

Jednoduchá regresní analýza a její předpoklady

Kvantitativní metody zkoumání politiky (JPB283)

Michal Parízek

28. 4. 2025

Institut politologických studií FSV UK

Obsah

- Zpátky k principům regresní analýzy
- Jak dobrá je naše regresní přímka?
- Předpoklady regresní analýzy a diagnostika

Zkouška

- 12+1 otázek
- Výpočty i teorie – výpočetní, uzavřené, otevřené otázky
- Limit: 2 hodiny, v PC učebně (zkontrolujte si přístup do PC [heslo]!)
- Během zkoušky odpojení od internetu (není možné pracovat s online materiály)
- Možnost přinést si offline materiály na USB
- Výpočetní úlohy znovu otevřeny pro trénink (bez bodů)

Zkouška - termíny

- 19.5. 8:30-10:30 a 13:00-15
- 26.5. 13:30-18 (dvě okna)
- 9.6. 9-14 (dvě okna)
- 23.6. 9-14 (dvě okna)

Přehled testů pro vztah různých druhů proměnných

Nezávislá proměnná	Závislá proměnná	
	Diskrétní (nominální, ordinální)	Spojitá (intervalová, poměrová)
Diskrétní (nominální, ordinální)	Chí-kvadrát	T-test (rozdíl průměrů)
Spojitá (intervalová, poměrová)	Logit, Probit	Korelace, Regrese

Zpátky k regresní analýze

Základní kroky regresní analýzy

Základní myšlenka: přímka zachycuje vztah mezi dvěma spojitými proměnnými

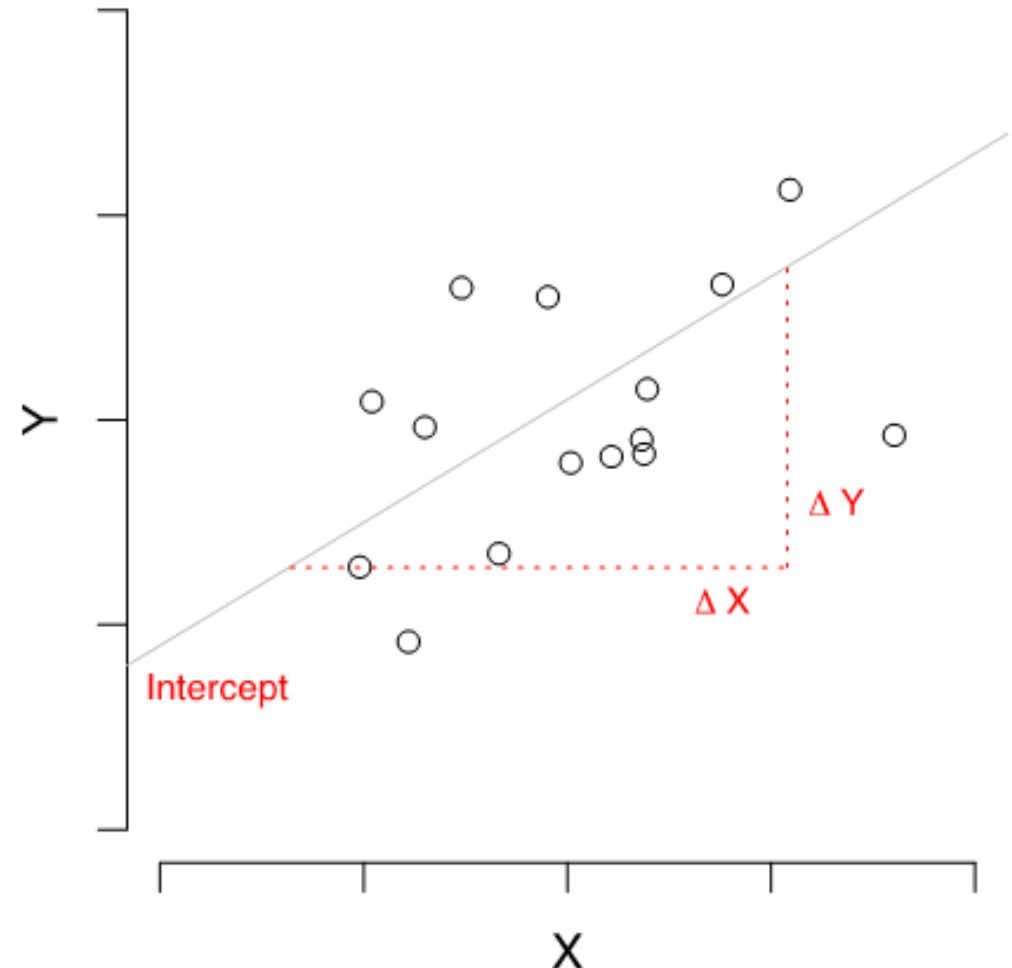
Kroky analýzy:

1. na základě dat odhadneme tvar (sklon) přímky
2. určíme **jestli je sklon odlišný od nuly**, protože nenulový sklon znamená vztah mezi nezávislou a závislou proměnnou
3. zjistíme, jak dobře náš model jako celek vysvětluje variaci na závislé proměnné (její chování)

