

PSY252

Statistická analýza dat v psychologii II

**Přednáška 2**

---

{*Mnohonásobná, vícenásobná*} **lineární regrese**

**Multiple linear regression**

---

# **REGRESE, JAK JSME SI JI PŘEDSTAVILI V PSY117**

---

# Dlouhodobá adaptace sluchu

---

Jak dlouho **vydrží** lidé nepříjemný hlasitý zvuk?

Lze využít informaci o tom, zda člověk poslouchá osobní přehrávač na vysokou **hlasitost** [% z maxima přehrávače] **k odhadu** výdrže nepříjemného zvuku?

hlasitost [%]	výdrž [s]
25	5
31	9
55	20
42	13
47	18
53	17
40	15
35	10
28	10

# 1. Popisné statistiky

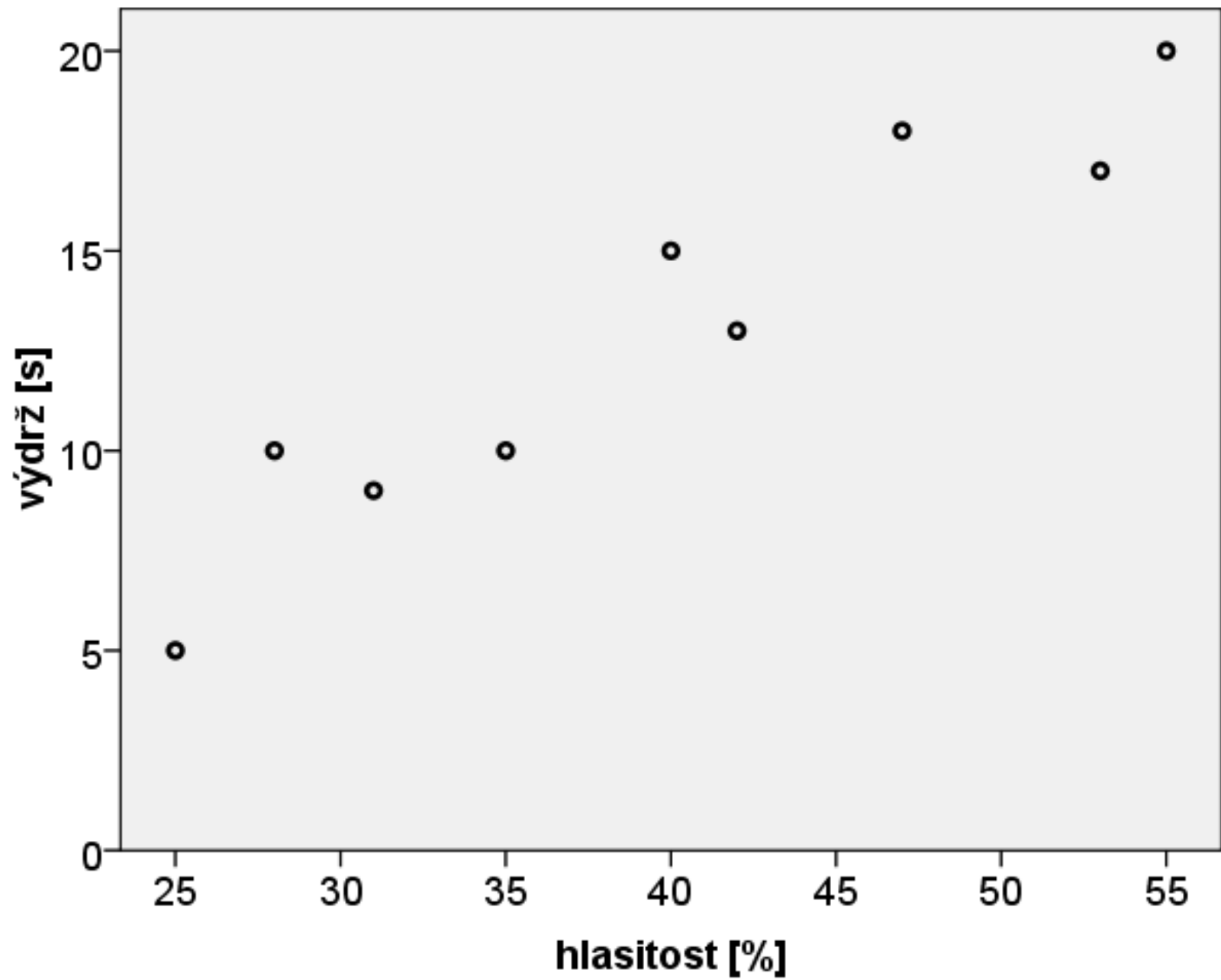
---

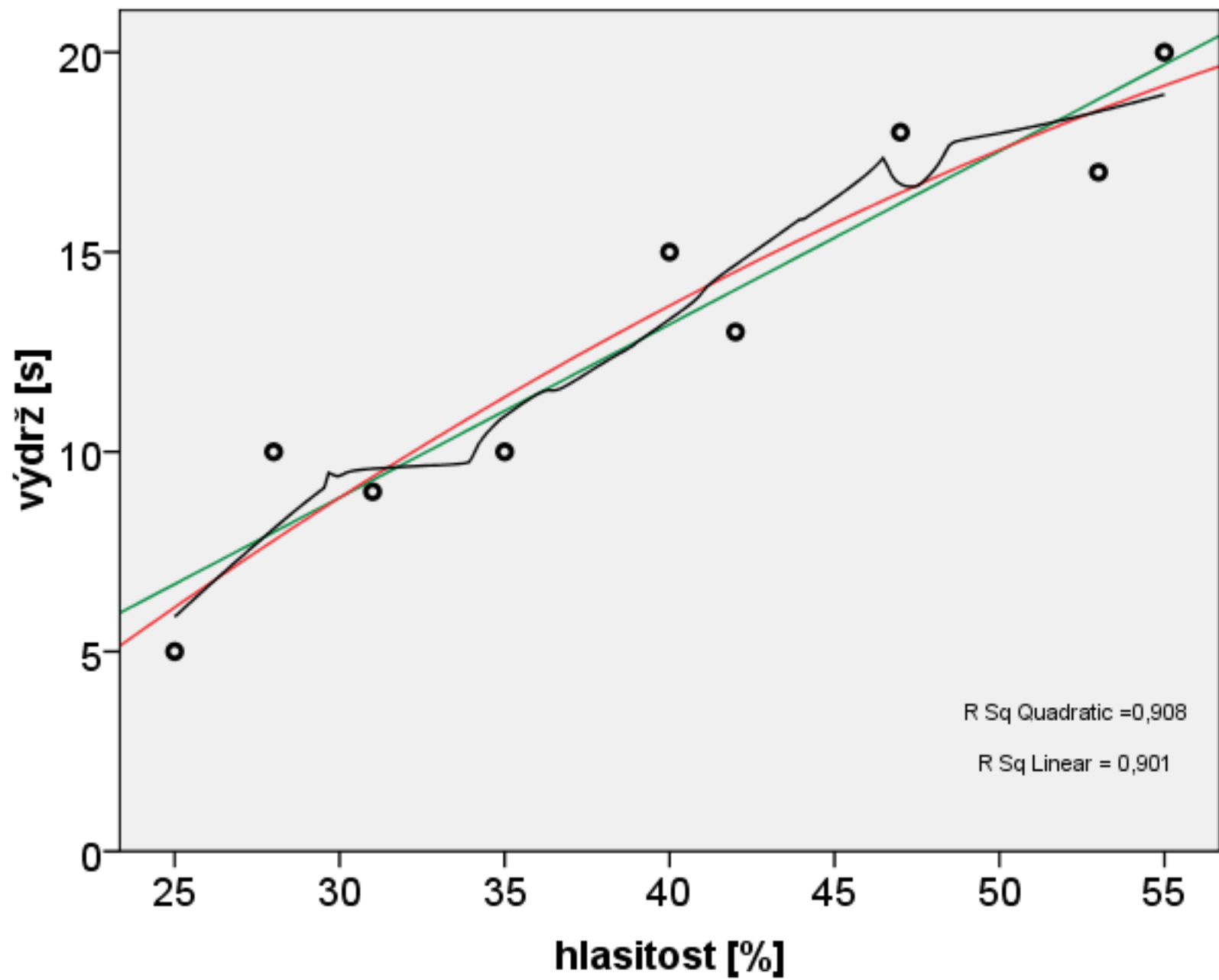
---

## 2. Popis vztahu

---

---





# Lineární regrese I. - **MODEL**

Je-li Pearsonova korelace dobrým popisem vztahu mezi hlasitostí a výdrží, lze vztah popsat, *modelovat* lineární funkcí:

