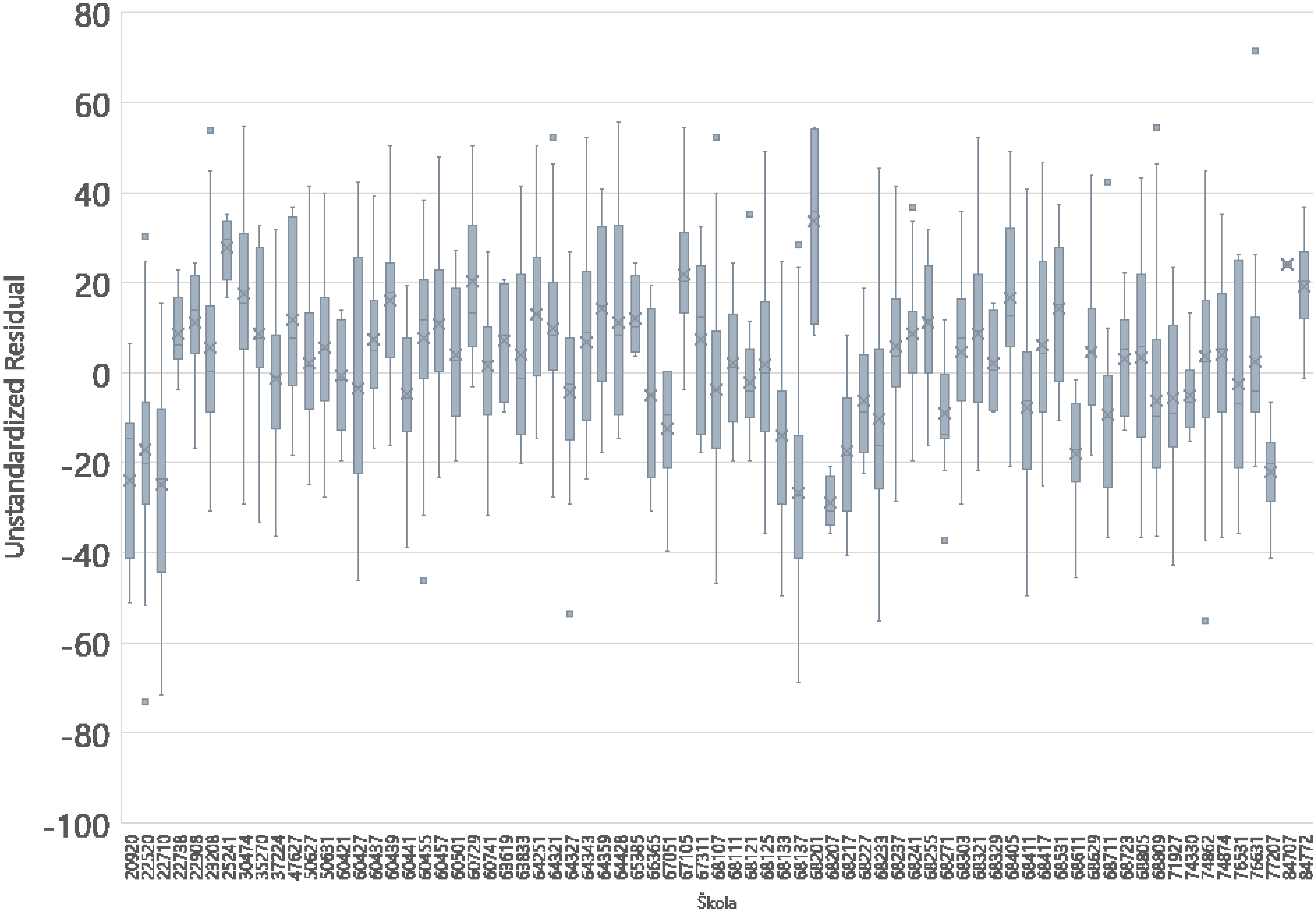

Podíl dívek napříč školami
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
% dívek
20920
22710
22908
25241
35270
47627
50631
60427
60439
60455
60501
60741
63833
64321
64343
64428
66365
67105
68107
68121
68133
68201
68217
68233
68241
68271
68321
68405
68417
68611
68711
68805
71927
74862
76531
77207
84772
Škola

Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

- Ano, m_B-m_G=-5,5 (t(1903)=5,57, d≈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

Boxploty reziduí podle školy



Příklad – Skotské zkoušky

Liší se holky a kluci ve výsledku testů?