Přímka a její sklon určí, jestli je mezi proměnnými vztah

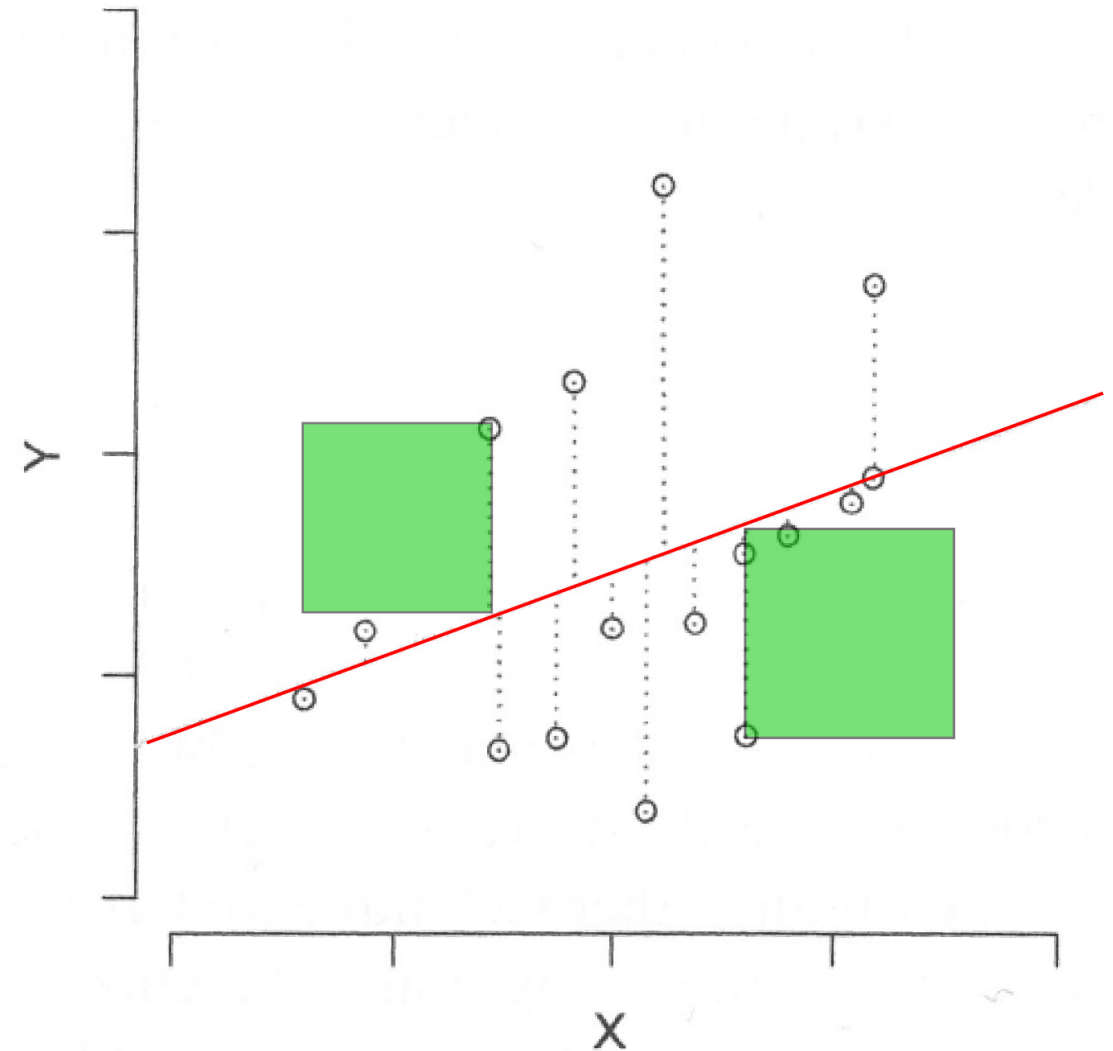
$$Y = a + bX + e$$

- Proložme daty přímkou
- a je konstanta
- b je sklon přímky, tedy změna hodnoty závislé proměnné Y v důsledku změny na X
- e je rozdíl mezi přímkou a pozorováním (error)



Metoda nejmenších čtverců (ordinary least squares, OLS)

- Volíme tu přímku, která **minimalizuje rezidua**, tj. rozdíl na Y mezi přímkou samotnou (kterou hledáme) a jednotlivými pozorováními
- Pro zamezení krácení $+/-$ chyb počítáme s druhými mocninami (čtverci)



Regresní rovnice

- Koeficienty b (sklon) a a (průsečík s Y) se dají snadno spočítat

- $$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

- $$a = \bar{Y} - b\bar{X}$$

- Výsledkem je regresní rovnice zachycující pro každé X predikovanou hodnotu $\hat{Y}$

$$\hat{Y} = a + bX$$

Jak dobrá je naše regresní
přímka?

Kolik nám toho o závislé proměnné naše analýza řekne?

Rovnice $\hat{Y} = a + bX$
zachycuje nejlepší možnou
přímku proloženou daty...



<http://memecreator.cz/meme/ptm3x9>

Kolik nám toho o závislé proměnné naše analýza řekne?

1) Kolik variability závislé proměnné vlastně přímka vysvětluje?

2) A můžeme věřit tomu, že sklon b je skutečně nenulový? Tedy že je mezi proměnnými skutečně vztah?



<http://memecreator.cz/meme/ptm3x9>

Ad 1) Koeficient determinace r^2 (většinou R^2)

- Jak zhodnotit míru vysvětlující síly regresní přímky vůči realitě?

- porovnáním **objemu vysvětlené variace Y** vůči **celkové variaci Y**

- $ESS_y = \sum(\hat{Y} - \bar{Y})^2$

- $SS_y = \sum(Y - \bar{Y})^2$

- ESS, explained sum of squares

- SS, total sum of squares

Ad 1) Koeficient determinace r^2 (většinou R^2)

$$r^2 = \frac{ESS_y}{SS_y}$$

- $r^2 = 0$ znamená, že regresní model nijak nezlepšuje naše chápání Y
- $r^2 = 1$ znamená, že model perfektně vysvětluje Y

Ad 2) Můžeme už prosím *konečně* říct něco o vztahu mezi X a Y ?

