

Vícenásobná regrese (a něco navíc)

Kvantitativní metody zkoumání politiky (JPB283)

Michal Parížek

5. 5. 2025

Institut politologických studií FSV UK

Obsah

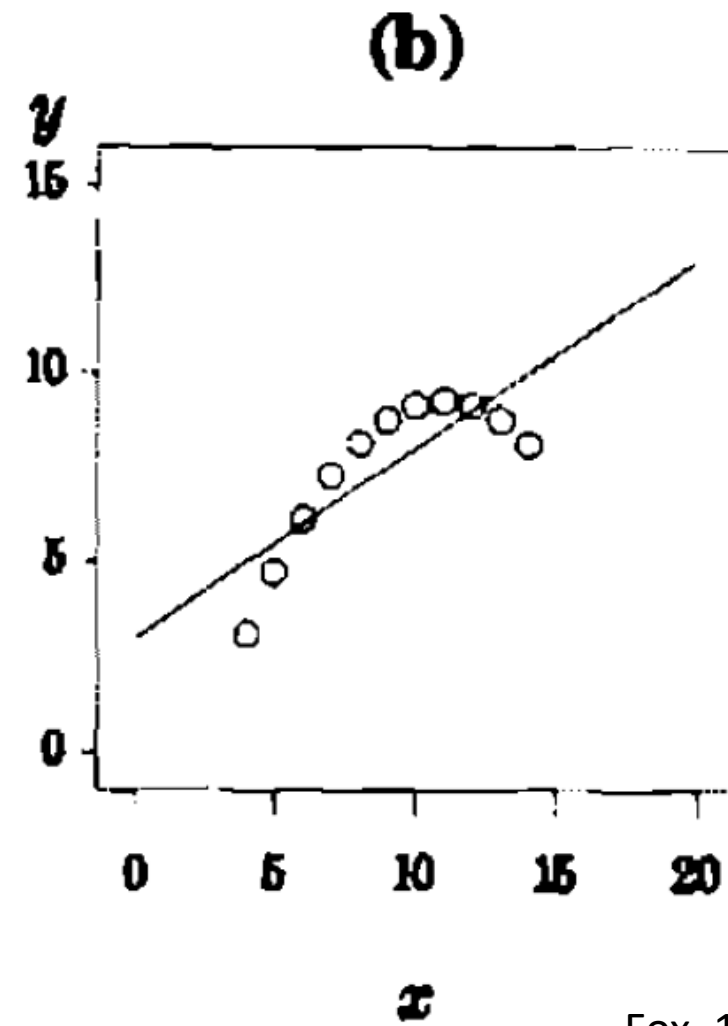
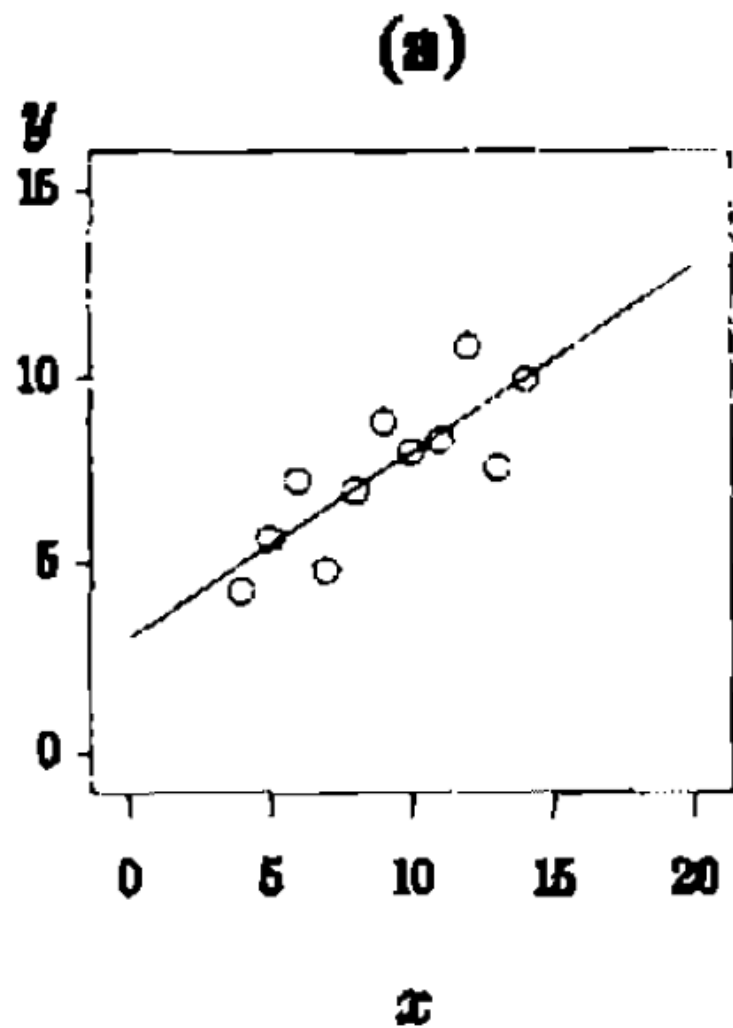
- Předpoklady regresní analýzy a diagnostika
- Vícenásobná regresní analýza
- Analýza na vyšší úrovni

