(Vícenásobná) lineární regrese

(Multiple) Linear Regression

Vít Gabrhel

vit.gabrhel@mail.muni.cz

vit.gabrhel@cdv.cz

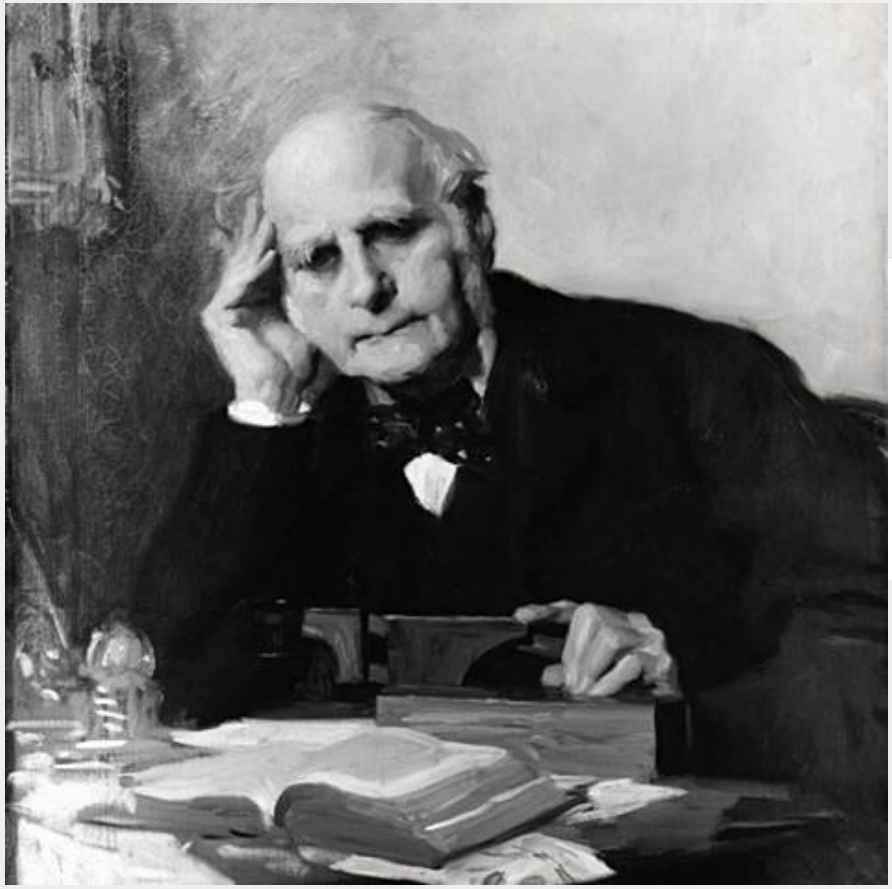
FSS MU,

14. 10. 2015

Harmonogram

1. Historie
2. Teorie
3. Předpoklady použití
4. Diagnostika
5. Dummy coding
6. Vkládání prediktorů
7. Reportování výsledků

