PSY252
Statistická analýza dat v psychologii II
Přednáška 2

{*Mnohonásobná, vícenásobná*} lineární regrese

Multiple linear regression

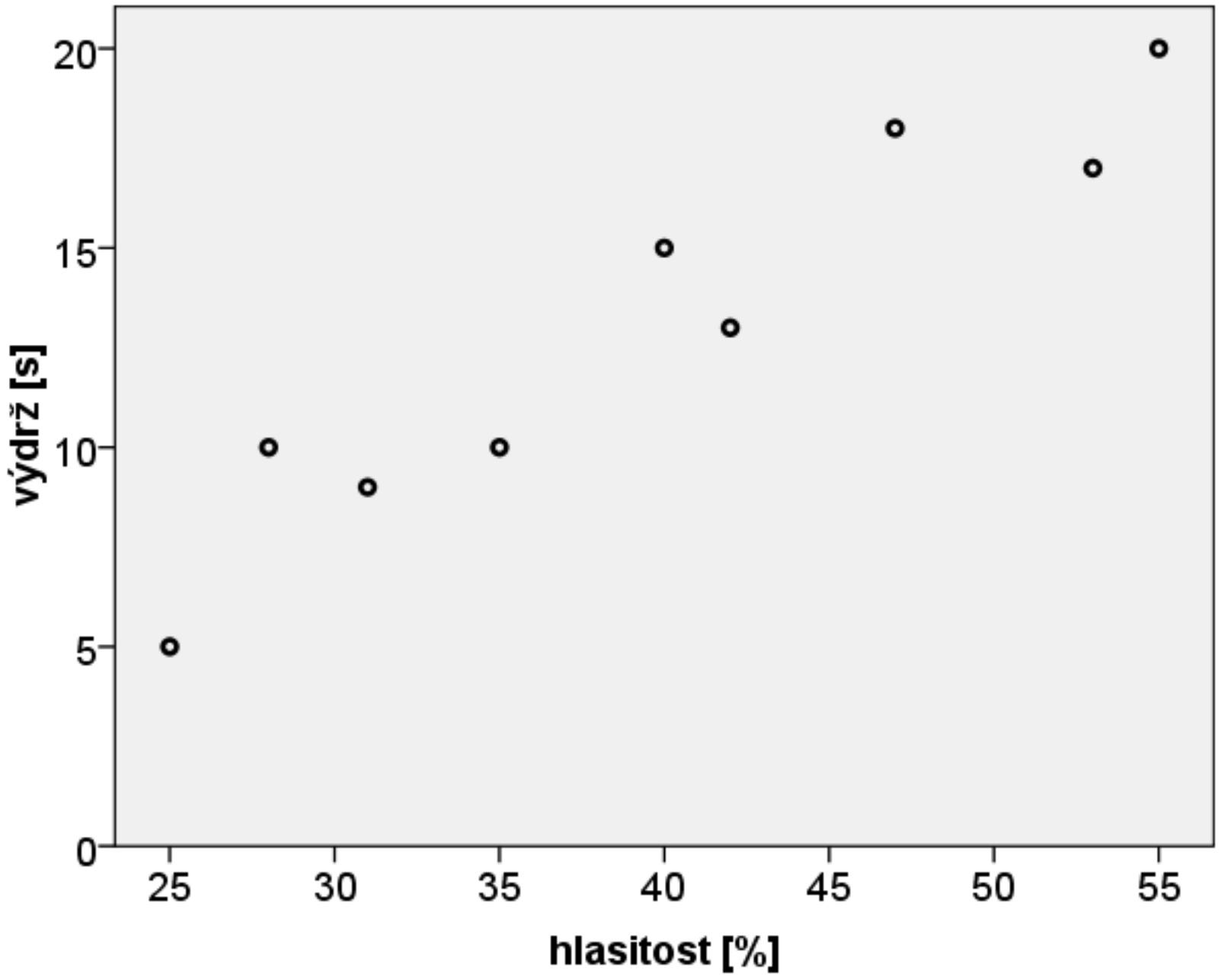
REGRESE, JAK JSME SI JI PŘEDSTAVILI V PSY117