Zkouška

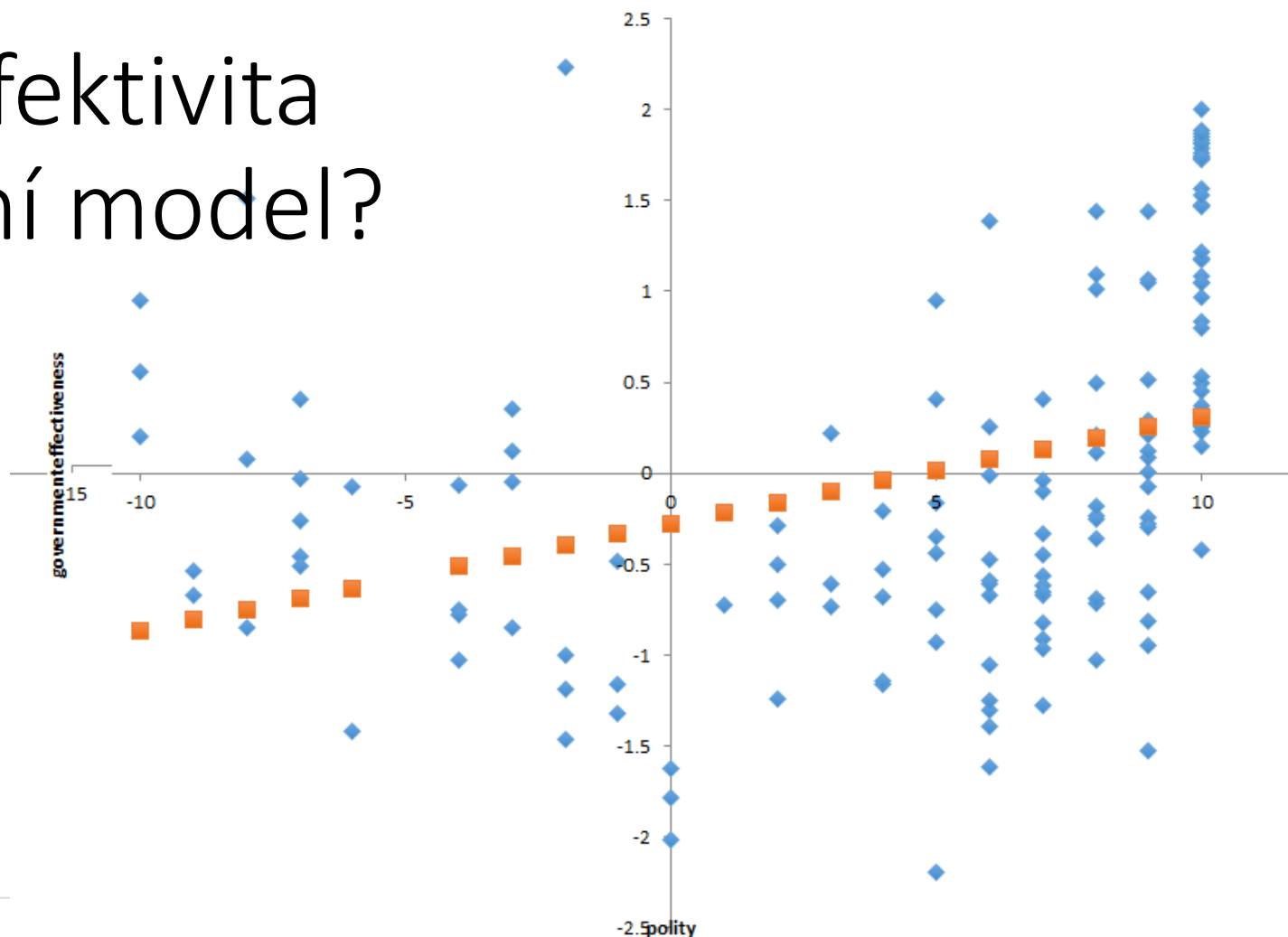
- Termíny v SIS
- Látky z celého semestru
- Materiály včetně možnosti opakování výpočetních úkolů budou předem v Moodlu
- Zkouška v PC místnosti B229, vždy dvě skupiny v daném zkouškovém termínu
- Bez přístupu k internetu!
- Možnost přinést si tištěné a digitální materiály na USB
- Povinnost ověřit s přístup do systému (login) a spuštění Excelu s modulem Analýza dat