$$V' = b_0 + b_1 H$$

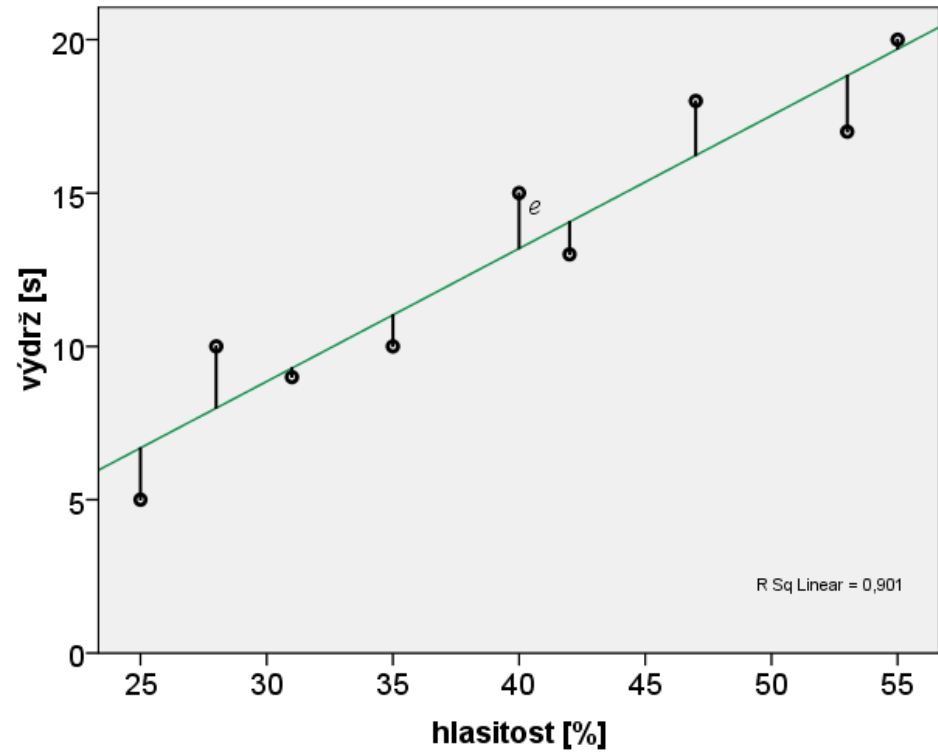
$b_1$  směrnice

$b_0$  průsečík

$$V = V' + e$$

$$V = b_0 + b_1 H + e$$

Pozorování = Model + Chyba



**Odhad** parametrů  $a$ ,  $b$ ?

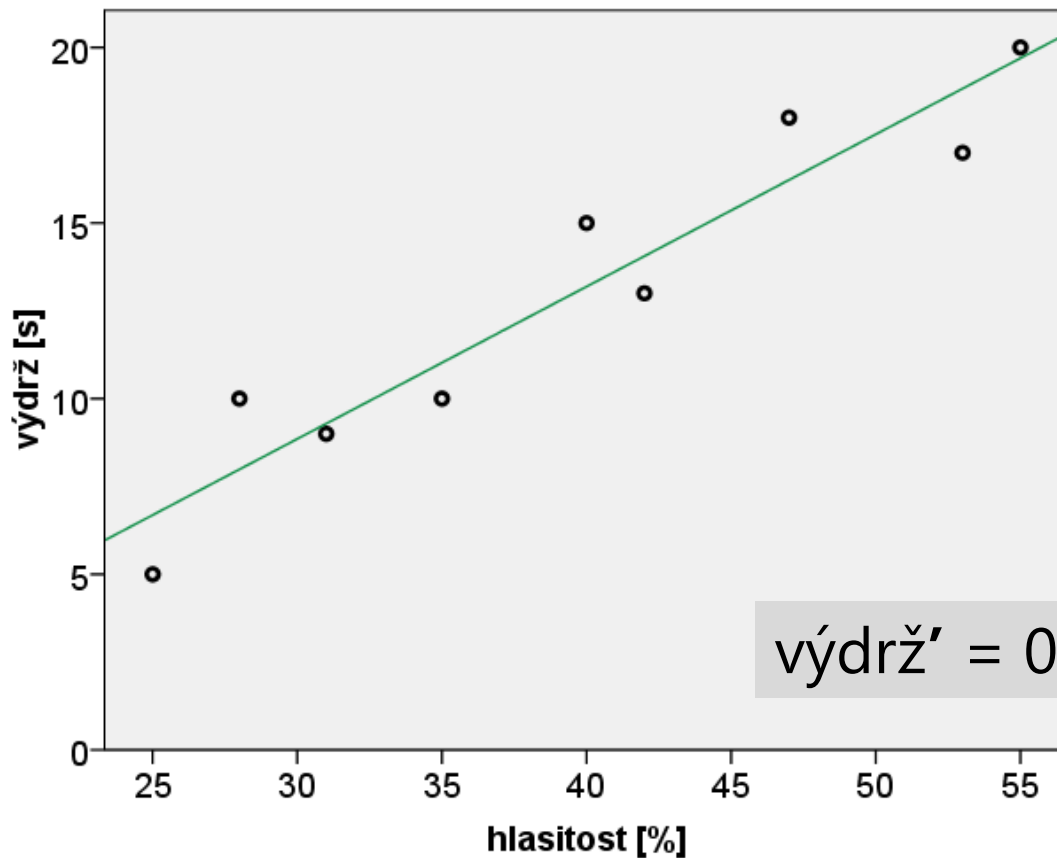
Metodou **nejmenších čtverců** (OLS)



# 3. Odhad parametrů modelu

---

# Lineární regrese II. – příklad



$$m_h = 39,6$$

$$s_h = 10,7$$

$$m_v = 13,0$$

$$s_v = 4,9$$

$$r = 0,95$$

$$\text{výdrž}' = 0,43 \cdot \text{hlasitost} - 4,15$$

# Novinky oproti PSY117

---

- Regr. koeficienty jsou  $\mathbf{b}_0$  (průsečík,  $a$ , (*constant*)) a  $\mathbf{b}_1$  (směrnice,  $b$ )
  - **Beta** – standardizovaný regresní koeficient.
    - O kolik víc násobku SD proměnné  $Y$  predikujeme člověku, který má o 1SD proměnné  $X$  víc. S jedním prediktorem =  $r$ .
  - Testy jednotlivých regresních koeficientů.
    - Testují  $H_0: b_k=0$ . ( $t=b/SE_b$ ,  $t$ -rozložení s  $df=N-k-1$ , )
-