1. O původu lineární regrese I.



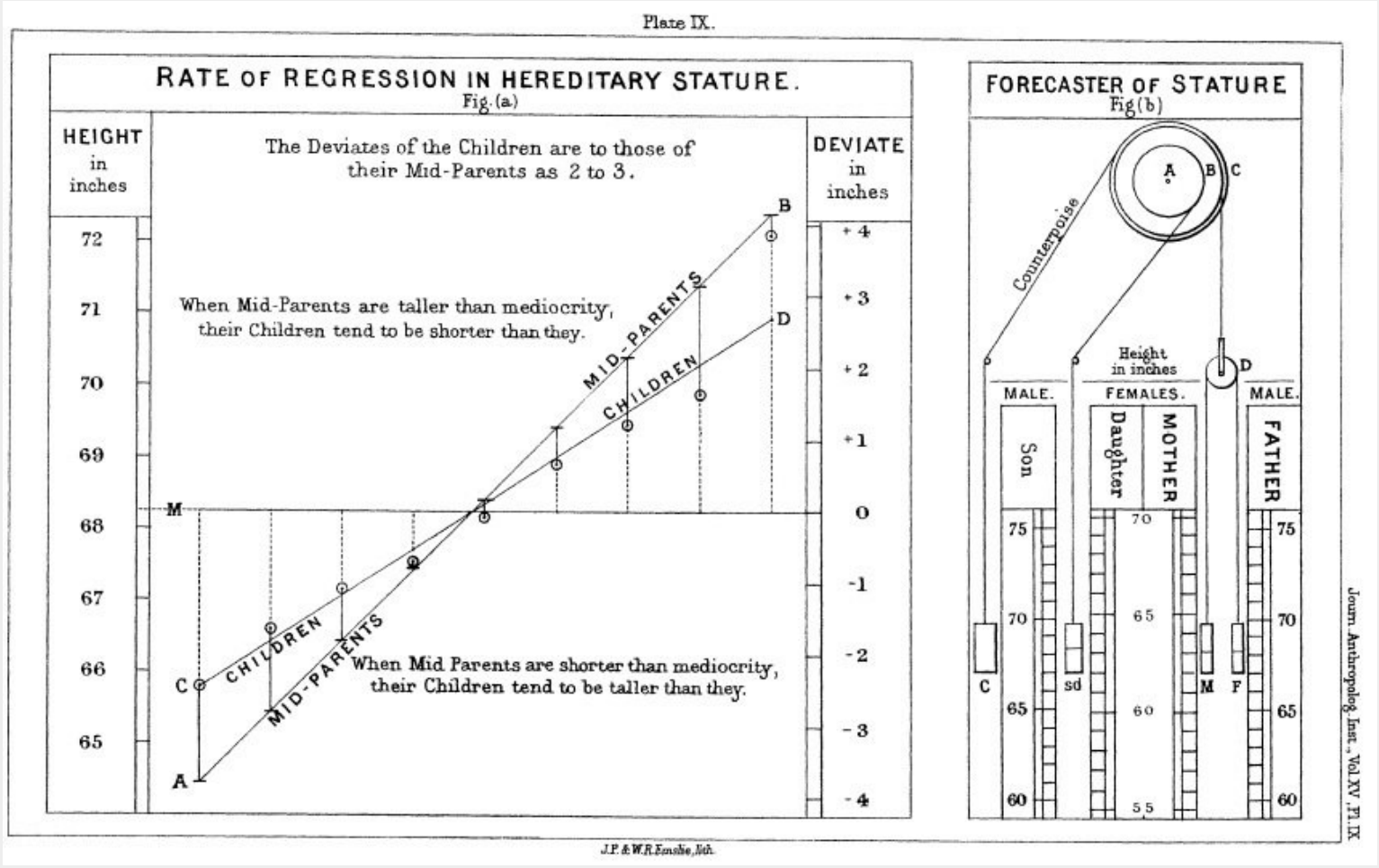
ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

1. O původu lineární regrese II.

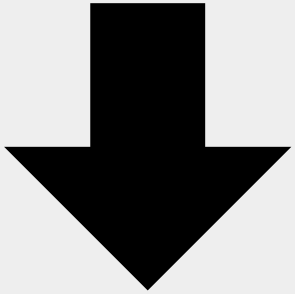


1. O původu lineární regrese III.

Jak to, že děti vysokých rodičů samy bývají vysoké, ale ne tak jako jejich rodiče?

Jak to, že děti útlých rodičů samy bývají útlé, ale ne tak útlé jako jejich rodiče?

Jak to, že nejlepší atlet minulé sezóny letos podává o něco horší výkon než loni?



Regrese k průměru (*Regression towards mediocrity*)

1. O původu lineární regrese IV.

"It appeared from these experiments that the offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they-to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small."

"The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens."