- Multilevel model, kde je zohledněno to, jaké školy žáci pochází
- 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

$Test_{ij} = b_{0Š} + b_1 Gender_i + e_{iŠ}$ <1. úroveň>

$b_{0Š} = b_{00} + u_{0Š}$ <2. 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

Průměr výkonu kluků v průměrné škole je 79,18

Rozdíl mezi pohlavími korigovaný na průměrnou úroveň škol ji -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.

ODBOČKA

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)

$$Test_{ij} = b_{0Š} + e_{iŠ} \qquad \langle 1.\ \text{úroveň} \rangle$$

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

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

$Y_{ij} = b_0 + b_{1j} X_{ij} + e_{ij}$ <1. úroveň>

$b_{1j} = b_{10} + u_{1j}$ <2. úroveň>

Alternativně (dosazením sloučeno)

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

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

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

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

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

$b_{1j} = b_{10} + u_{1j}$ <2. úroveň>

Alternativně (dosazením sloučeno)

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

Random intercept and slope model $Y_{ij} = b_0 + b_1 X_{ij} + (e_{ij} + u_{0j} + u_{1j})$

Efekty (fixed effects)

b_{00} průměrný průsečík napříč skupinami

b_{10} průměrný efekt pro všechny skupiny

Struktura reziduí (kovarianční parametry)

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

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

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

Model má 5 odhadovaných parametrů.

Šestý 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

$$Test_{iŠ} = b_{00} + b_{10}Gender_{iŠ} + (e_{ij} + u_{0Š} + u_{1Š})$$

Příklad – Skotské zkoušky

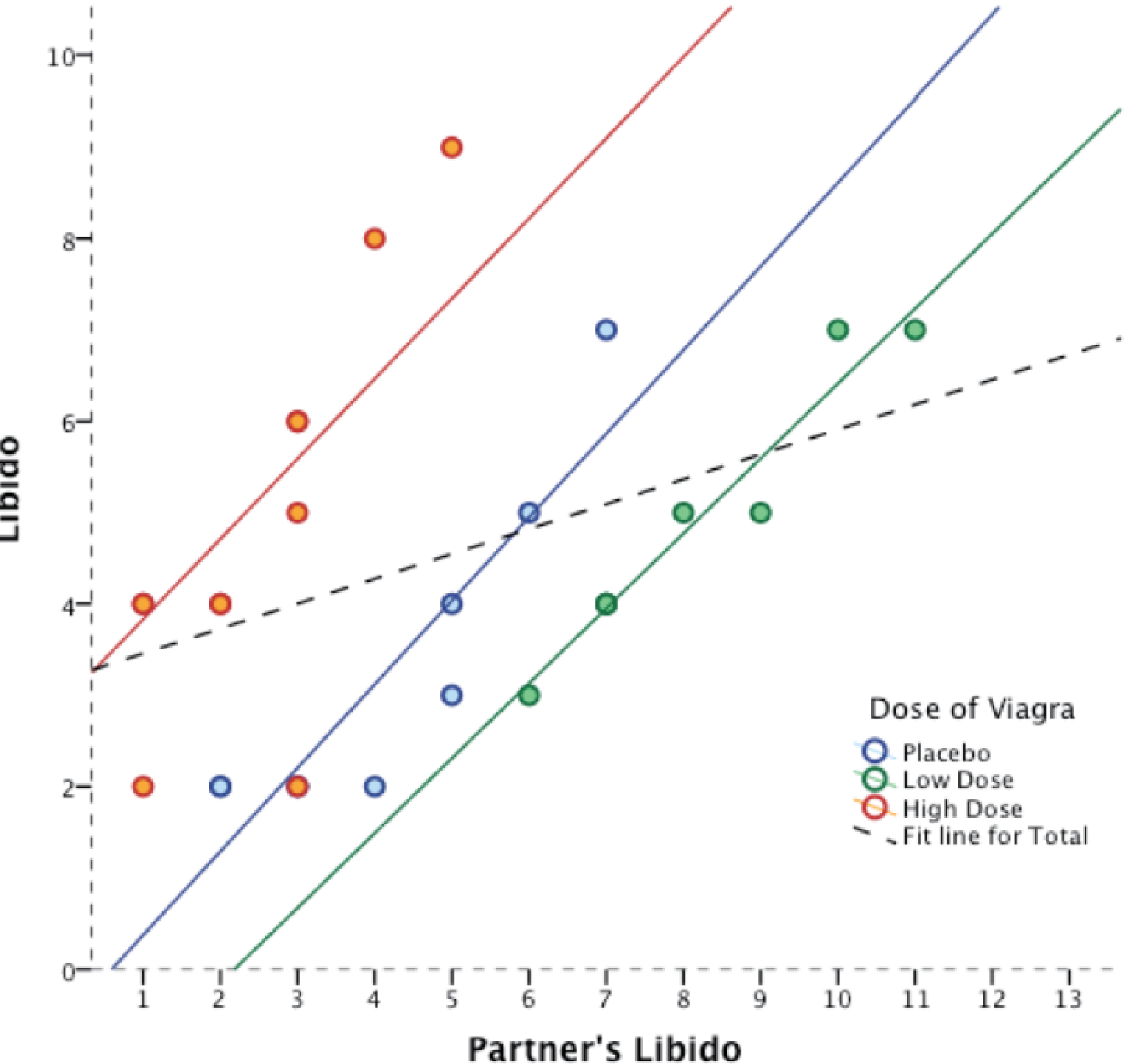
- $Test_i = (79 \pm 12) - (4{,}0 \pm 2{,}7) Gender_i \pm 17{,}8$