# *Jak dobrý je model?*

## Predikované hodnoty a rezidua

hlasitost [%]	výdrž [s]	výdrž' [s]	reziduum [s]
25	5	6,69	-1,69
31	9	9,29	-0,29
55	20	19,70	0,30
42	13	14,06	-1,06
47	18	16,23	1,77
53	17	18,83	-1,83
40	15	13,19	1,81
35	10	11,02	-1,02
28	10	7,99	2,01

## 4. Rezidua a jejich rozložení

---

- Lineárně regresní model předpokládá normálně rozložená rezidua.
  - Nenaplnění tohoto předpokladu
    - způsobuje nepřesný odhad S.E.
    - Může být způsobeno...
      - nenormalitou závislé
      - nelinearitou vztahu
    - lze mnohdy korigovat transformací
-

# Lineární regrese III. – úspěšnost predikce

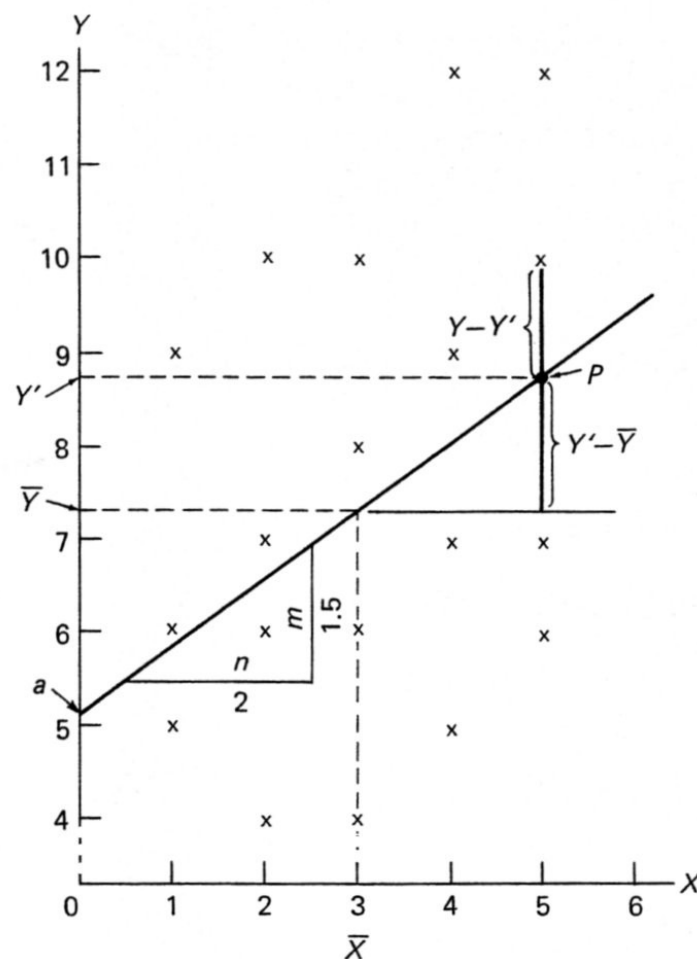
Kritériem kvality modelu jsou nyní nejmenší čtverce – jak malé jsou nejmenší čtverce?

Pozorování = Model + Chyba

Chyba = Pozorování – Model

Suma čtverců chyb (deviance,  $ss_{res}$ ) =  $\sum(V_i - V_i')^2$

Rozptyl chyb ( $s^2_{res}$ ) =  $\sum(V_i - V_i')^2 / (N-1) =$   
= deviance / df

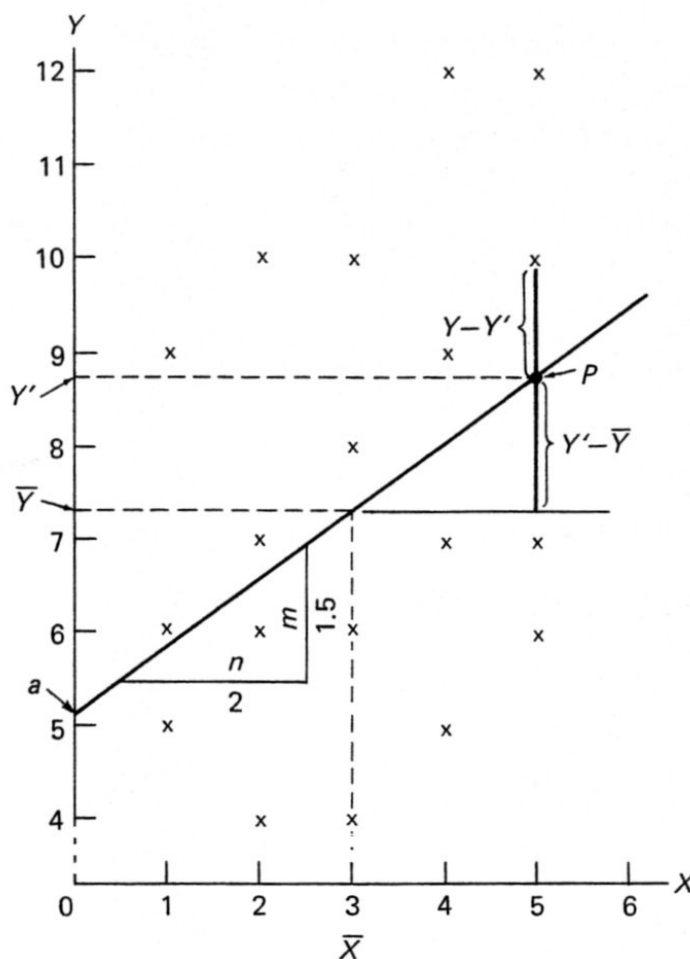


# Lineární regrese III. – úspěšnost predikce

$$R^2 = s_{Y'}^2 / s_Y^2$$

Koeficient determinace ( $R^2$ )