Galton, 1886, s. 246

2. K čemu slouží lineární regrese?

Lineární regrese

- *Nakolik lze z **IQ skóru** usuzovat o **výkonu** v matematice?*
  - **Predikce**

Vícenásobná lineární regrese

- *Přispívá k **výši platu** kromě **úrovně vzdělání** také **pohlaví**?*
  - **Predikce**
  - **Inkrementální validita**
  - **Statistická kontrola**

2. Notace

$$\mathbf{Y = Y' + e}$$

Lineární regrese

$$Y' = a + bX$$

$$Y' = b_0 + b_1X_1$$

Vícenásobná lineární regrese

$$Y' = a + b_nX_n$$

$$Y' = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

Y = *Predikovaná (= závislá; outcome) proměnná*

Y' = *Náš model*

e = *Chyba měření*

a nebo $\mathbf{b_0}$ = *průsečík (= intercept)*

b nebo $\mathbf{b_{1...n}}$ = *směrnice (= slope)*

$\mathbf{X_{1...n}}$ = *Prediktor (= nezávislá proměnná; predictor)*

2. Grafické znázornění

$$Y = Y' + e$$

$$Y' = a + bX$$

$$Y' = b_0 + b_1X_1$$

dle Field, 2009, s. 199

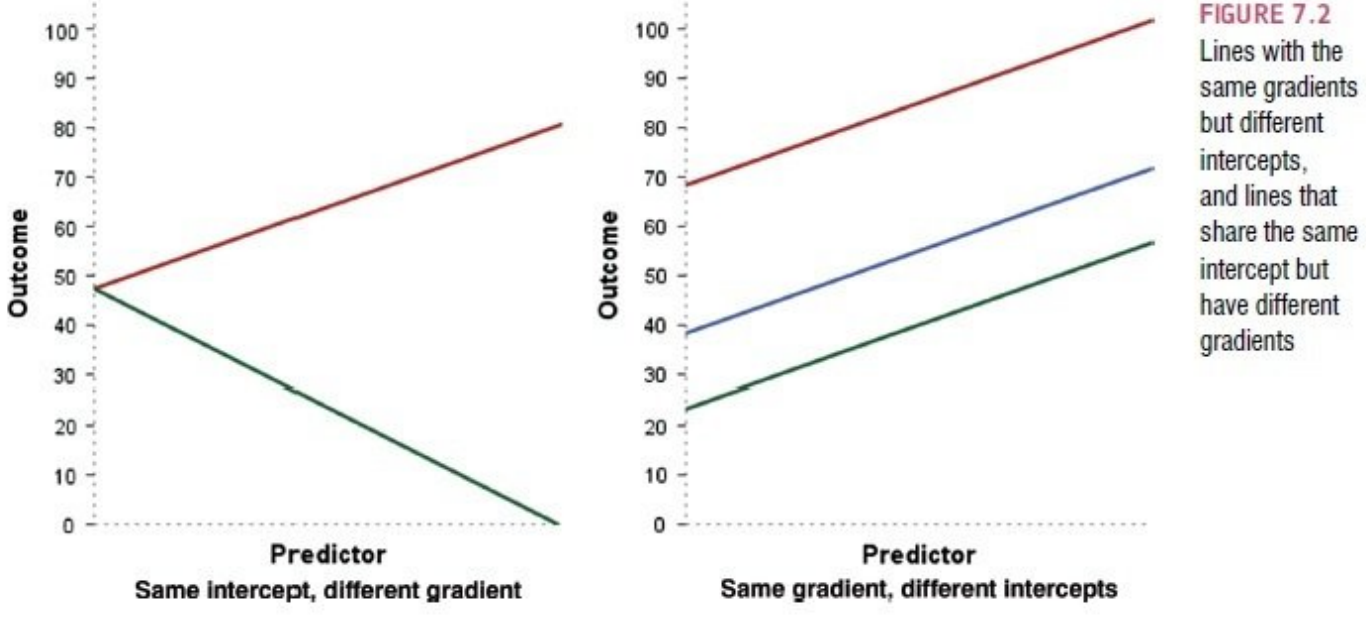


FIGURE 7.2
Lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

2a. Příklad

Skupina pracovníků v podniku BD Technologies si stěžuje vedení firmy, že se roky, které odpracovali ve firmě (X) nepromítají do výše jejich mzdy (Y). Psycholog pracující v tomtéž podniku dostane za úkol zjistit, zda je stížnost pracovníků oprávněná.