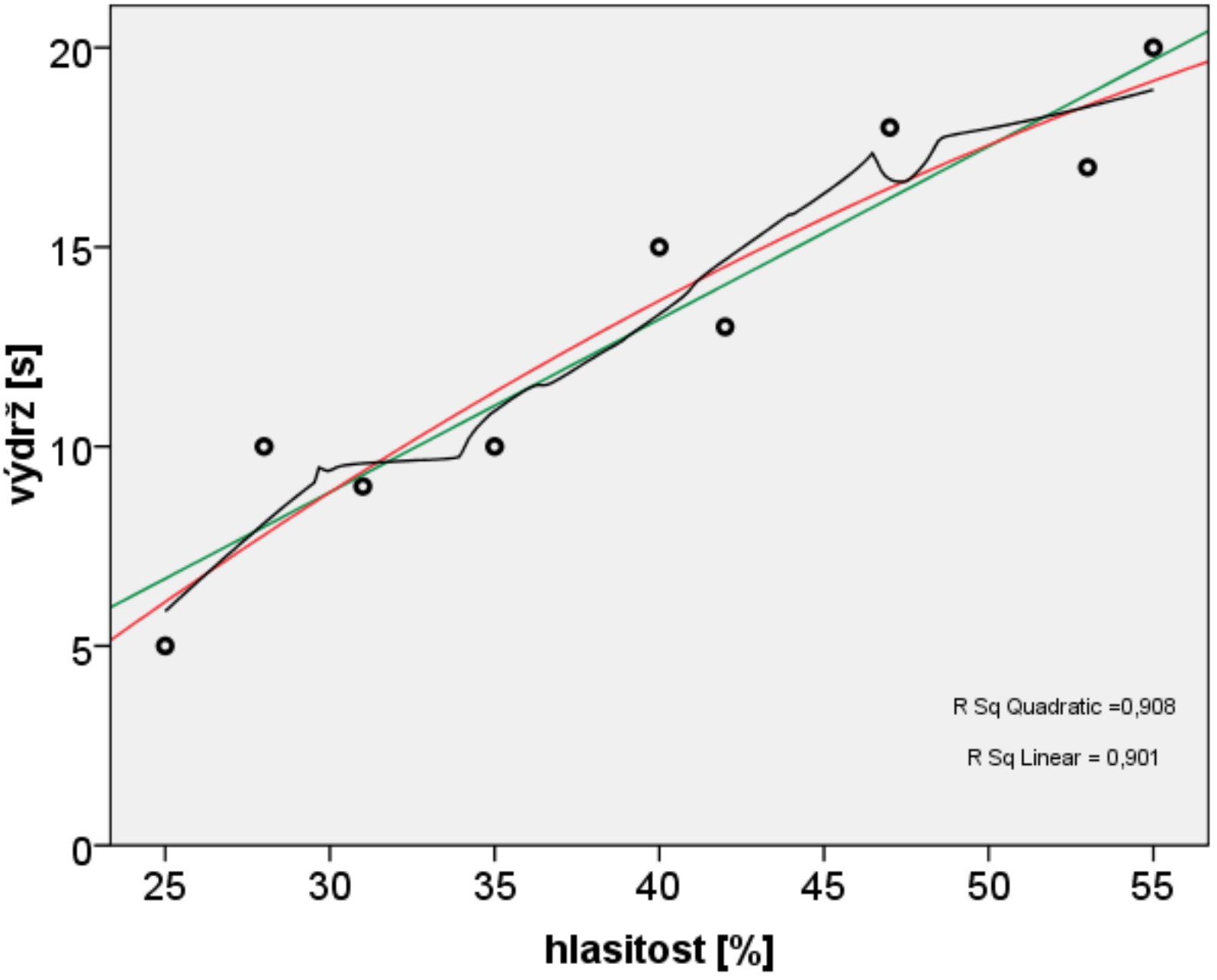
Dlouhodobá adaptace sluchu

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


20
15
10
5
0
výdrž [s]
25
30
35
40
45
50
55
hlasitost [%]


20
15
10
5
0
výdrž [s]
25
30
35
40
45
50
55
hlasitost [%]
R Sq Quadratic =0,908
R Sq Linear = 0,901