Jsou-li kluci 0 a holky 1, pak…

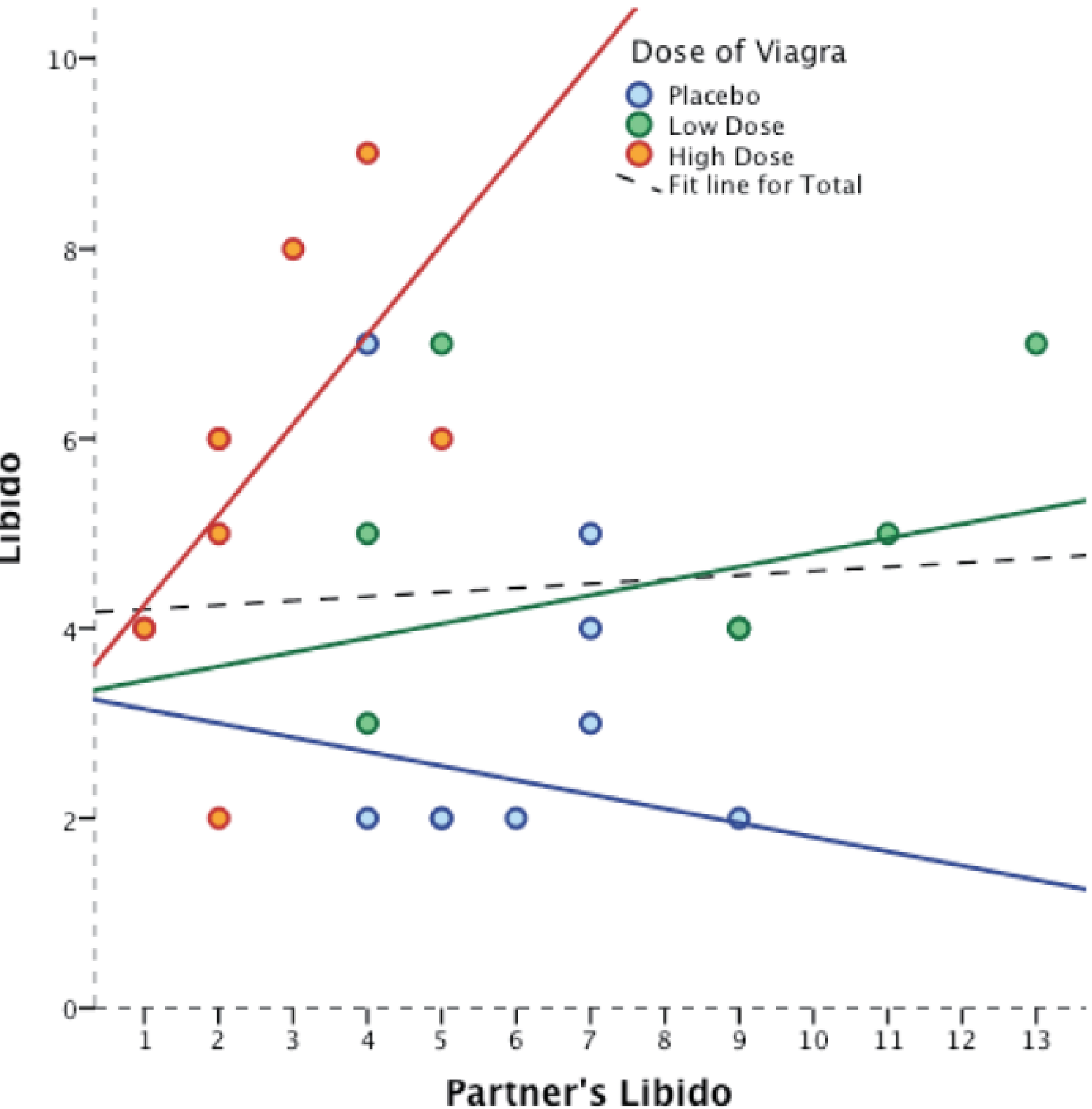
- Výkon průměrného kluka v průměrné škole je 79, přičemž školy se liší tak, že výkony průměrných kluků mají SD=12.
- Průměrná holka má v průměrné škole o 4,0 bodu míň.
- I když napříč školami mají rozdíly mezi průměrnou holkou a průměrným klukem SD=2,7, rozptyl efektů není signifikantně odlišný od 0.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