- Mezi proměnnými **existuje vztah**, pokud má identifikovaná nejlepší přímka **nenulový sklon**, tj. $b \neq 0$
- Ovšem pracujeme s nedokonalými odhady na základě vzorku a nikoliv dokonalou znalostí populace (reality)
- Platí $b \neq 0$ s **dostatečnou mírou jistoty** vzhledem k tomu, kolik je v našich datech prostoru pro náhodné chyby?

Ad 2) Můžeme už prosím *konečně* říct něco o vztahu mezi X a Y ?

- Celková průměrná míra chyby naší predikce Y ...

$$MSe = \frac{\sum (Y - \hat{Y})^2}{N - 2}$$

- ...vztažená k celkovému rozptylu na X

$$s_b = \sqrt{\frac{MSe}{s_X^2 \times N - 1}}$$

- *Dává...* směrodatnou chybu odhadu sklonu b

Ad 2) Můžeme už prosím *konečně* říct něco o vztahu mezi X a Y ?

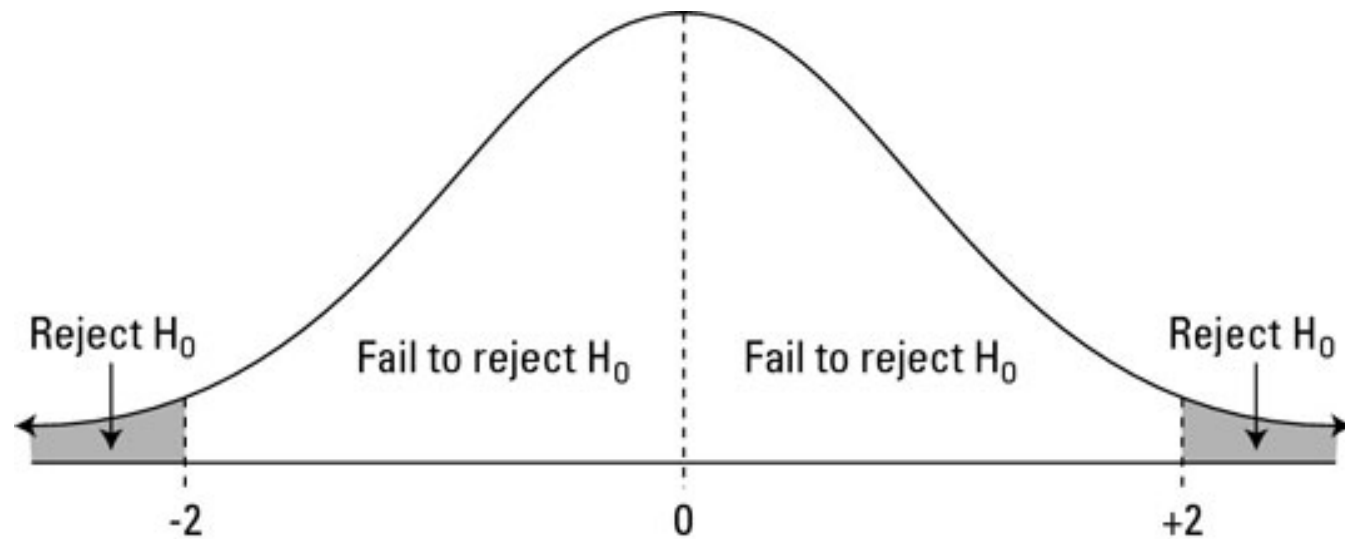
- a konečně spočítáme t-statistiku koeficientu b (jako v t-testu)

$$t = \frac{b}{s_b}$$

- Ta nám řekne, jak daleko je naše b od 0
- a podle statistické tabulky zjistíme, s jakou pravděpodobností je toto získané b skutečně odlišné od 0

Co to sakra znamená?

- Je hodnota koeficientu b dostatečně daleko od 0? Pokud ano, je mezi nezávislou a závislou proměnnou pravděpodobně **statistický vztah a vyvrácíme H_0**
- Jak měříme „dostatečnost“ vzdálenosti od 0? Porovnáním hodnoty b s mírou chyby v odhadu velikosti b .



Aplikace: predikce vztahu k EU na základě vztahu strany k imigraci

- Predikuje pozice strany vůči imigraci (na škále 0-10, kde 10 je silná restrikce imigrace) pozici strany vůči EU (na škále 1-7, kde 7 je silná podpora integraci)?
- Data z Chapel Hill Expert Survey (2014), 268 stran s kompletními daty
- Hypotéza? Negativní vztah...

	A	B	C
1	eu_position	immigrate_policy	
2	6	2.8	
3	6	3.4000001	
4	6.25	1.8	
5	6.1999998	2.2	
6	6.4000001	6.1999998	
7	6.5999999	6.5999999	
8	6.4000001	4.1999998	
9	6.5999999	5.4000001	
10	5	7.5999999	
11	6.4000001	4.4000001	
12	2.5999999	9.6000004	
13	3.4000001	1.8	
14	2.5	8.5	
15	6	5.5	
16	7	2.5999999	
17	5.5	7.0999999	
18	4.6999999	2.8	

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.296859493							
R Square	0.088125559							
Adjusted R Square	0.084684523							
Standard Error	1.703059704							
Observations	267							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	74.28009692	74.2801	25.610185	7.82155E-07			
Residual	265	768.6092746	2.900412					
Total	266	842.8893715						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>ower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	5.977427739	0.252457526	23.67696	2.4065E-67	5.480349908	6.47450557	5.48035	6.47450557
immigrate_policy	-0.221459424	0.043761058	-5.06065	7.8216E-07	-0.307623033	-0.135295815	-0.30762	-0.135295815

[illegible]

Regression Statistics

Multiple R	0.296859493
R Square	0.088125559
Adjusted R Square	0.084684523
Standard Error	1.703059704
Observations	267

[illegible]