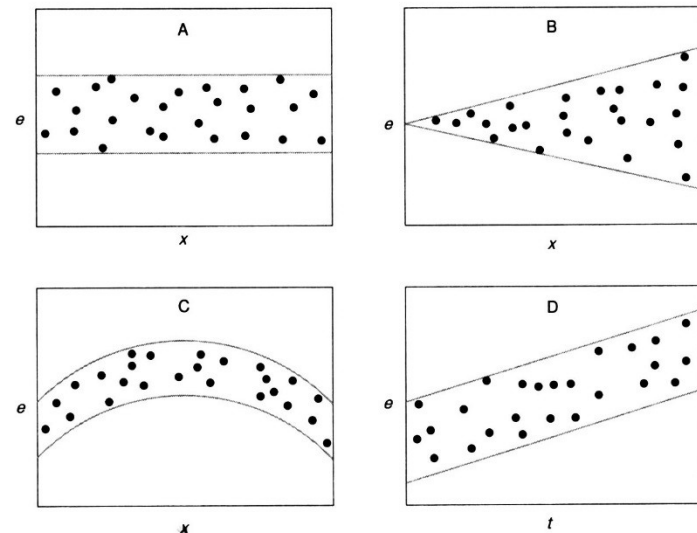
- Podíl rozptylu vysvětleného modelem
- Je ukazatelem kvality, úspěšnosti regrese
- Vyjadřuje shodu modelu s daty



# Lineární regrese IV. – předpoklady, platnost

Předpoklady oprávněnosti použití lineárního modelu

- jako u Pearsonovy korelace
- konceptuální předpoklad: vztah je ve skutečnosti lineární
- rezidua mají normální rozložení s průměrem 0
- homoskedascita
  - =rozptyl reziduí (chyb odhadu) se s rostoucím  $X$  nemění



- Platnost modelu je omezena daty, z nichž byl získán, a teorií.
  - Extrapolace, neoprávněná extrapolace ( $\approx$ jako generalizace nad rámec empirických dat)
  - Pozor na odlehlé hodnoty – jako u všech ostatních momentových statistik



# Konstanta jako model

---

- $M$ : všem predikujeme stejnou hodnotu  $c$
  - $Y' = c$  ,  $Y = c + e$
  - Deviance =  $\sum(Y_i - c)^2$
  - Deviance je nejnižší, když  $c = m_Y$
  - Deviance =  $\sum(Y_i - m_Y)^2$
  - $s^2_{\text{res}} = \sum(Y_i - m_Y)^2 / (N-1)$  ... tedy  $s^2_Y$
  - $s^2_{\text{reg}} = 0$  a tedy i  $R^2 = 0$
  - Nulový model
-

# Novinky oproti PSY117

---

- Adjusted  $R^2$  – jak velké  $R^2$  bychom čekali, kdybychom analýzu dělali na celé populaci (ne vzorku). Overfitting.
- ANOVA – test  $H_0: R^2=0$ .
- Standard error of the estimate -  $s_{\text{res}}$



# Mnohonásobná lineární regrese

---

Více prediktorů, lepší model?

K čemu je?

- Jak moc přispívá proměnná  $X$  k predikci jevu  $Y$ ?
    - Inkrementální validita
  - Liší se muži a ženy v proměnné  $Y$ , i když zohledníme intervenující proměnnou  $Z$ ?
    - Statistická kontrola
  - Je měřítko  $A$  lepším prediktorem než  $B$ ? (lépe pomocí  $r$ )
-

# Mnohonásobná lineární regrese

---

- Počet prediktorů není teoreticky omezen

- $Y = (b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k) + e$

- Problémy plynoucí z většího množství prediktorů

- Výpočetní komplikace
    - Korelace mezi prediktory komplikují interpretaci – (multi)kolinearita
    - Otázka „pořadí“ prediktorů
    - Možnost neintuitivních výsledků – př. suprese
  
  - Více příležitostí k rybaření
    - Méně příležitostí si uvědomit omezenost modelu
    - Množství dat více motivuje k přeskočení detailního se seznamování s daty a prozkoumávání naplnění předpokladů
    - Zapomínání na to, že prioritou je model jako celek
-