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

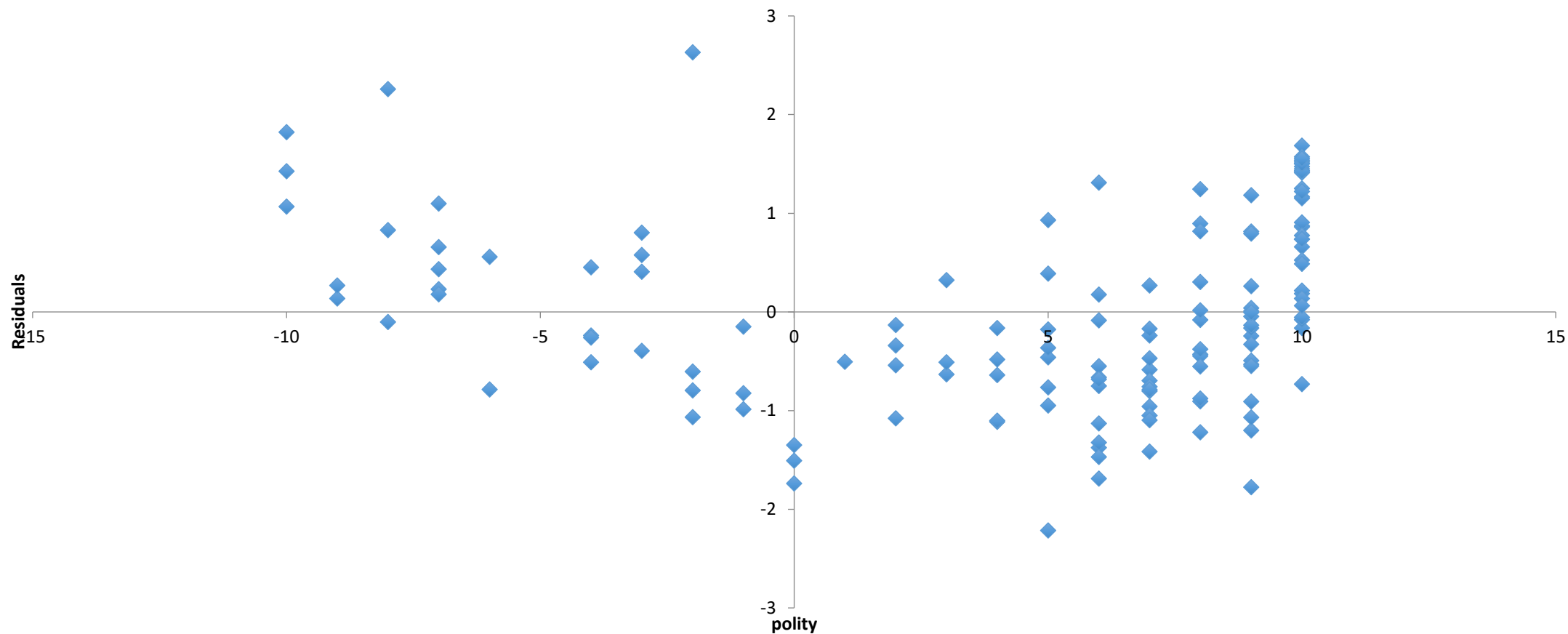
Adjusted R 0.111898907

Standard Error 0.942931017

Observations 144

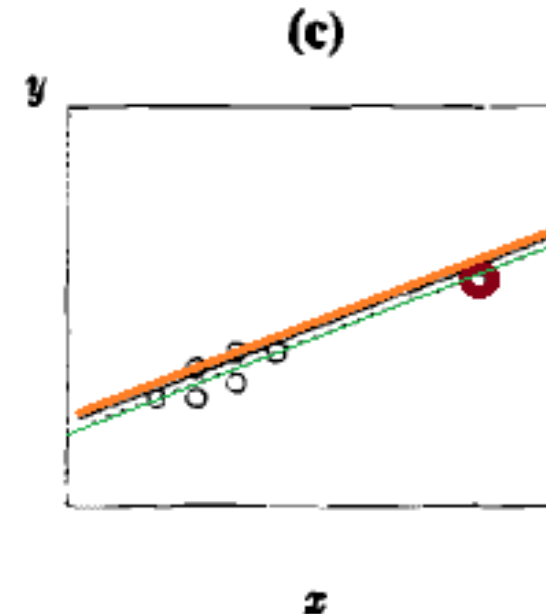
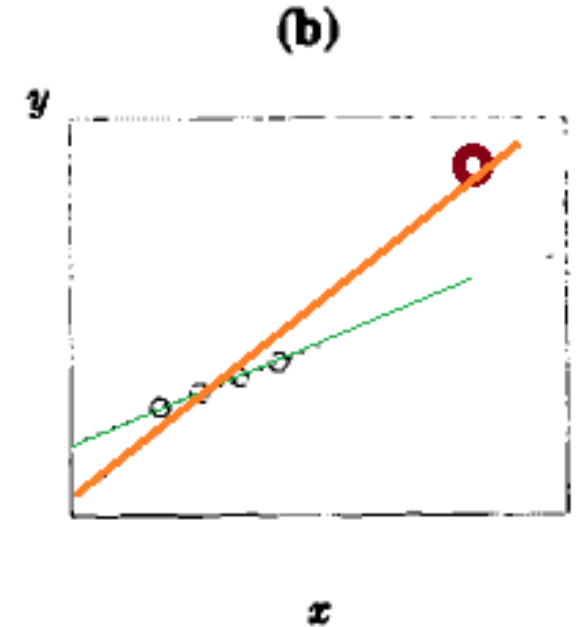