Shrnutí

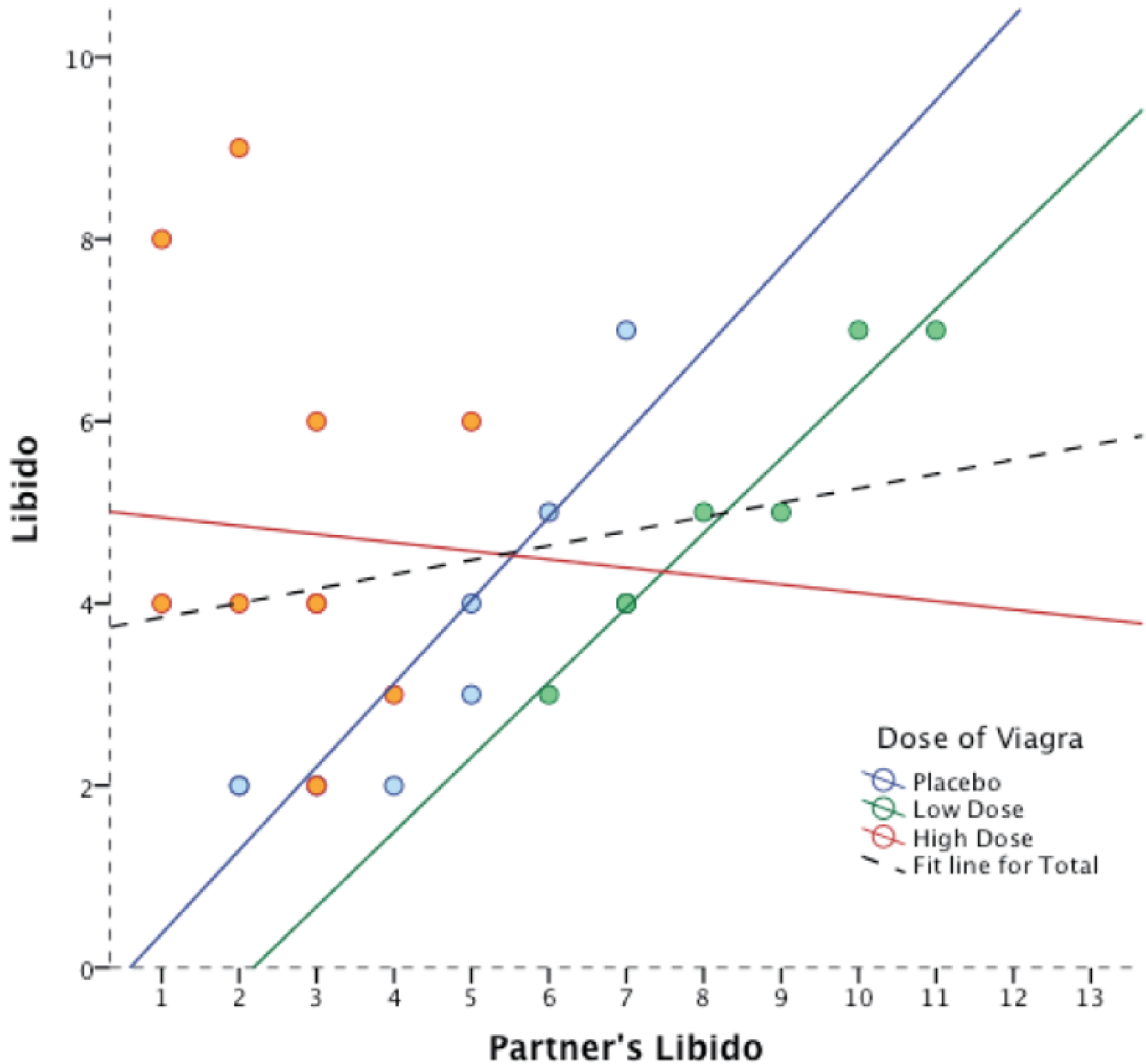
- Multilevel modely nám umožňují modelovat to, že některé parametry regresního modelu se mohou pro různé skupiny lišit.
- Od moderace se to liší tím, že různost parametrů má podobu normálního rozložení. Nezajímáme se o hodnoty pro jednotlivé skupiny – ze vzorku skupin usuzujeme na populaci skupin
- S 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 G

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

$b_{0j} = b_{00} + b_{01}G_{ij} + u_{0j}$ <*2. úroveň*>

$b_{1j} = b_{10} + b_{11}G_{ij} + u_{1j}$ <*2. úroveň*>

- *G* 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?

$Test_{ij} = b_{0\check{S}} + b_1 Gender_i + e_{i\check{S}}$ <1. úroveň>

$b_{0\check{S}} = b_{00} + b_{01}\, Nghbr_{\check{S}} + u_{0\check{S}}$ <2. úroveň>

$$Test_{i\check{S}} = b_{00} + b_{01}\, Nghbr_{\check{S}} + b_{10} Gender_{i\check{S}} + (e_{ij} + u_{0\check{S}} + u_{1\check{S}})$$

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*)

Multilevel alternativy R^2

ICC – vnitrotřídní korelační koeficient

- Random means model dělí rozptyl na reziduální rozptyl a rozptyl způsobený rozdílnými průměry skupin
- ICC=rozptyl interceptů/(rozptyl interceptů+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 → $R^2_{\text{within}}=1-(\sigma^2_{e\ (s\ prediktorem)} / \sigma^2_{e(bez\ prediktoru)})$
  - Interpretujeme jako R^2 v běžné regresi
- L2 prediktor by měl snížit rozptyl náhodného efektu → $R^2_{\text{between}}=1-(\sigma^2_{u\ (s\ prediktorem)} / \sigma^2_{u(bez\ prediktoru)})$
  - Interpretujeme: prediktor vysvětlil x% rozptylu průsečíků

Typy kovariančních struktur

- 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

Čtěte opatrně, Andy tu nejistě mlží. (i v 5. vydání)

Předpoklady

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

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)