Proměnné

$N = 30$

Roky:

- M (5.6), SE (3.7),
- Min (1), Max (15)

Plat:

- M (21 400), SE (9 828),
- Min (10 000), Max (50 000)

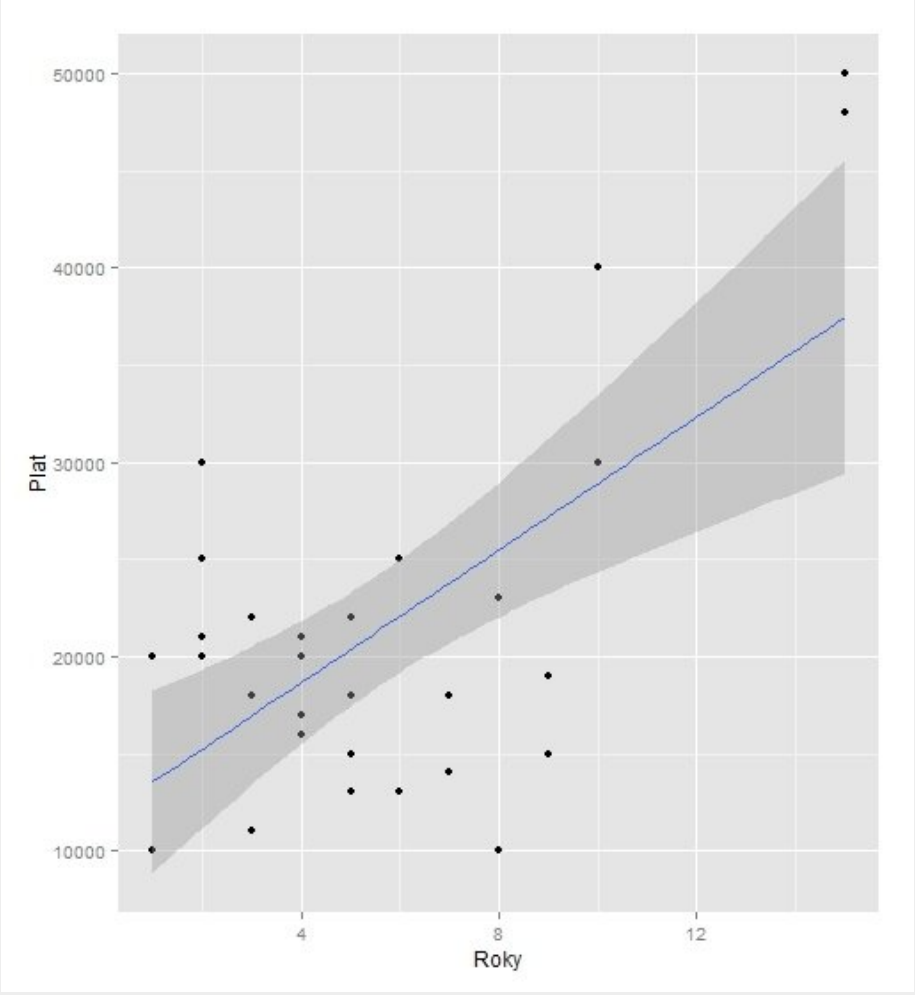
Průsečík a směrnice

$Y' = a + bX$

$Y' = b_0 + b_1X_1$

a, resp. $b_0 = 11829$

b, resp. $b_1 = 1709$



2. Model

Přímka (model) je proložena daty tak, aby jim co nejlépe odpovídala.

