

PSY252

Statistická analýza dat v psychologii II

Seminář 3

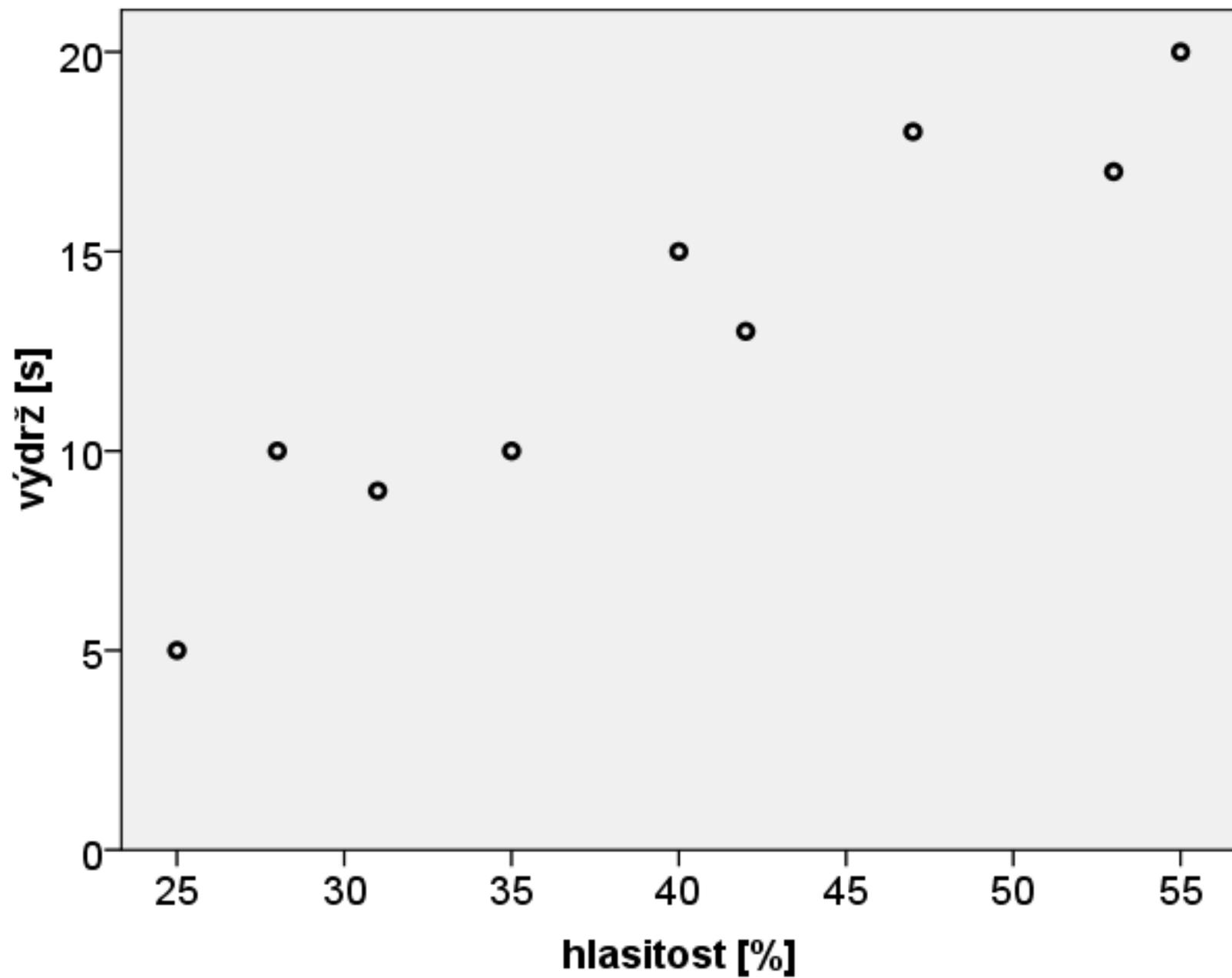
{*Mnohonásobná, vícenásobná*} **lineární regrese**