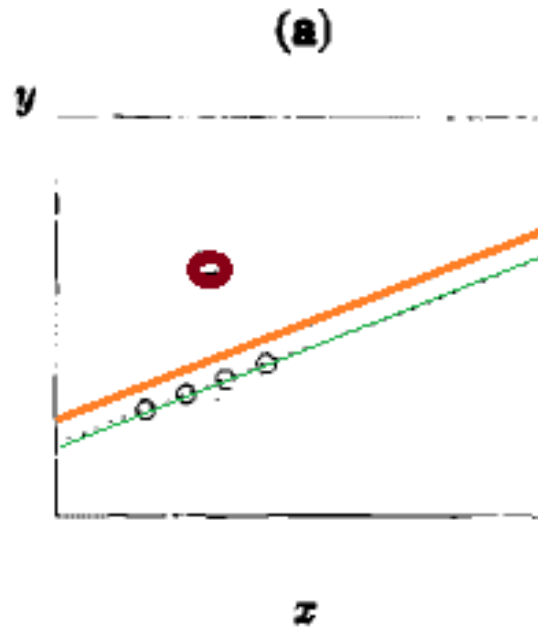
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.276914247	0.101708749	-2.72262	0.007289	-0.47797	-0.07586	-0.47797	-0.07586
polity	0.058943139	0.013516187	4.360929	2.47E-05	0.032224	0.085662	0.032224	0.085662