Lineární regrese I. - **MODEL**

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

$V' = b_0 + b_1 H$

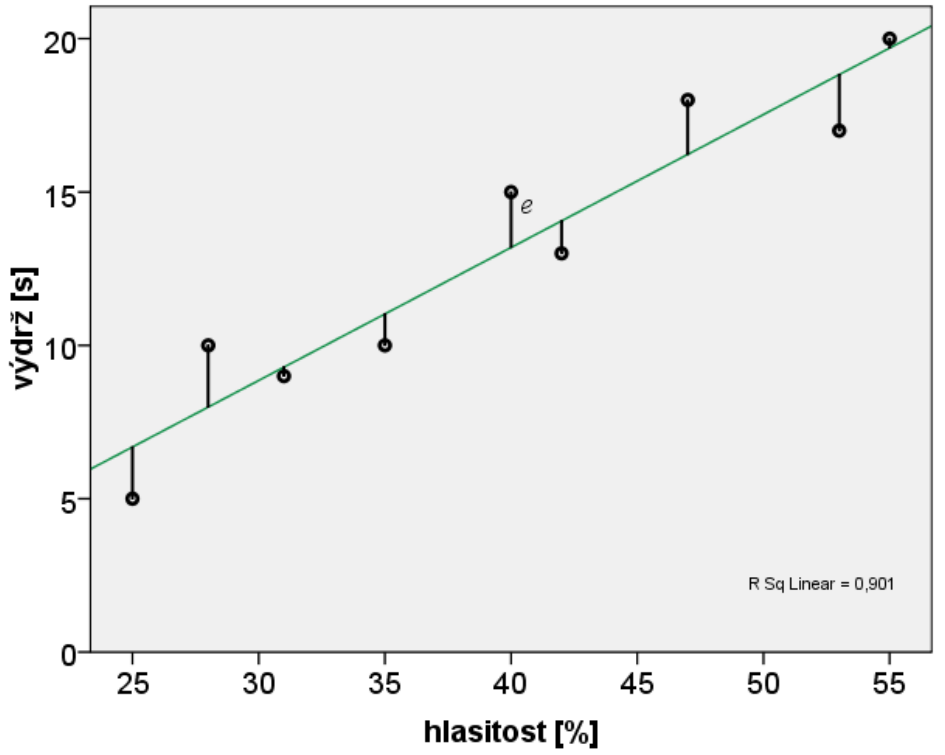
$\boldsymbol{b_1}$ směrnice

$\boldsymbol{b_0}$ průsečík

$V = V' + e$

$V = b_0 + b_1 H + e$

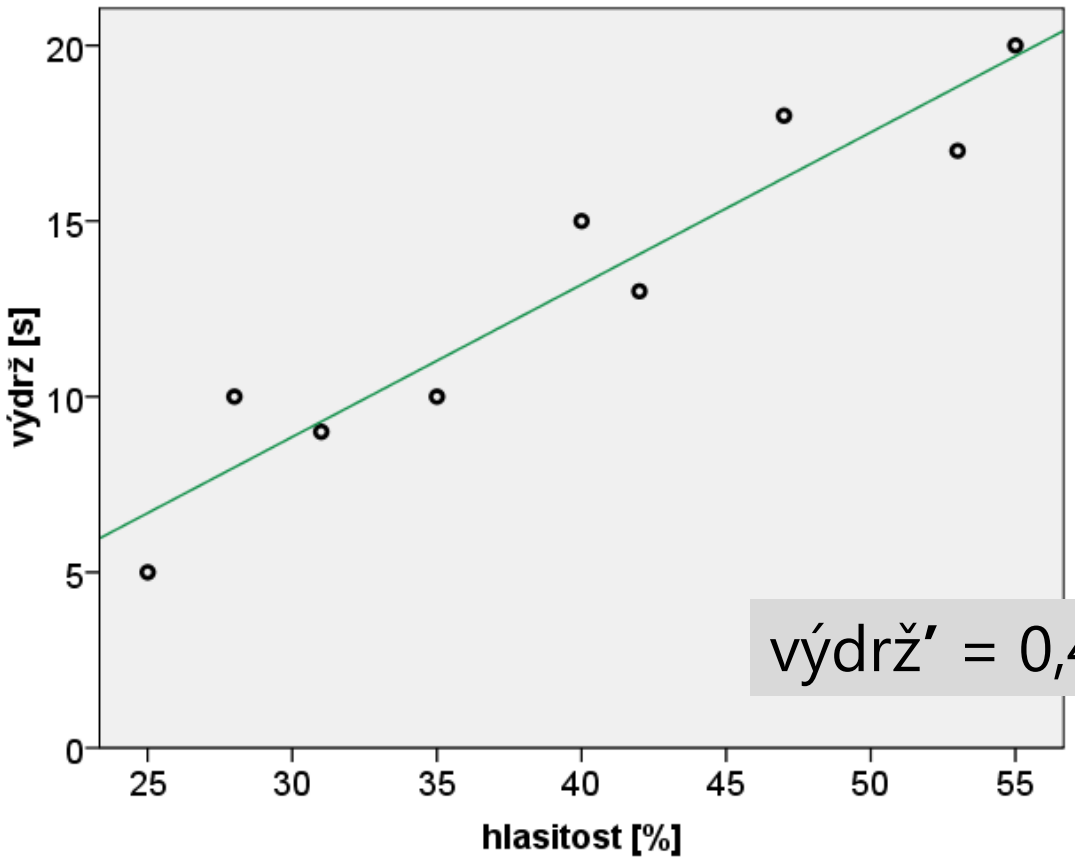
Pozorování=Model+Chyba



Odhad parametrů *a*, *b*?