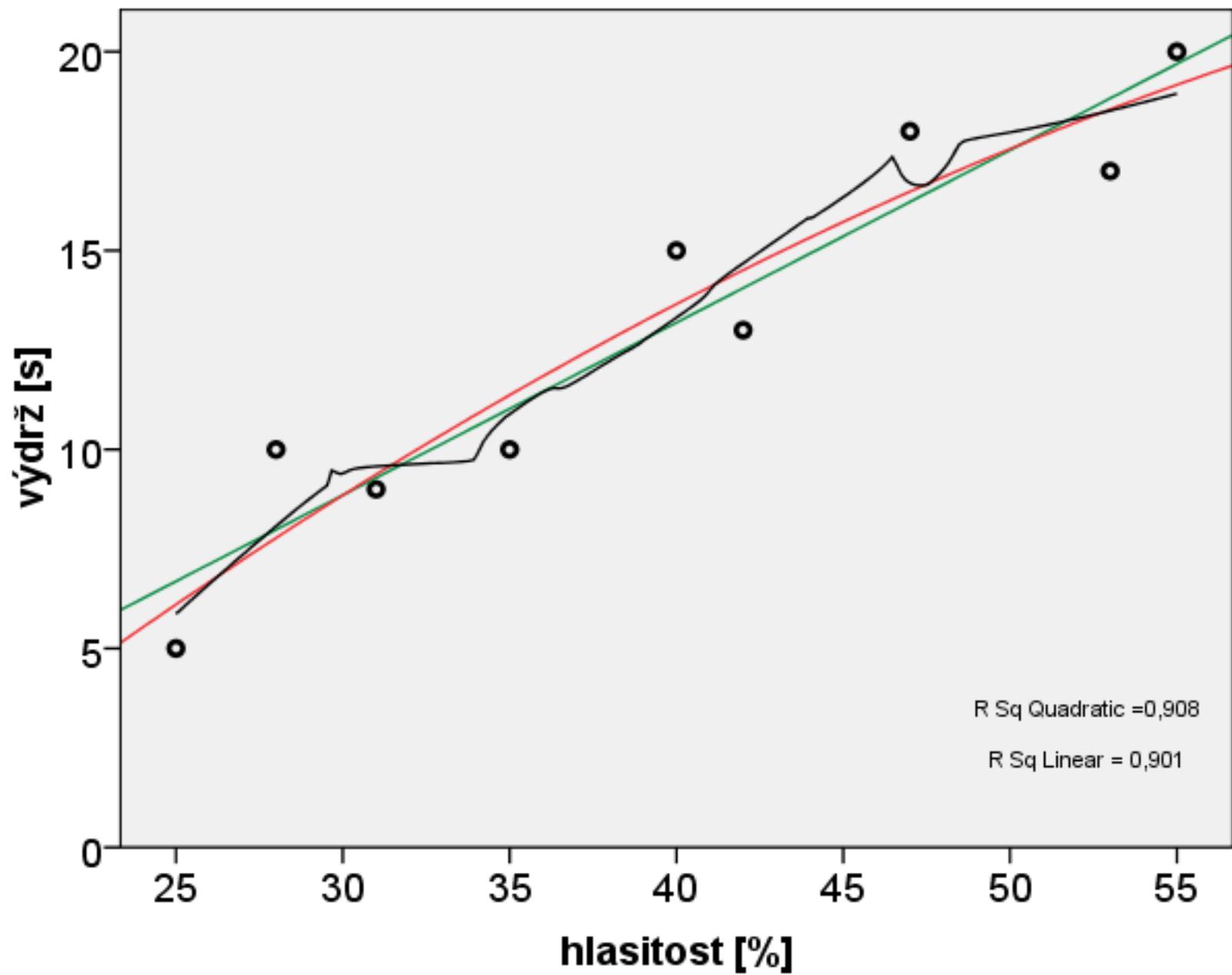
Multiple linear regression

Dhodobá adaptace sluchu

Lidé, kteří poslouchají osobní přehrávač na vysokou **hlasitost** [% z maxima přehrávače], **vydrží** nepříjemný hlasitý zvuk déle?

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





Lineární regrese I. - MODEL

Je-li Pearsonova korelace dobrým popisem vztahu mezi dvěma proměnnými, lze popsat vztah mezi nimi lineární funkcí