Metoda odhadu nejmenších čtverců (***Least Squares Estimation***)

Suma (druhých mocnin) vzdáleností modelu od dat je nejmenší možná

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

$$SS_R = \frac{\sum(Y - Y')^2}{n-1}$$

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

$$s_T^2 = s_M^2 + s_R^2 \ (\textit{neboli}\ SS_T = SS_{res} + SS_{reg})$$

$$R^2 = SS_M^2 \ / \ SS_T^2$$

SS_M = Rozdíl mezi **nulovým modelem** (průměr Y) a námi **stanoveným modelem** (přímkou)

SS_R = Rozdíl mezi **daty** a námi **stanoveným modelem** (přímkou)

SS_T = Rozdíl mezi **daty** a **nulovým modelem** (průměr Y)

R^2 = Podíl rozptylu závislé (outcome) proměnné **vysvětlené modelem** (= ***koeficient determinance*)**

2. Metoda nejmenších čtverců graficky

dle Field, 2009, s. 203

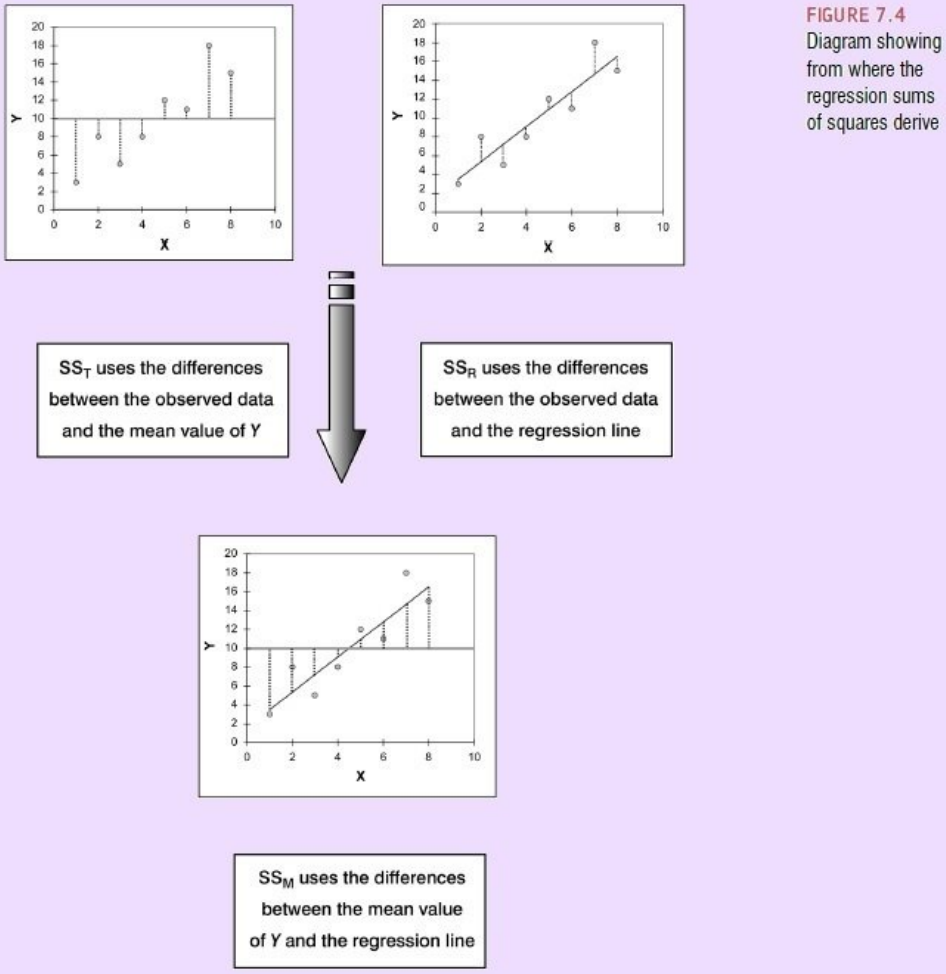


FIGURE 7.4
Diagram showing from where the regression sums of squares derive

2. Koeficienty

b_i

Vyjadřuje nárůst Y′ při nárůstu X_i o jednu jednotku v jednotkách Y, při kontrole všech ostatních prediktorů (tj. semiparciální korelace); jedinečný přínos