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	74.28009692	74.2801	25.610185	7.82155E-07
Residual	265	768.6092746	2.900412		
Total	266	842.8893715			

a

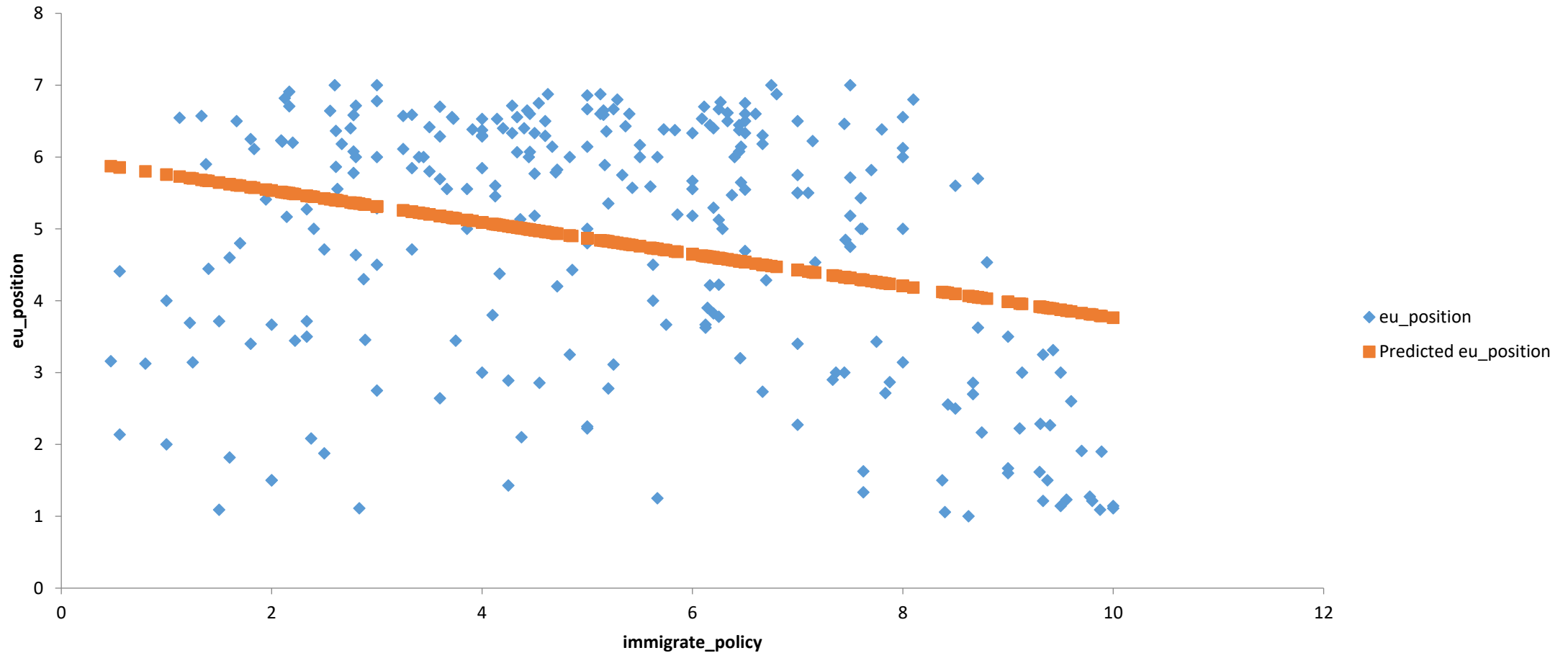
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	5.977427739	0.252457526	23.67696	2.4065E-67	5.480349908	6.47450557	5.48035	6.47450557
immigrate_policy	-0.221459424	0.043761058	-5.0606	7.8216E-07	-0.307623033	-0.135295815	-0.30762	-0.135295815