$$Y' = a + bX$$

b – směrnice

a – průsečík

$$Y = Y' + e$$

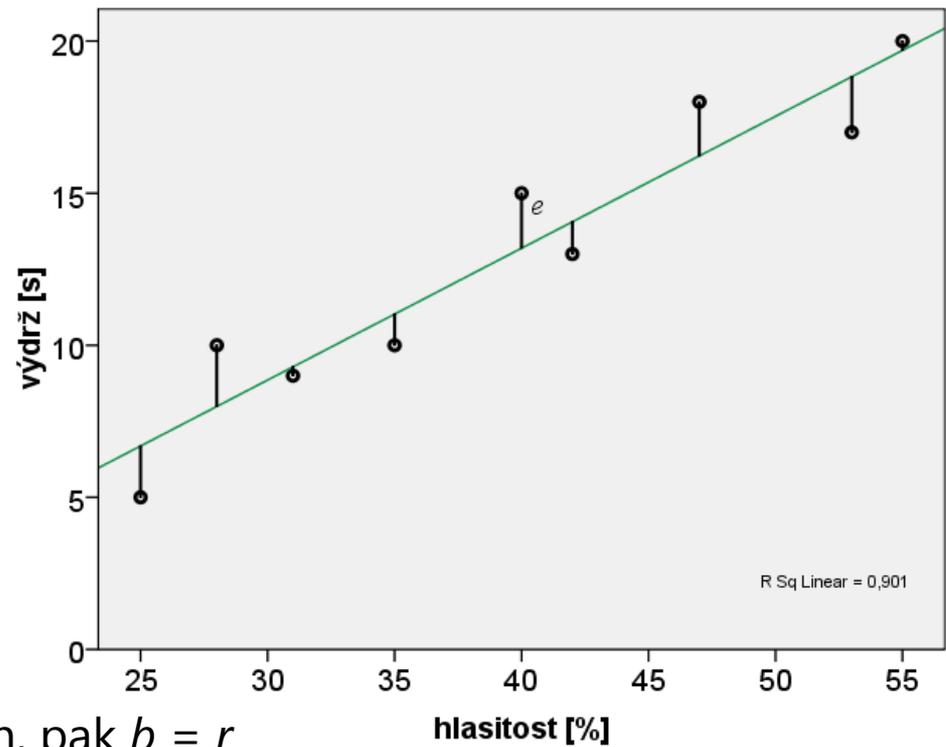
$$Y = a + bX + e$$

Odhad metodou
nejmenších čtverců

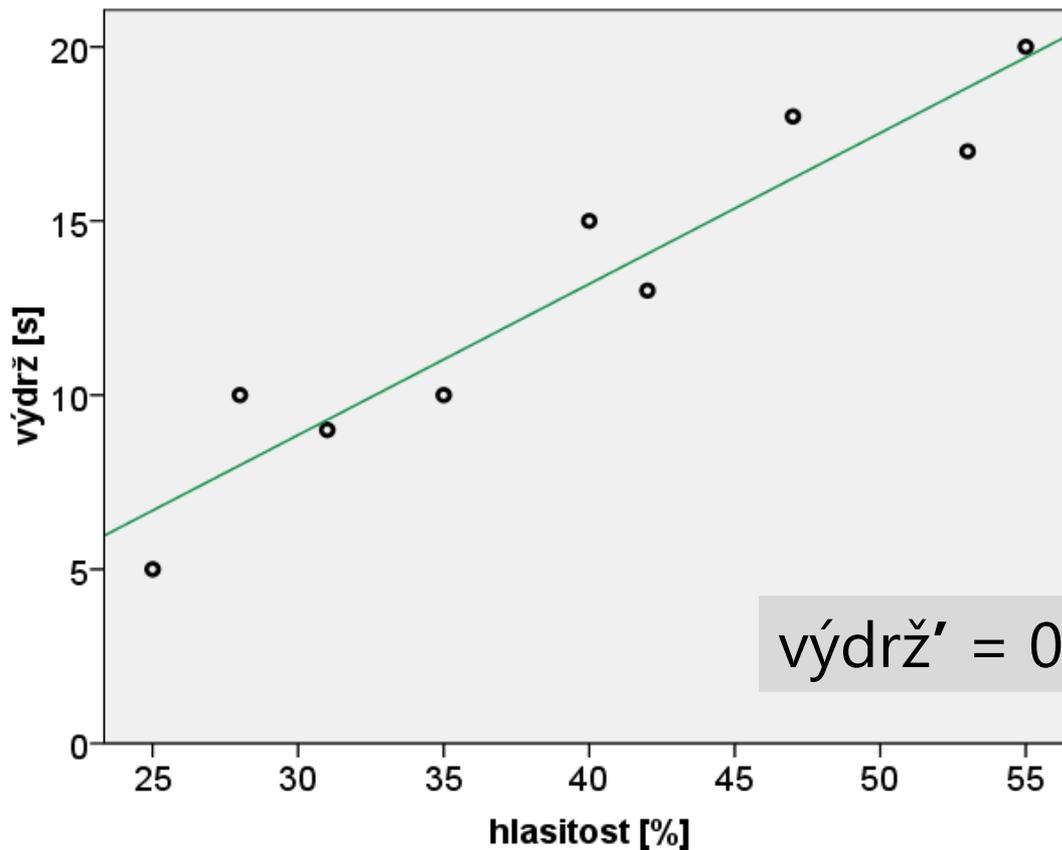
$$b = r_{xy}(s_y/s_x)$$

$$a = m_y - bm_x$$

Jsou-li X a Y vyjádřeny v z-skórech, pak $b = r_{xy}$



Lineární regrese II. – příklad



$$m_h = 39,6$$

$$s_h = 10,7$$

$$m_v = 13,0$$

$$s_v = 4,9$$

$$r = 0,95$$

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

Predikované hodnoty a rezidua

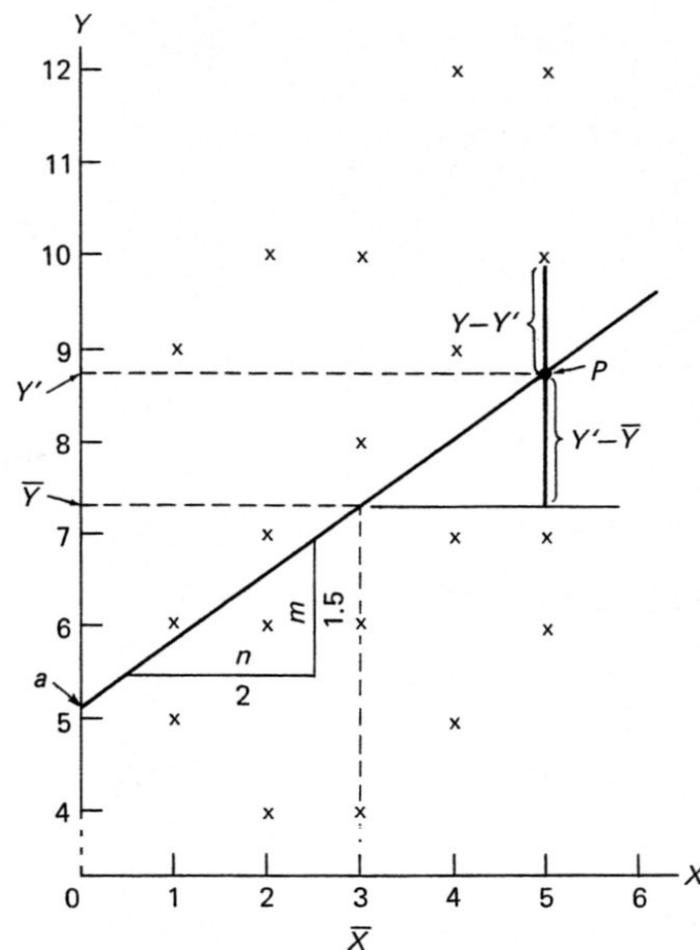
hlasitost [%]	výdrž [s]	výdrž' [s]	reziduum [s]
25	5	6,69	-1,69
31	9	9,29	-0,29
55	20	19,70	0,30
42	13	14,06	-1,06
47	18	16,23	1,77
53	17	18,83	-1,83
40	15	13,19	1,81
35	10	11,02	-1,02
28	10	7,99	2,01

Lineární regrese III. – úspěšnost predikce

$$s_{reg}^2 = \frac{\sum (m_y - Y')^2}{n-1} \quad s_{res}^2 = \frac{\sum (Y - Y')^2}{n-2}$$

$$s_y^2 = \frac{\sum (Y - m_y)^2}{n-1}$$

- $s_y^2 = s_{reg}^2 + s_{res}^2$ ($SS_y = SS_{res} + SS_{reg}$)
- $R^2 = s_{reg}^2 / s_y^2$
- Koeficient determinace (R^2)
 - Podíl rozptylu vysvětleného modelem
 - Je ukazatelem kvality, úspěšnosti regrese
 - Vyjadřuje shodu modelu s daty
- Pro jednoduchou lin. regr. platí $R^2 = r^2$