- K porovnání síly prediktoru v různých skupinách, modelech, vzorcích

β_i ; Beta

Vyjadřuje nárůst Y′ při nárůstu X_i o 1; jsou-li X_i i Y standardizovány, při kontrole všech ostatních prediktorů (tj. 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

Po vycentrování (odečtení průměru od všech hodnot X1) odpovídá průměru Y.

2a. Příklad - Model a koeficienty

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,635[a]	,404	,383	7723,007	,404	18,965	1	28	,000

a. Predictors: (Constant), q2 Roky
b. Dependent Variable: q1 Plat

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11828,512	2611,302		4,530	,000
	q2 Roky	1709,194	392,481	,635	4,355	,000

a. Dependent Variable: q1 Plat

$$Y' = b_0 + b_1X_1$$

Model

- Zvolený model vysvětluje 40 % rozptylu, tedy
- Délka praxe odpovídá za výši platu ze 40 %

Prediktory

- Y' = 11 828 + 1709*X
- Pracovník, který ve firmě působí 1 rok, by si dle modelu měl vydělat **13537 Kč**

3. Předpoklady použití I.

"To draw conclusions about a population based on a regression analysis done on a sample, several assumptions must be true." (Field, 2009 , s. 220)

Proměnné

1. **Povaha proměnných** - spojité, kvantitativní a kardinální nebo dummy (jen v případě prediktorů).
2. Nenulová **variabilita** prediktorů (tj. nejde o konstantu).

Prediktory

3. Absence (dokonalé) **multikolinearity** - prediktory by spolu neměly **vysoce** korelovat.
4. Prediktory nekorelují s vnějšími proměnnými - **absence třetí** (intervenující, vnější) **proměnné**.

3. Předpoklady použití II.

Rezidua

5. **Homoskedascita** - rozptyl reziduí by měl být konstantní napříč různými úrovněmi prediktoru

6. **Nezávislost reziduí** - Reziduální hodnoty kterýchkoliv dvou případů by spolu neměly souviset.

7. **Normálně rozložená rezidua** - jejich rozložení by mělo být náhodné

Outcome

8. **Nezávislost** kterýchkoliv dvou hodnot závislé proměnné (každá hodnota v rámci ní pochází z unikátního zdroje)

9. **Linearita** - přímka jako vhodný model popisu dat.