b

 $P(b=0)$

$$\widehat{eu_position} = 5.98 - 0.22 \times immigrate_policy$$

immigrate_policy Line Fit Plot

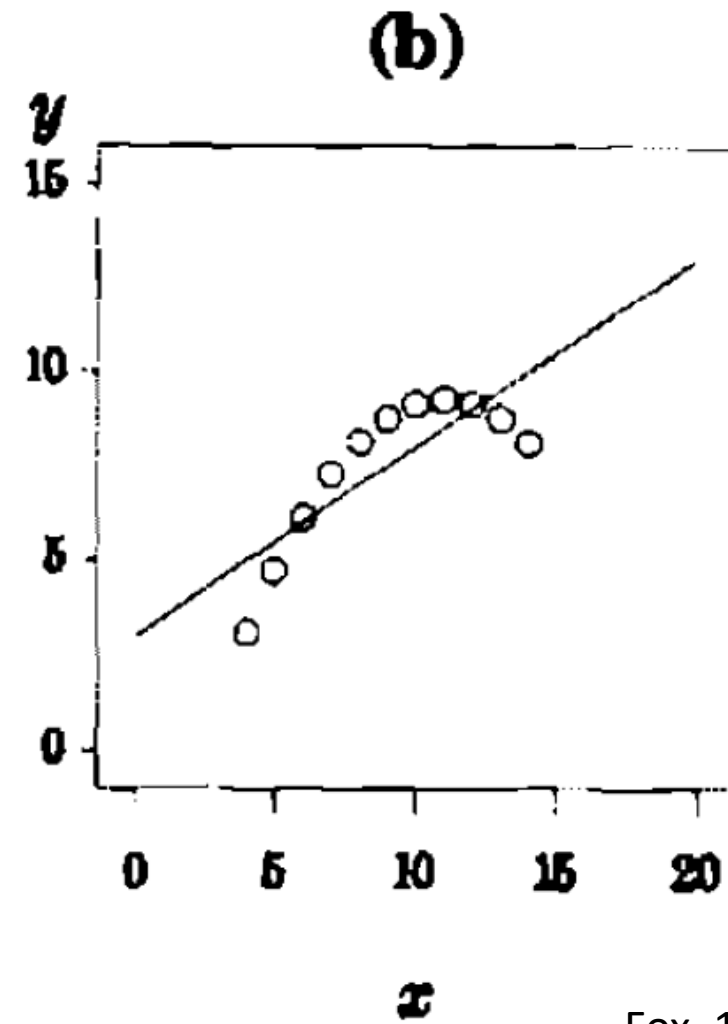
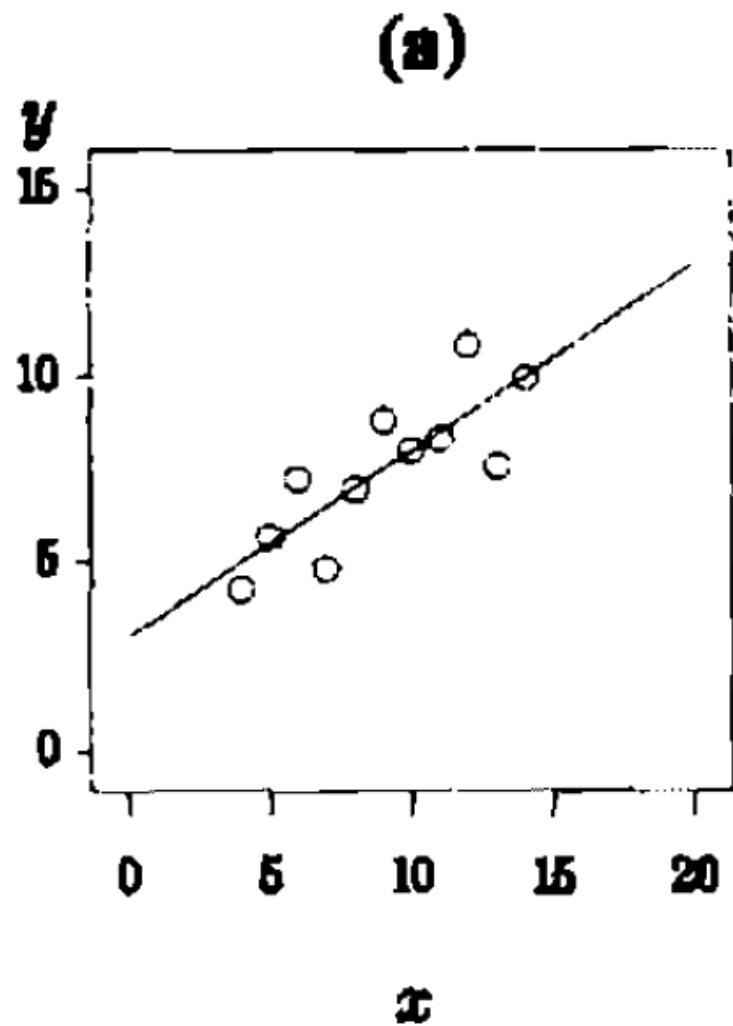


$$\widehat{eu_position} = 5.98 - 0.22 \times immigrate_policy$$

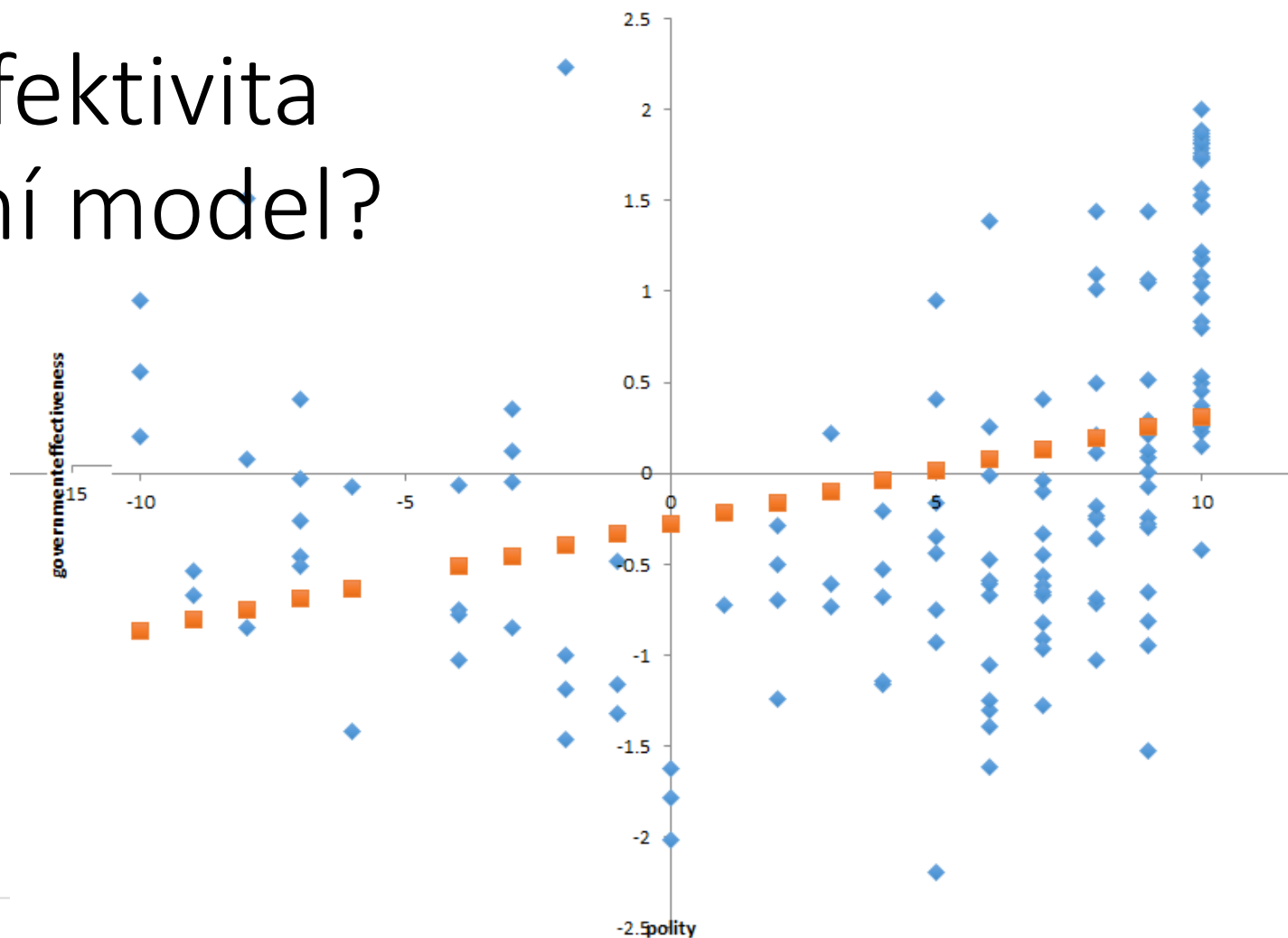
- Jakou hodnotu pozice vůči EU **v průměru** očekáváme u strany, která je extrémně proti-imigrační? ($? = 5.98 - 0.22 \times 10 = 5.98 - 2.2 = 3.78$)
- Jakou hodnotu pozice vůči EU **v průměru** očekáváme u strany, která je extrémně pro-imigrační? ($? = 5.98 - 0.22 \times 0 = 5.98$)
- Jakou hodnotu pozice vůči EU **v průměru** očekáváme u strany, která je průměrně pro/proti-imigrační? ($immigrate_policy=5.25$) ($? = 5.98 - 0.22 \times 5.25 = 5.98 - 1.155 = 4.83$)
- Kolik % variability na pozici vůči EU vysvětluje u stran jejich pro/proti-imigrační postoj?