Metodou **nejmenších čtverců** (OLS)

Lineární regrese II. – příklad



m_h=39,6

s_h = 10,7

m_v=13,0

s_v = 4,9

r = 0,95

výdrž' = 0,43.hlasitost – 4,15

Novinky oproti PSY117

- Regr. koeficienty jsou $\boldsymbol{b_0}$ (průsečík, *a, (constant)*) a $\boldsymbol{b_1}$(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

Lineární regrese III. – úspěšnost predikce

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

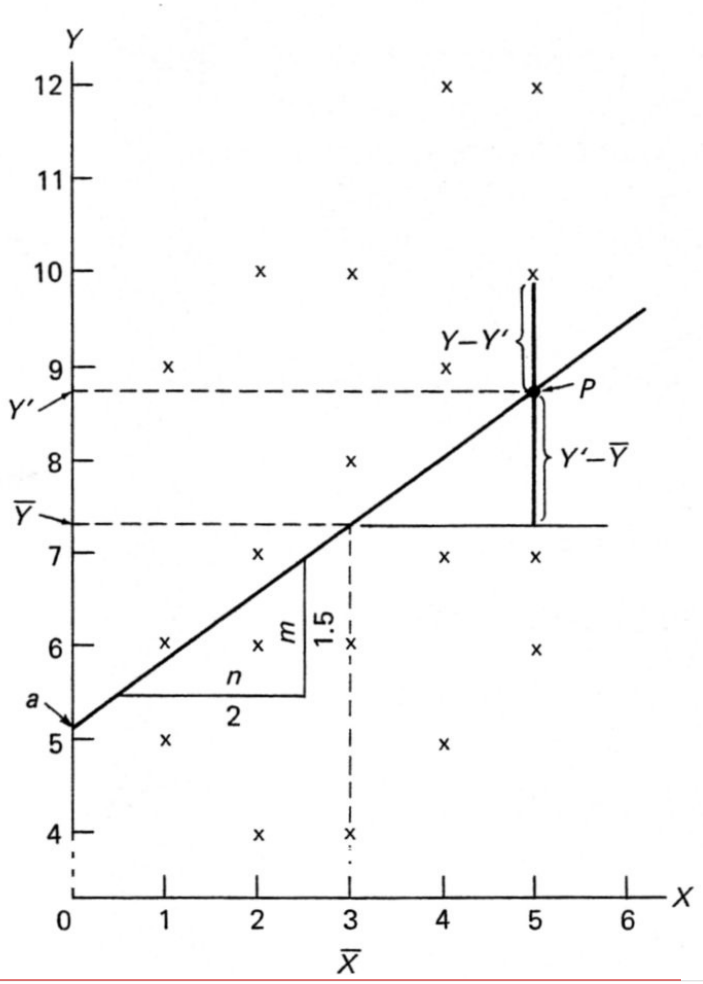
Pozorování = Model + Chyba

Chyba = Pozorování – Model

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

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

= deviance / df

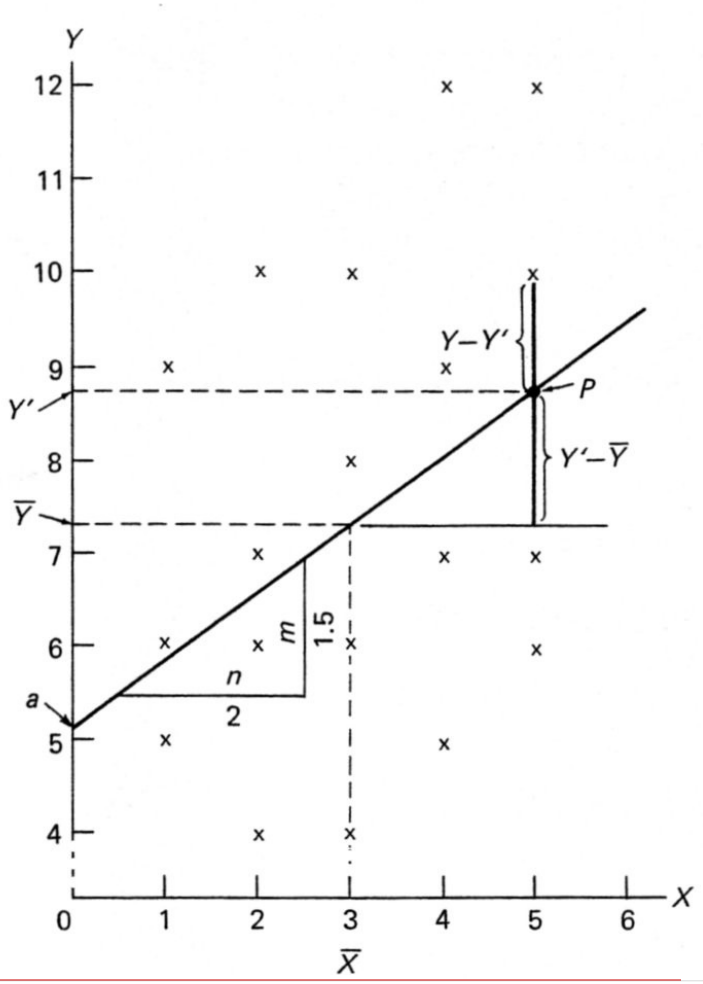


Lineární regrese III. – úspěšnost predikce

$R^2 = s_{V'}^2 / s_V^2$

Koeficient determinace (R^2)

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



Konstanta jako model

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

Novinky oproti PSY117

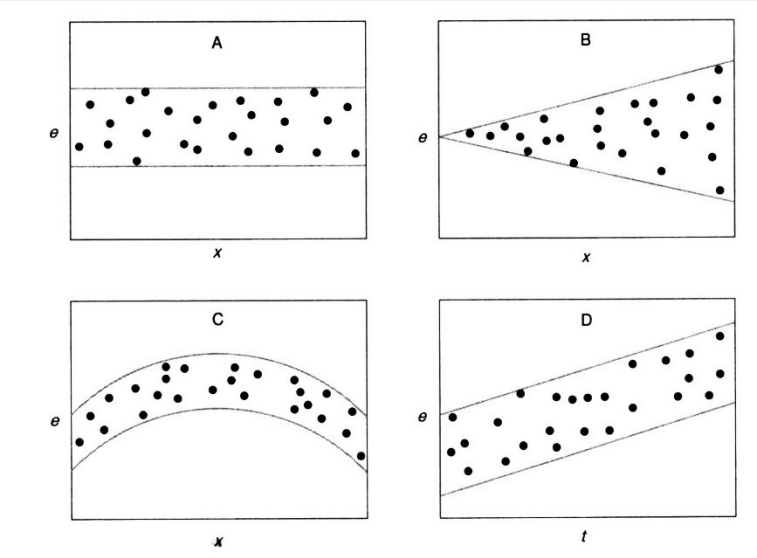
- Adjusted R^2 – jak velké R^2 bychom čekali, kdybychom analýzu dělali na celé populaci (ne vzorku). Overfitting.
- ANOVA – test H_0: $R^2=0$.
- Standard error of the estimate - s_{res}

https://en.wikipedia.org/wiki/Overfitting

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

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
  - $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e$

- 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

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

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

MLR: Interpretace regresních koeficientů

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

$\boldsymbol{B_i}$; $\boldsymbol{b_i}$ vyjadřuje nárůst *Y'* při nárůstu X_i o jednu jednotku; v jednotkách *Y*, při kontrole všech ostatních prediktorů (≈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 R2
- 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 ±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**
  - …leda by šlo o zjevnou chybu v datech či vzorku
  - …leda 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)