3a. Příklad

dle Field, 2009, s. 248

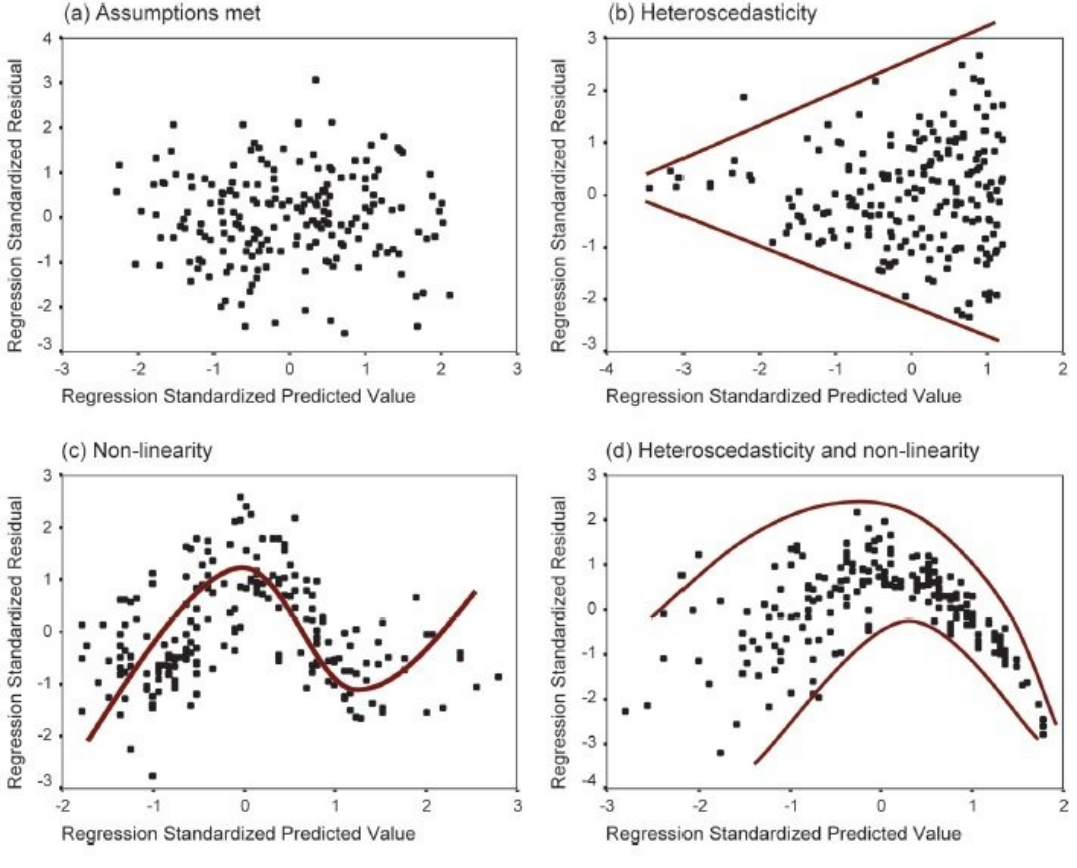


FIGURE 7.19 Plots of *ZRESID against *ZPRED

4. Diagnostika I. - Outliery a vlivné případy

Nemají některé případy příliš velký vliv na výsledky regrese?

- **Outliery** – mohou zvyšovat i snižovat b
  - **Rezidua** – případy s vysokými rezidui regrese predikuje nejhůř, standardizovaná, ± 3
  - **Vlivné případy** – případy, které nejvíc ovlivňují parametry modelu
    - 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)

4. Diagnostika II. - Kolinearita

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

5. Dummy coding I. - obecně a postup

Dummy proměnné - kategorické proměnné **upravené** tak, aby mohly vstoupit do (vícenásobné) lineární regrese

Postup (dle Field, 2009, s. 254)

1. Count the number of groups you want to recode and subtract 1.
2. Create as many new variables as the value you calculated in step 1. These are your dummy variables.
3. Choose one of your groups as a baseline (i.e. a group against which all other groups should be compared). This should usually be a control group, or, if you don't have a specific hypothesis, it should be the group that represents the majority of people (because it might be interesting to compare other groups against the majority).
4. Having chosen a baseline group, assign that group values of 0 for all of your dummy variables.
5. For your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
6. For the second dummy variable assign the value 1 to the second group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
7. Repeat this until you run out of dummy variables.
8. Place all of your dummy variables into the regression analysis!

5. Dummy coding II. - Kódování

Indikátorové kódování (*Indicator coding*)

- Referenční kategorie = 0

Efektové kódování (*Effect coding*)

- Referenční kategorie = -1

Úroveň vzdělání	Původní hodnota	Indikátorové kódování		Efektové kódování	
		Vysokoškolské	*Středoškolské*	*Vysokoškolské*	*Středoškolské*
Vysokoškolské	1	1	0	1	0
Středoškolské	2	0	1	0	1
Základní	3	0	0	-1	-1

5. Dummy coding III. - Interpretace

$$Y = b_0 + b_{A1}X_{A1} + b_{A2}X_{A2} + \ldots + b_mX_m + e$$

- Po dosazení do regresní rovnice predikujeme případu **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 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