Demokracie a efektivita vládnutí: nelineární vztah v reziduích



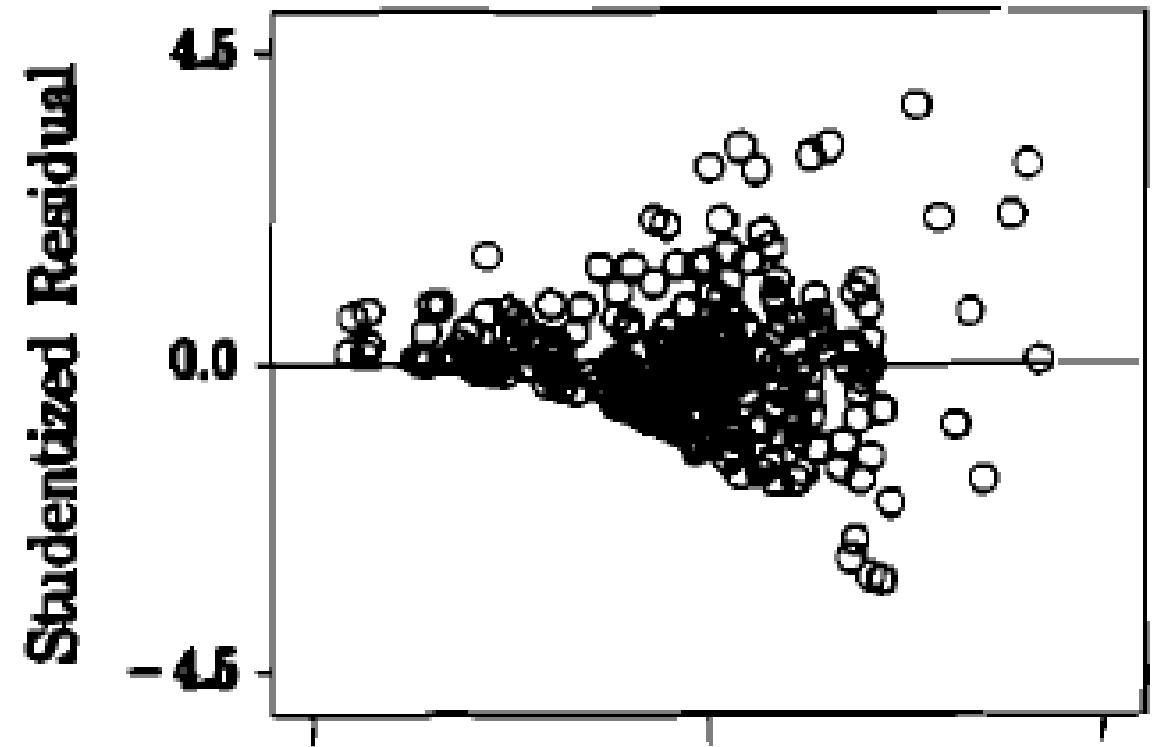
Vlivná pozorování

- 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



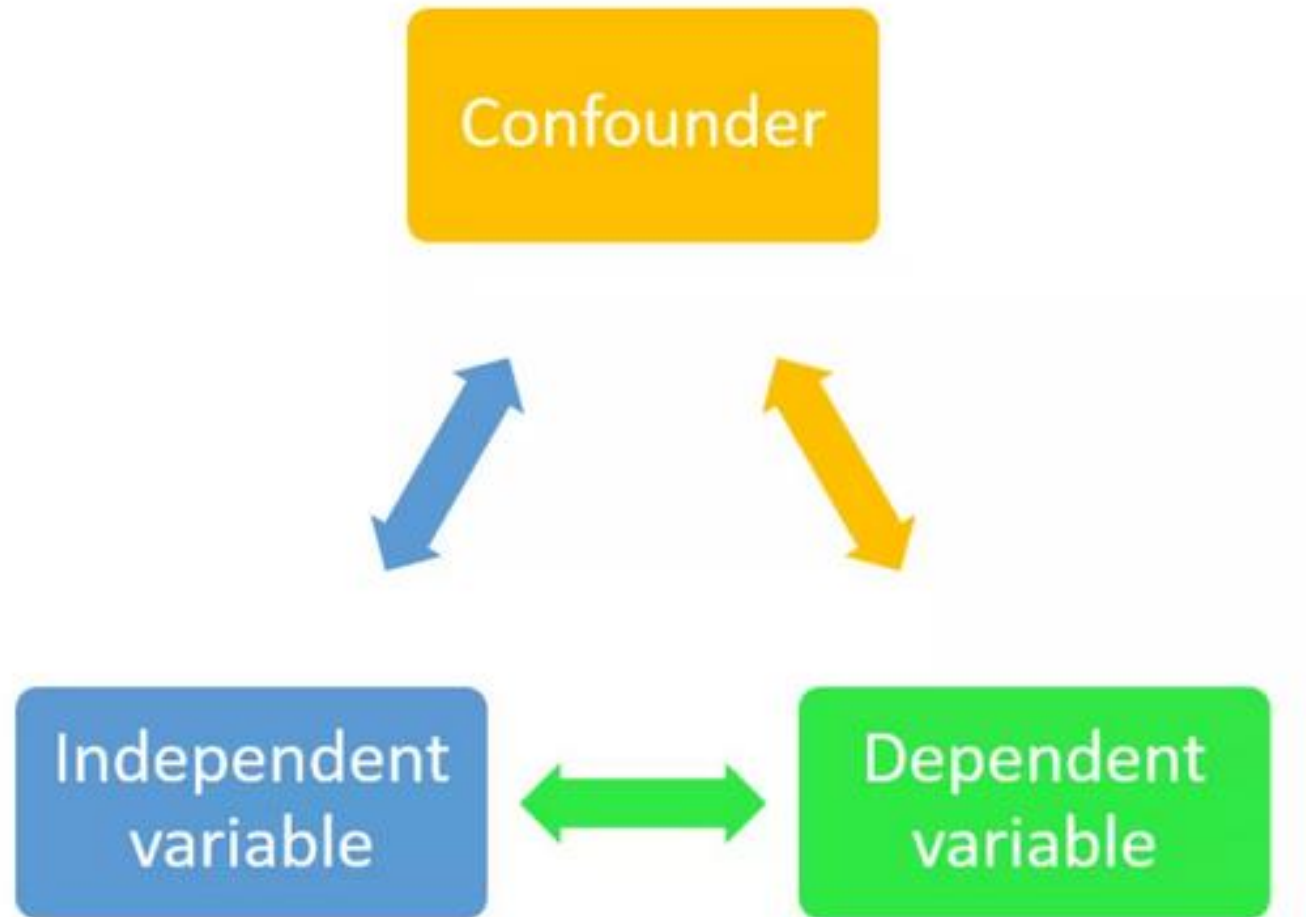
Vzájemná nezávislost chyb (a tedy i pozorování)

- 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."

A pozor: Zkreslení kvůli chybějící proměnné!



Vícenásobná regresní analýza

Základní cíl kauzálního výzkumu

- Identifikace vztahu mezi nezávislou proměnnou X a závislou proměnnou Y , tj.

$$Y \leftarrow X$$

- V jednoduché regresi jsme pro to použili rovnici $Y = a + bX + \epsilon$
- Ale co když existuje Z , které je spojené s Y i X ?