# Příklad Long1

---

- záv: deprese
  - pred: selfe, effi, duv\_r, duv\_v
  
  - Celý soubor
-

# MLR: Interpretace regresních koeficientů

---

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

- **$B_i$ ;  $b_i$**  vyjadřuje nárůst  $Y'$  při nárůstu  $X_i$  o jednu jednotku; v jednotkách  $Y$ , při kontrole všech ostatních prediktorů ( $\approx$ semiparciální korelace); jedinečný přínos
    - K porovnání síly prediktoru v různých skupinách, modelech, vzorcích
  - **$\beta_i$ ;  $b_i^*$ ; **BETA**** vyjadřuje nárůst  $Y'$  při nárůstu  $X_i$  o 1; jsou-li  $X_i$  i  $Y$  standardizovány, při kontrole všech ostatních prediktorů ( $\approx$ semiparciální korelace); jedinečný přínos
    - k porovnání prediktorů mezi sebou v rámci jednoho modelu
    - k porovnání různě operacionalizovaného prediktoru v různých modelech
    - ukazatel velikosti účinku
  - **$b_0$**  – obtížně interpretovatelný průsečík ... leda by prediktory byly **centrované**
  - V různých modelech nemusí být vliv prediktoru stejný
-

# MLR: Interpretace regresních koeficientů

---

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

**$B_i$** ;  **$b_i$**  vyjadřuje nárůst  $Y$  při nárůstu  $X_i$  o jednu jednotku; v jednotkách  $Y$ , při kontrole všech ostatních prediktorů ( $\approx$  semiparciální korelace); jedinečný přínos

- Význam  $b$  lze vysoudit i dosazením do regresní rovnice
  
  - Centrování prediktorů usnadňuje přímou interpretaci regresních koeficientů
    - Průsečík pak udává predikci pro člověka, který má průměrnou hodnotu všech prediktorů
-



# Hrátky s prediktory

---

Prediktory lze do modelu vložit všechny najednou, jednotlivě, nebo po skupinkách

Porovnáváme tak vlastně mnoho modelů lišících se zahrnutými prediktory.

- Vše najednou = ENTER
  - Postupně po jednom = FORWARD
  - Vše a postupně ubírat = BACKWARD
  - Po blocích, blockwise = ENTER + další blok
-

# Hierarchická lineární regrese

---

- Bloková, se sadami (sets) prediktorů
  - Prediktory vkládáme po skupinách (popř. jednotlivě) v teoreticky zdůvodněném pořadí
  - Teoreticky zdůvodněné pořadí umožňuje rozdělit rozptyl  $Y$  na smysluplné části (variance partitioning)
    - Změna pořadí prediktorů změní velikost těch částí
  - Zajímá nás schopnost sady prediktorů vylepšit model
    - Srovnání různých oblastí vlivu na zkoumaný jev
    - Zkoumání inkrementální validity
-

# Obvyklá řazení bloků

---

- Dle času, kauzální priority
    - Př. od dispozičním k situačním...
  - Od známých k neznámým vlivům
    - kontrola intervenujících proměnných
    - Minimalizace chyby 1. typu
  - Podle výzkumné relevance
    - Od ústředních po „co kdyby“; maximalizace síly
-

# Obvyklý postup regresní analýzy

---

- Na základě teoretických rozvah stanovíme různé modely, jejichž srovnání je potenciálně zajímavé
  - Nejjednodušší srovnání je u hierarchických modelů, kdy je jeden model plně vnořen do následujícího – to umožňuje testovat inkrement (nárůst)  $R^2$
  - Až v druhé řadě se zabýváme jednotlivými regresními koeficienty v modelu, který je nejúplnější/nejlepší
-



# Diagnostika 1: Outliery a vlivné případy

---

Nemají některé případy příliš velký vliv na výsledky regrese?

- ❑ Outliery – mohou zvyšovat i snižovat  $b$ 
    - **Rezidua** – případy s vysokými r. regrese predikuje nejhůř, standardizovaná, studentizovaná  $\pm 3$
    - **Vlivné případy** – případy, které nejvíc ovlivňují parametry
      - ❑ Co se stane s parametry regrese, když případ odstraníme?
      - ❑ DFBeta – rozdíl mezi parametrem s a bez, standardizované  $> 1$
      - ❑ DFFit – rozdíl mezi predikovanou hodnotou a predikovanou hodnotou bez případu (adjustovanou)
      - ❑ Cookova vzdálenost  $> 1$
      - ❑ Leverage  $> 2(k+1)/n$ , kde  $k$  = počet prediktorů,  $n$  = velikost vzorku
  - ❑ Případy s vysokými rezidui či vlivné případy **NEODSTRAŇUJEME**
    - ❑ ...leđa by šlo o zjevnou chybu v datech či vzorku
    - ❑ ...leđa by nám šlo výhradně o zpřesnění predikce (nikoli o testy hypotéz)
-

# Daignostika 2: Kolinearita

---

- Když 2 prediktory vysvětlují tutéž část variability závislé, jeden z nich je téměř zbytečný
- Komplikuje porovnávání síly preditorů
- Snižuje stabilitu odhadu parametrů
- V extrému (když lze jeden prediktor přesně vypočítat z ostatních) regresi úplně znemožňuje
  
- Korelace nad 0,9
- **Tolerance (=  $1/VIF$ ) cca pod 0,1**
- (VIF (=  $1/tolerance$ ) cca nad 10)

I při korelacích kolem 0,5 komplikuje interpretaci!!