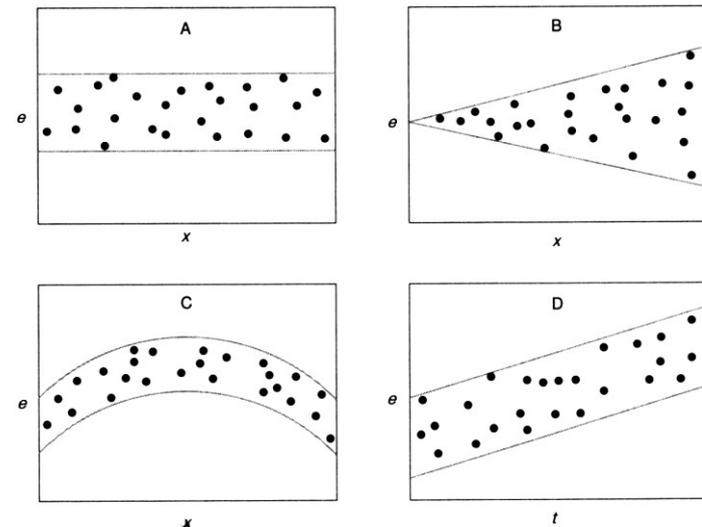


AJ: regression and residual variance (sum of squares), explained variance, **model fit with the data**, coefficient of determination (R square)

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 (≈jako generalizace nad rámec empirických dat)
 - Pozor na odlehlé hodnoty – jako u všech ostatních momentových statistik

Mnohonásobná lineární regrese

- Počet prediktorů není omezen

- $$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_mX_m + e$$

- Problémy plynoucí z většího množství prediktorů

- Výpočetní komplikace
 - Korelace mezi prediktory komplikují interpretaci – (multi)kolinearita
 - Otázka „pořadí“ prediktorů
 - Možnost a problémy porovnávání prediktorů mezi sebou
 - b_i vyjadřuje nárůst Y' při nárůstu X_i o jednu jednotku; v jednotkách Y
 - β_i vyjadřuje nárůst Y' při nárůstu X_i o 1; jsou-li X_i i Y standardizovány
 - K porovnání prediktorů mezi sebou v rámci regrese slouží β_i
 - K porovnání síly prediktoru v různých skupinách slouží b_i

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
-

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 ± 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
-

Diagnostika 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
 - VIF (= $1/\text{tolerance}$) cca nad 10
-

Diagnostika 3: Předpoklady regrese

- ❑ Závislá alespoň intervalová
 - ❑ Prediktory intervalové i kategorické
 - ❑ Nenulový rozptyl prediktorů
 - ❑ Absence vysoké kolinearity (žádné $r > 0.9$)
 - ❑ Neexistence intervenující proměnné, která by korelovala se závislou i prediktory
 - ❑ Homoscedascita (scatterplot ZRESID x ZPRED, parciální scatter)
 - ❑ Nezávislost reziduí (Durbin-Watson = 2)
 - ❑ Normálně rozložená rezidua (histogram, P-P)
 - ❑ Nezávislost jednotlivých případů
 - ❑ Linearita vztahu
-

Síla testu v regresi (Hair, 7th ed.)

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 (α) = .01 No. of Independent Variables				Significance Level (α) = .05 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.*

Zapojení kategorických prediktorů

Dummy coding -> dummy variables

- Pomocí $k-1$ kategorických proměnných
- Indikátorové kódování (indicator coding)
 - Referenční kategorie = 0
- Efektové kódování (effect coding)
 - Referenční kategorie = -1

Člen rodiny	Původní proměnná	Indikátorové kódování		Efektové kódování	
		Matka	Otec	Matka	Otec
Matka	1	1	0	1	0
Otec	2	0	1	0	1
Dítě	3	0	0	-1	-1

Interpretace vah dummy proměnných

- $Y = b_0 + b_{A1}X_{A1} + b_{A2}X_{A2} + \dots + b_mX_m + e$
- Po dosažení do regresní rovnice predikujeme člověku průměr jeho skupiny (pokud nejsou žádné další prediktory).
- Indikátorové kódování
 - b_{Ai} udává rozdíl průměrných hodnot Y mezi indikovanou skupinou a referenční skupinou; sig b_{Ai} znamená sig rozdílu
 - b_{Ai} udává o kolik nám členství ve skupině zvyšuje/snižuje predikovanou hodnotu oproti referenční skupině
 - b_0 udává (při absenci jiných prediktorů) průměr Y v referenční skupině
- Efektové kódování
 - b_{Ai} udává rozdíl průměrných hodnot Y mezi indikovanou skupinou a celkovým průměrem
 - b_0 udává (při absenci jiných prediktorů) celkový průměr

□ záv: deprese

□ pred: selfe, effi3, duv_r, duv_v