Předpoklady regresní analýzy a diagnostika

Linearita vztahu



Demokracie a efektivita vládnutí: lineární model?



SUMMARY OUTPUT

Regression Statistics

Multiple R 0.343670487

R Square 0.118109404

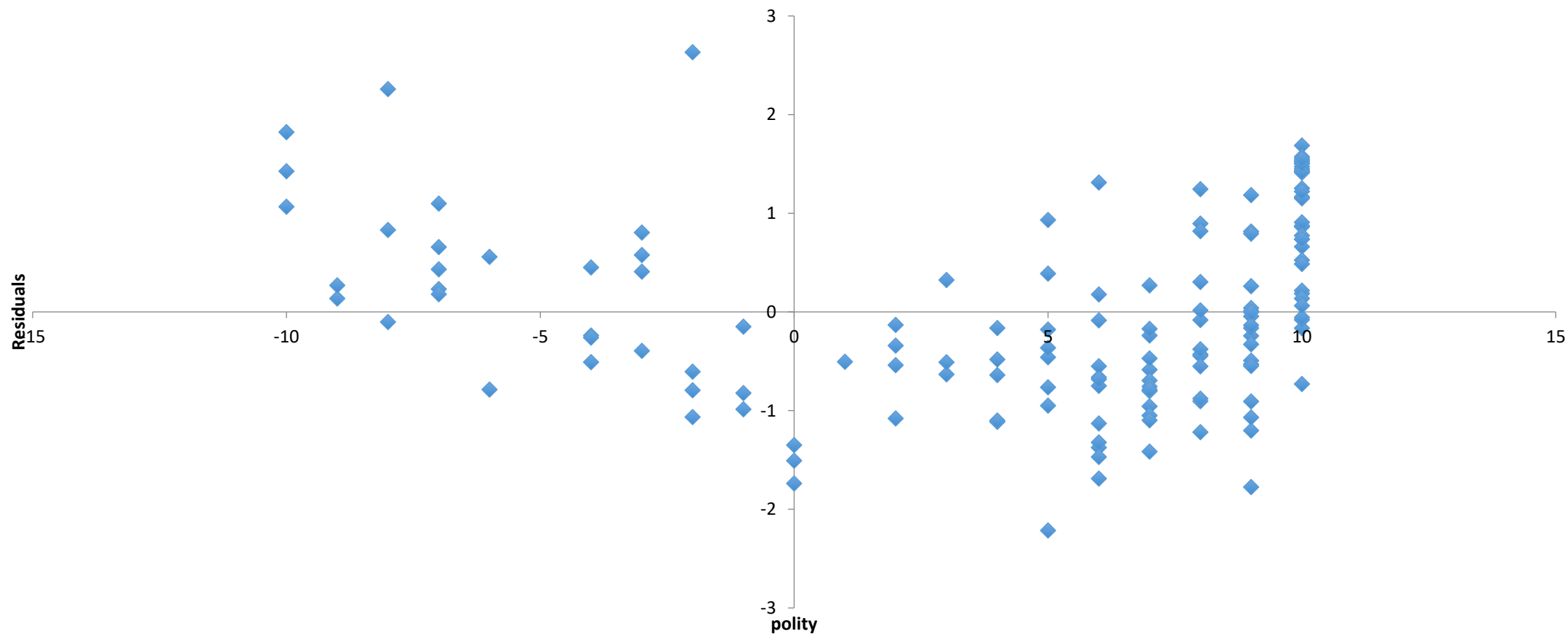
Adjusted R 0.111898907

Standard Error 0.942931017

Observations 144

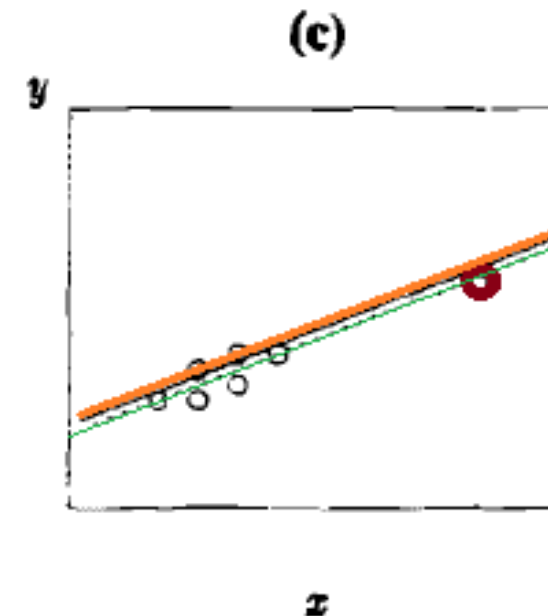
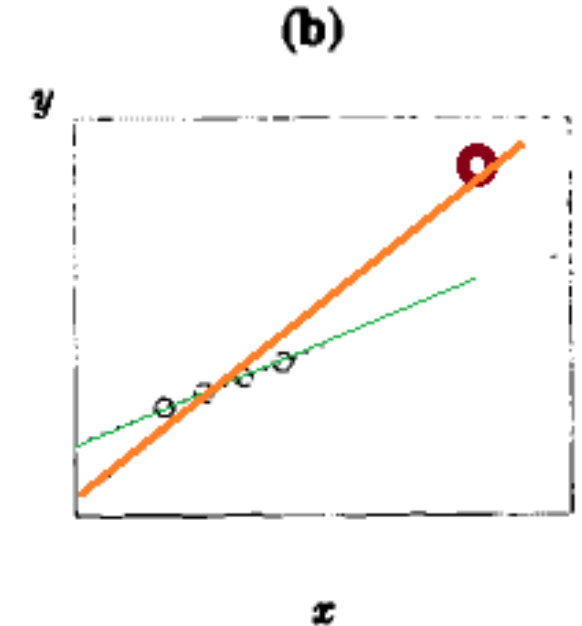
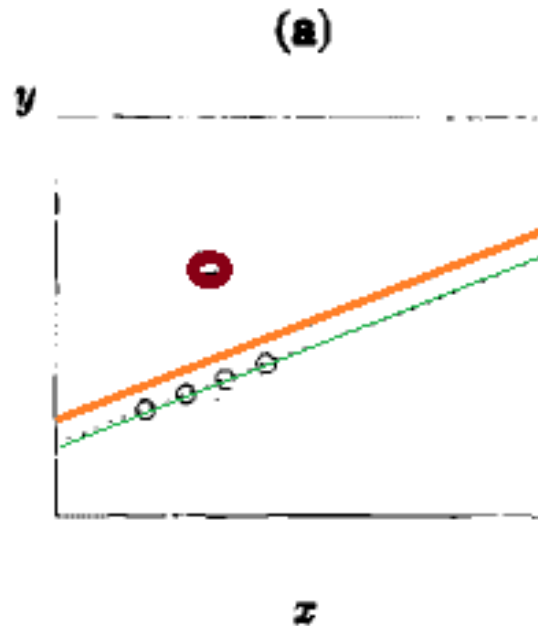
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.276914247	0.101708749	-2.72262	0.007289	-0.47797	-0.07586	-0.47797	-0.07586
polity	0.058943139	0.013516187	4.360929	2.47E-05	0.032224	0.085662	0.032224	0.085662