5a. Příklad

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,894[a]	,799	,776	4655,458	,799	34,416	3	26	,000

a. Predictors: (Constant), qVS qVS, q2 Roky, qSS qSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4952,782	2296,642		2,157	,040
	q2 Roky	1612,338	240,268	,599	6,711	,000
	qSS qSS	4754,252	2236,356	,241	2,126	,043
	qVS qVS	15044,797	2250,920	,750	6,684	,000

a. Dependent Variable: q1 Plat

$$Y = b_0 + b_1 Plat + b_2 SŠ + b_3 VŠ + e$$

Interpretace - Model:

- Přidání stupně vzdělání zlepšilo predikční vlastnosti modelu na 80 %.
- Výše platu ve firmě BD Technologies se tedy z 80 % odvíjí od let praxe a dosaženého stupně vzdělání.

Interpretace - Prediktory

- *Středoškolské vzdělání* garantuje **ve srovnání** s tím základním **průměrně** o 5 tisíc Kč větší plat.
- *Vysokoškolské vzdělání* garantuje **ve srovnání** s tím základním **průměrně** o 0,75 směrodatnou odchylku větší plat.

6. Vkládání prediktorů I.

SPSS nabízí 4 způsoby:

ENTER (Forced entry)

Vloží všechny prediktory najednou

BLOCKWISE

Vkládání sady prediktorů po blocích

STEPWISE

FORWARD

Vybere prediktory, které nejlépe odpovídají datům - až po stanovenou mez

BACKWARD

Vyřadí prediktory nejhůře odpovídající datům - až po stanovenou mez

6. Vkládání prediktorů - dovětek k BLOCKWISE I.

- 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

Obvyklé řazení bloků

- Od známých k neznámým vlivům
    - kontrola intervenujících proměnných
    - Minimalizace chyby 1. typu
- Podle výzkumné relevance
    - Od ústředních po „co kdyby“; maximalizace statistické síly

6. Vkládání prediktorů - dovětek k BLOCKWISE II.

Obvyklý postup

- Na základě teoretických rozvah stanovíme různé modely, jejichž srovnání je potenciálně zajímavé
  - Možnost testovat nárůst (inkrement) R^2
- Až v druhé řadě se zabýváme jednotlivými regresními koeficienty v modelu, který je nejúplnější/nejlepší

7. Reportování (více např. dle APA, 2001)

1. Popisné statistiky

- Y, X
  - **Spojité** - *N, Min, Max, M, SD, Me*
  - ***Kategorické*** - *N, %, dummy coding*
- Korelační matice

2. Předpoklady použití

- Konstatování (např. o povaze proměnných
- Výpočet (např. outliery a vlivné příklady)

3. Model

- F-test
- Koeficient determinance (R^2)
- *p*

4. Prediktory

- B
- SE či intervaly spolehlivosti
- Beta
- *p*

Děkuji za pozornost!

Zdroje

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

Field, A. (2009). *Discovering statistics using SPSS*, 3th Ed. Los Angeles: Sage.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models,* 3th Ed. Los Angeles: Sage.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, pp. 246-63. Dostupné online z "http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf"

Robotková, A., & Ježek, S. (2012). Vícenásobná lineární regrese. Prezentace ke kurzu PSY252.

Úkol

Na zvolených datech proveďte vícenásobnou lineární regresi.

Požadavky:

- **Minimálně** jeden prediktor ***spojité*** (kardinální etc.) povahy a minimálně jeden prediktor kategorické (***dummy***) povahy
- Celý proces - od ***popisu proměnných*** přes ***předpoklady*** a ***interpretaci výsledků*** po ***diagnostiku modelu***

Bonus:

- **1 bod** obdrží ten, kdo
  - ***Představí dataset*** a ***proměnné*** na úrovni ***konceptů*** (např. z jakých položek se skládá škála well-beingu) a zároveň
  - Volbu ***proměnných*** (a hypotéz) a ***interpretaci výsledků*** podloží *odbornými zdroji*