Izolace efektu X: experiment

- Experimentální design: **izolace všech potenciálně relevantních prediktorů**, které mohou být zároveň korelované s nezávislou proměnnou
- Jak? Náhodným výběrem do experimentální (treatment) a kontrolní skupiny
- Zjistíme **čistý efekt** nezávislé proměnné

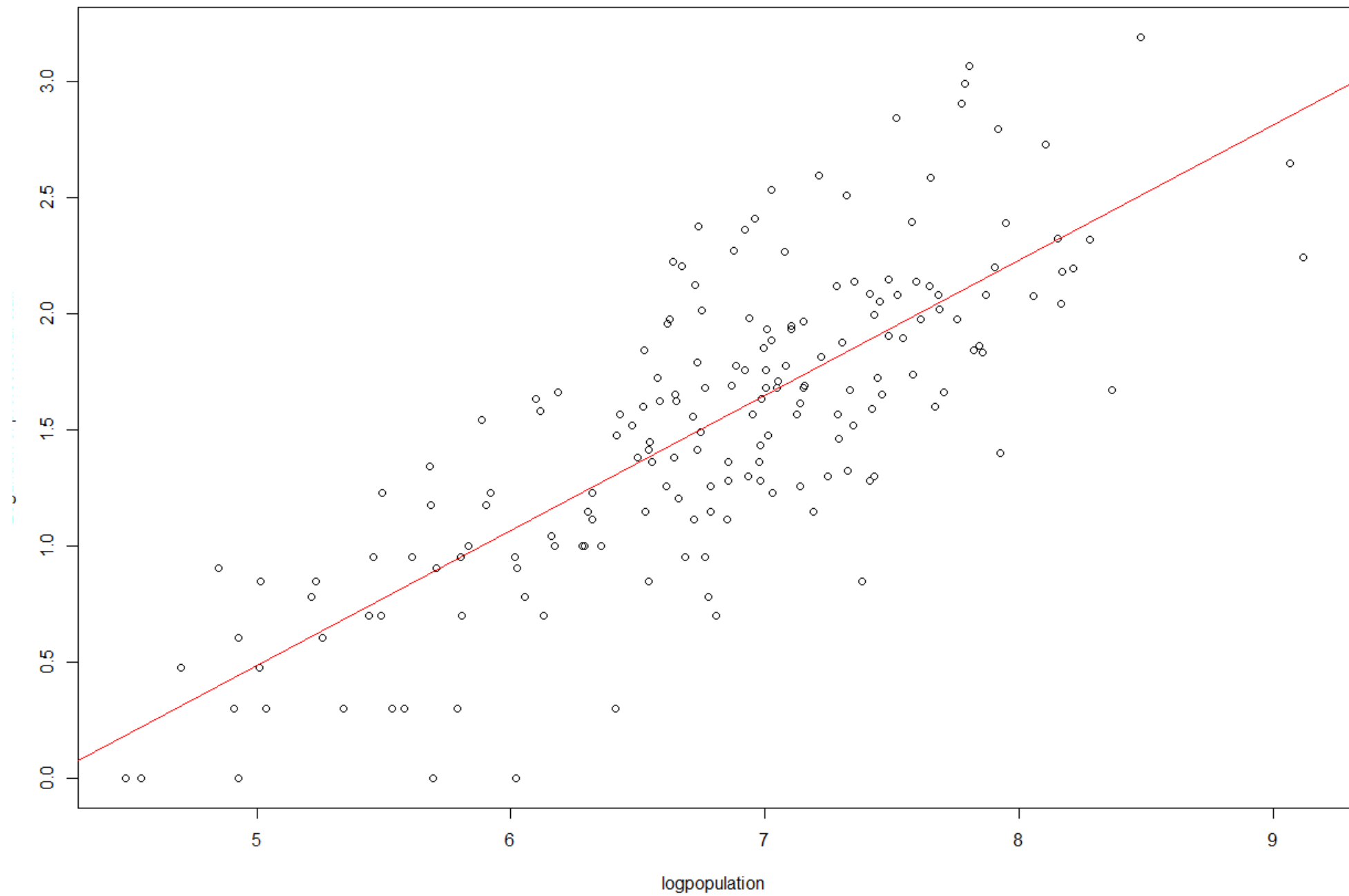


Izolace efektu X, když experiment *není možný*

- Co když experiment nemůžeme udělat?
- Pak musíme všechny potenciálně relevantní prediktory **změřit a zahrnout do analýzy** – „kontrolovat“ je
- Zjistíme **čistý efekt** nezávislé proměnné



Staff, given by GNI (all logs)