---

# Diagnostika 3: Předpoklady regrese

---

- ❑ Závislá alespoň intervalová, prediktory intervalové i kategorické
  - ❑ Nenulový rozptyl prediktorů
  - ❑ Absence vysoké kolinearity (žádné  $r > 0,9$ , tolerance  $< 0,1$ )
  - ❑ Neexistence intervenující proměnné, která by korelovala se závislou i prediktory
  - ❑ Homoskedascita (scatterplot ZRESID x ZPRED, parciální scatterplot)
  - ❑ Nezávislost reziduí (Durbin-Watson = 2)
  - ❑ Normálně rozložená rezidua (histogram, P-P)
  - ❑ Nezávislost jednotlivých případů
  - ❑ Linearita vztahu
-

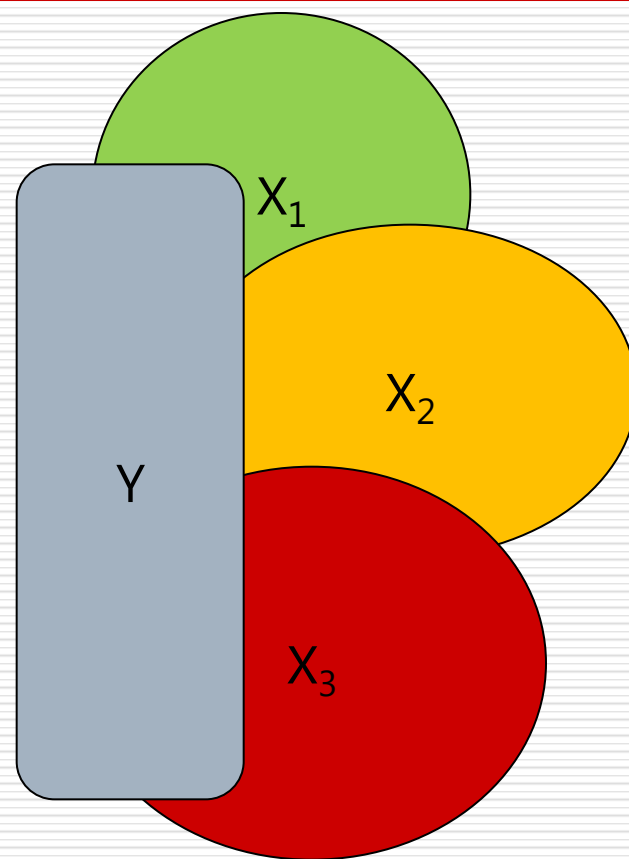




# MLR: Shoda modelu s daty: $R^2$

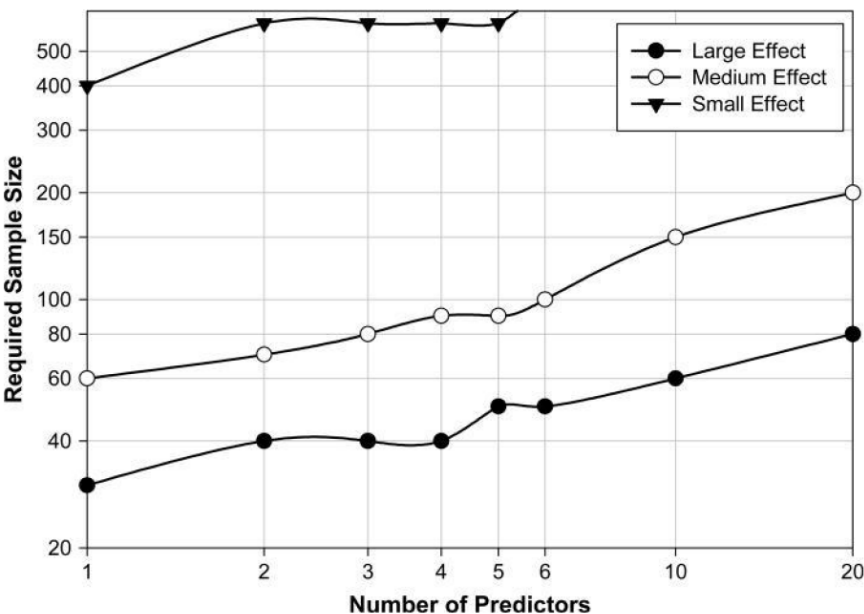
---

- Část rozptylu  $Y$  vysvětleného dohromady všemi prediktory
- Predikční síla sady prediktorů
- Ukazatel velikosti účinku
- $R$ : Mnohonásobná (mutiple) korelace
- Vždy nadhodnocuje >> při replikaci vychází nižší  $R^2$ 
  - shrinkage correction – Adjusted (upravené)  $R^2$ 
    - Wherry (SPSS, Statistica) –kdybychom model dělali z cenzových dat
  - cross-validation
    - Stein (Field) – očekávané  $R^2$  při replikaci
    - split-sample analýza



# Síla testu a velikost vzorku v MLR

Přibývá nový faktor síly testu: **množství prediktorů**



**TABLE 5** Minimum  $R^2$  That Can Be Found Statistically Significant with a Power of .80 for Varying Numbers of Independent Variables and Sample Sizes

Sample Size	Significance Level ( $\alpha$ ) = .01				Significance Level ( $\alpha$ ) = .05			
	No. of Independent Variables				No. of Independent Variables			
	2	5	10	20	2	5	10	20
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1,000	1	2	2	3	1	1	2	2