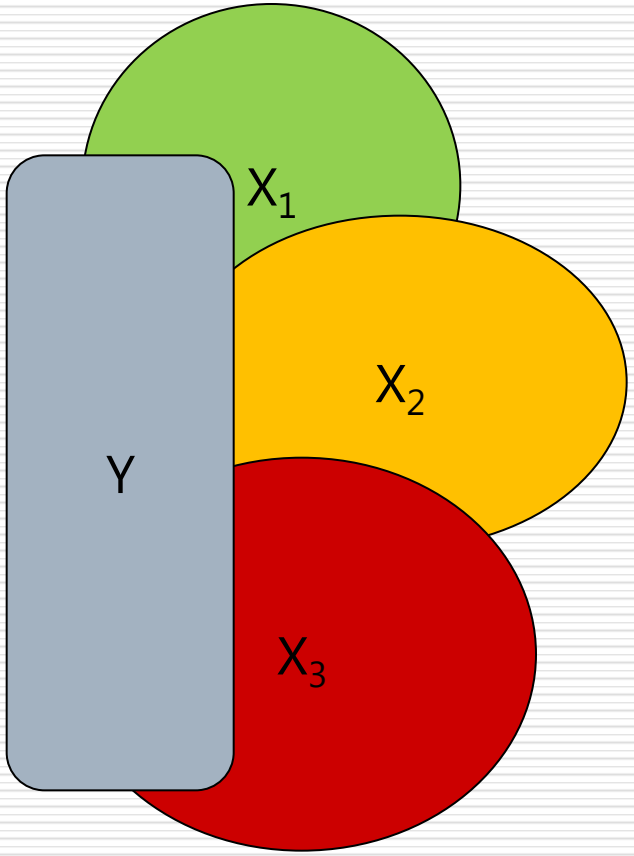
I při korelacích kolem 0,5 komplikuje interpretaci!!

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 R2 při replikaci
    - split-sample analýza



Síla testu a velikost vzorku v MLR

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

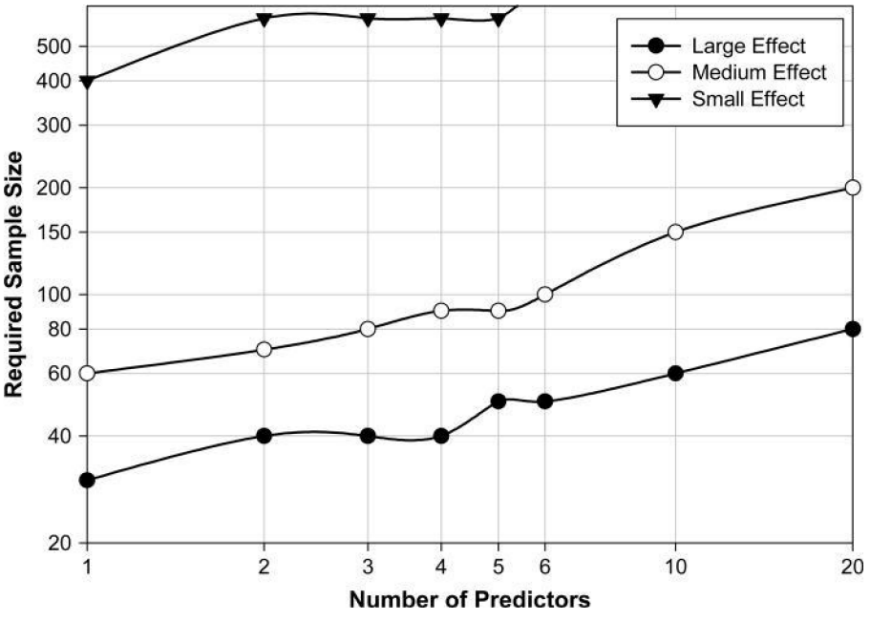


TABLE 5 Minimum R^2 That Can Be Found Statistically Significant with a Power of .80 for Varying Numbers of Independent Variables and Sample Sizes

	Significance Level (α) = .01 *No. of Independent Variables*				*Significance Level (α) = .05* *No. of Independent Variables*			
Sample Size	**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 často s korelační maticí
  - Ujištění o naplnění předpokladů
  - Popis shody modelu s daty – R^2 , p (někdy i s Ftestem)
  - Přehled regresních koeficientů, b, β s jejich SE, popř. s intervaly spolehlivosti, nebo p

- záv: deprese
- pred: selfe, effi3, duv_r, duv_v, pohlavi a mat99
- Split podle kohorty

Úkol

Vytvořte model predikující sebehodnocení dospívajících (položky k01–k12).

Jako prediktory zařaďte ve zdůvodněném pořadí po blocích následující proměnné:

> Deprese (n01 – n20), Vřelost matky (b01 – b22 – liché položky), Zdraví (l01 – l15), Vřelost otce (b01 – b22 – sudé položky), Optimismus (h01 – h08), Důvěrnost s přáteli (d01 – d12 – sudé položky) Proměnné si v datech vytvořte.

Podívejte se na model odděleně u chlapců a dívek. Zkuste se zamyslet nad možnými odlišnostmi a jejich vysvětlením.

Z analýz sepište zprávu v souladu s konvencemi.

Odevzdejte do pondělí – do 14 hodin.