Demokracie a efektivita vládnutí: nelineární vztah v reziduích



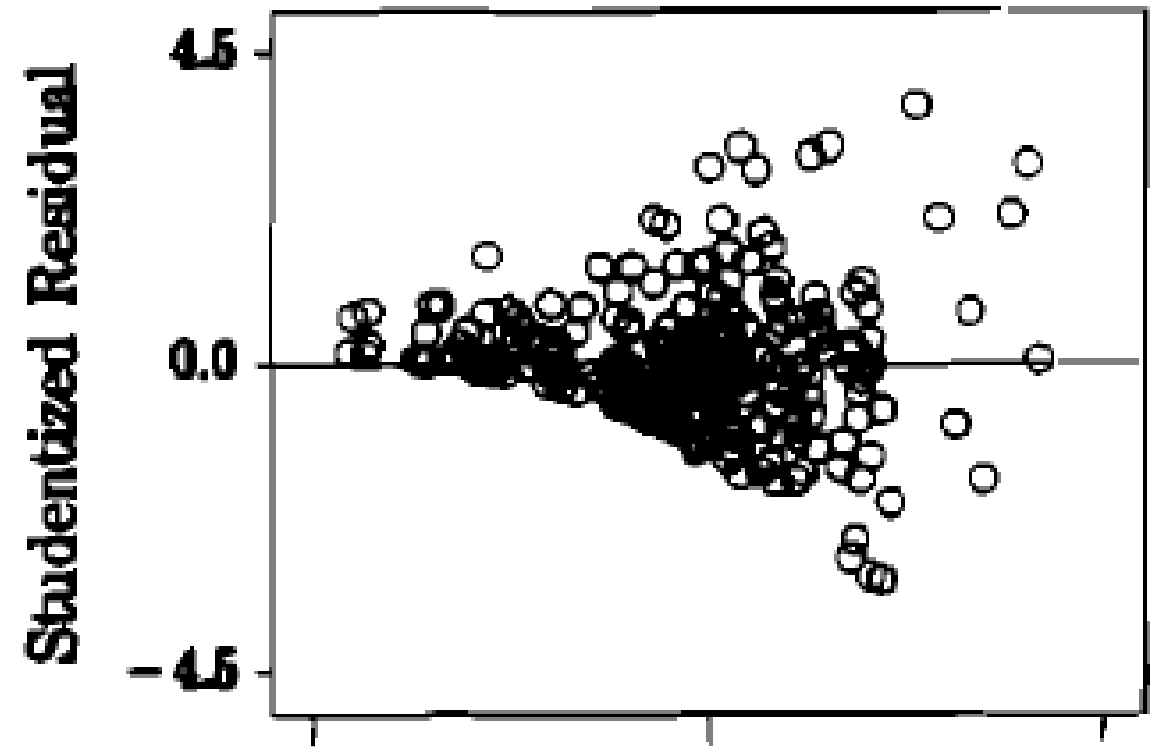
Vlivná pozorování

- odlehlá pozorování (outliers) jsou pozorování s neobvyklou hodnotou na závislé proměnné vzhledem k hodnotě na nezávislé proměnné
- odlehlé pozorování není nutně vždycky vlivné pozorování (influential observation)



Homoskedasticita

- konstantní směrodatná odchylka reziduí
- popření předpokladu je známé jako **heteroskedasticita** reziduí
- chyby modelu se mění v závislosti na hodnotě predikce nebo hodnotě nezávislé proměnné



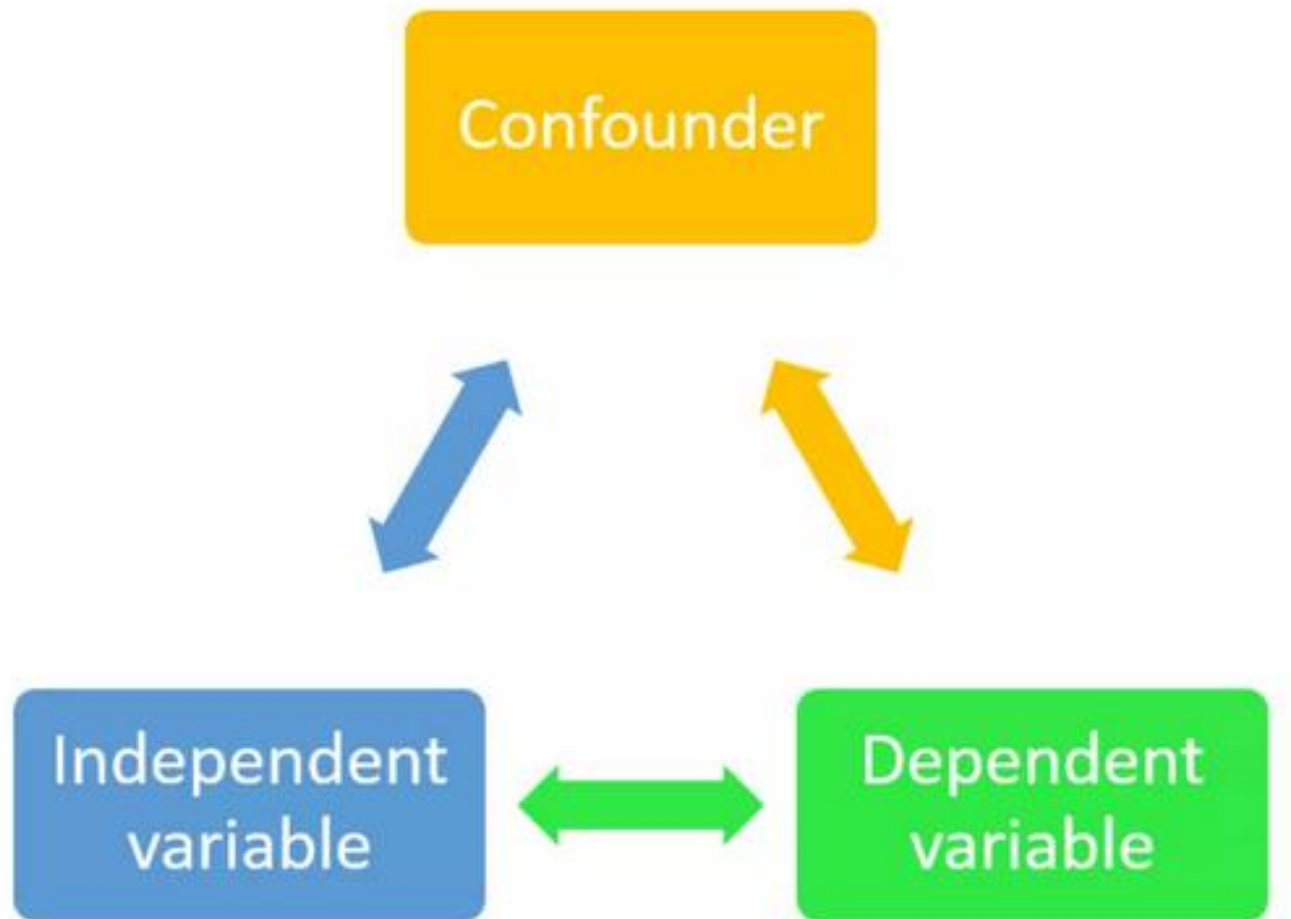
Vzájemná nezávislost chyb (a tedy i pozorování)

- každé pozorování nese informaci
- ale vzájemně závislá pozorování nesou **méně informace**
- časové řady, časové vazby
- domácnosti, prostorové vazby



"Yes, I've heard long-married couples tend to look alike, but I'm not sure about the dress alike part."

Zkreslení kvůli chybějící proměnné





Odkazy

- Fox, John. *Regression Diagnostics: An Introduction*. 1st ed. SAGE Publications, Inc, 1991.
- Gerring, John, and Dino Christenson. *Applied Social Science Methodology: An Introductory Guide*. Cambridge, United Kingdom ; New York: Cambridge University Press, 2017..