*Note: Values represent percentage of variance explained.  
NA = not applicable.*

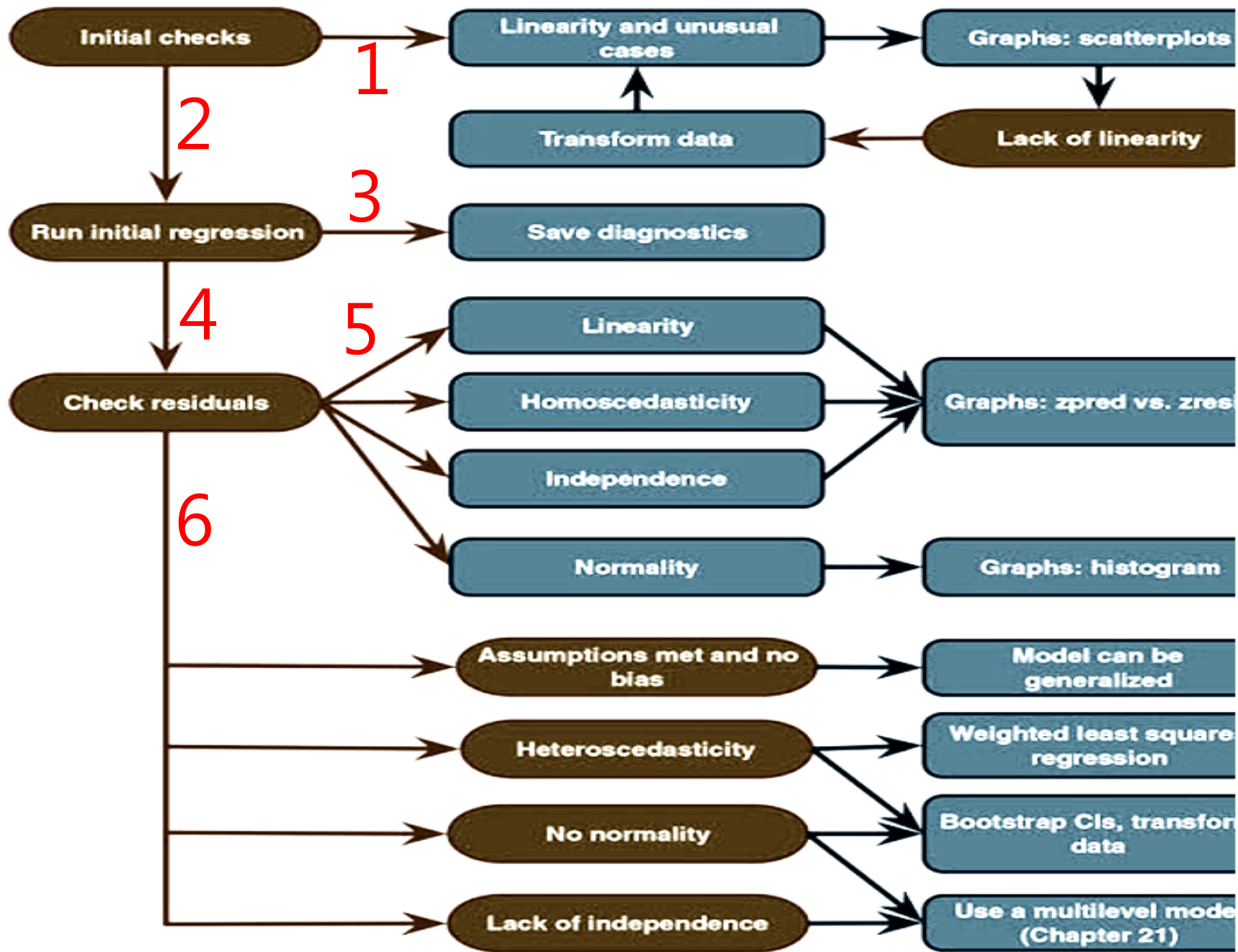
# Reportování MLR

---

## Základ:

- Popisné statistiky  $Y$  a  $X_i$  obvykle s korelační maticí všech
  - Ujištění o naplnění předpokladů
  - Popis shody modelu s daty –  $R^2$ ,  $p$  (někdy i s  $F$ -testem)
  - Přehled regresních koeficientů,  $b$ ,  $\beta$  s jejich  $SE$ , popř. s intervaly spolehlivosti, nebo  $p$
  - Limity, např. možný dopad nedokonalého naplnění předpokladů, vlivných případů apod.
-

- 
- záv: deprese
  - pred: selfe, effi3, duv\_r, duv\_v, pohlavi a mat99
  - Split podle kohorty
-



# Úkol

---

Field Smart Alex's task Ch. 9 Task 7 s. 435

Data:

[http://study.sagepub.com/sites/default/files/spss\\_files\\_0.zip](http://study.sagepub.com/sites/default/files/spss_files_0.zip)

Kontext:

<https://onlinelibrary.wiley.com/doi/full/10.1111/j.1469-7610.2006.01655.x>

Z analýz sepište zprávu v souladu s konvencemi.

Odevzdejte do pondělí – do 14 hodin.

---