Vícenásobná regrese zahrnuje a tak kontroluje další prediktory

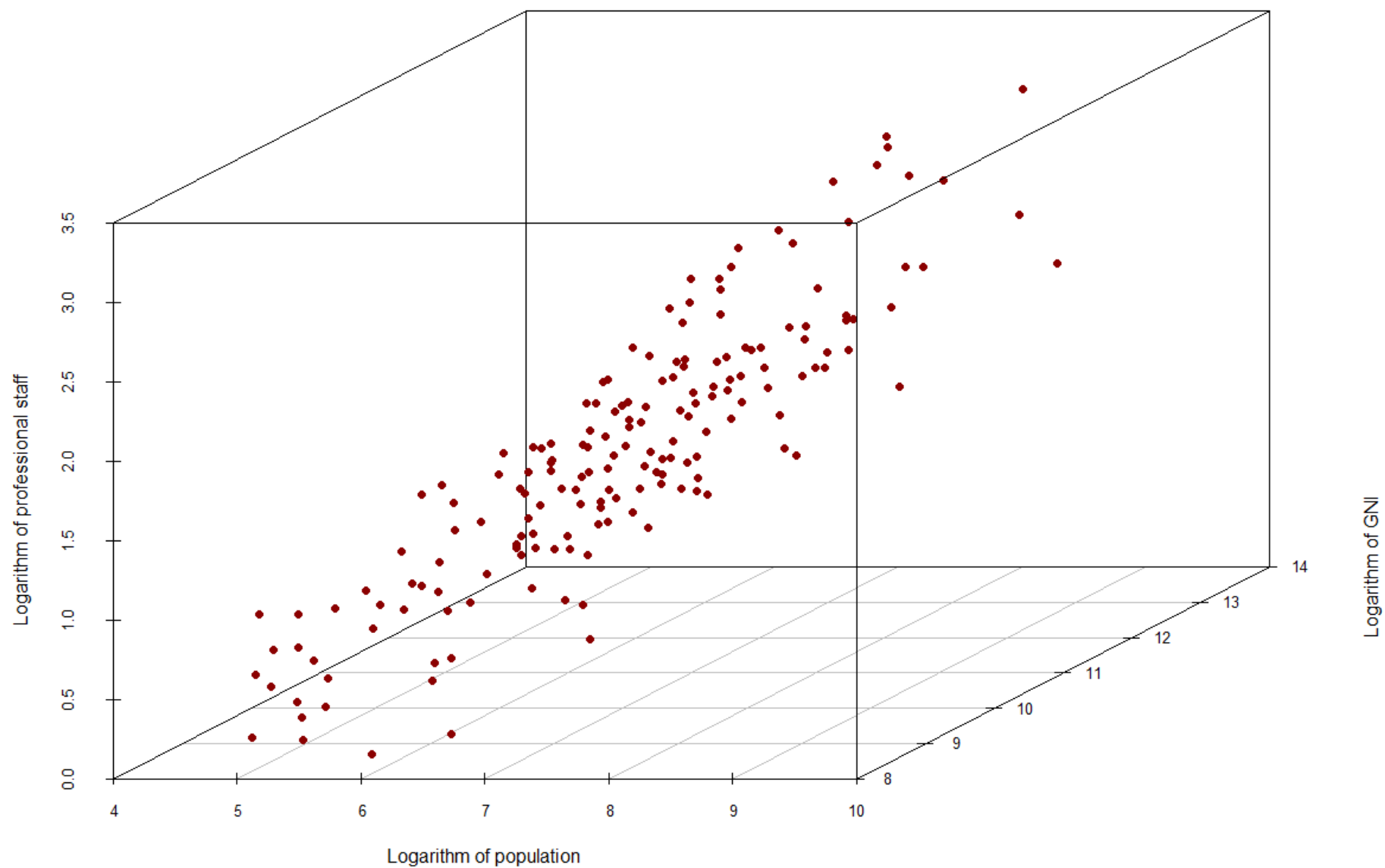
$$Y = a + bX_1 + e$$

$$Y = a + b_1X_1 + b_2X_2 + e$$

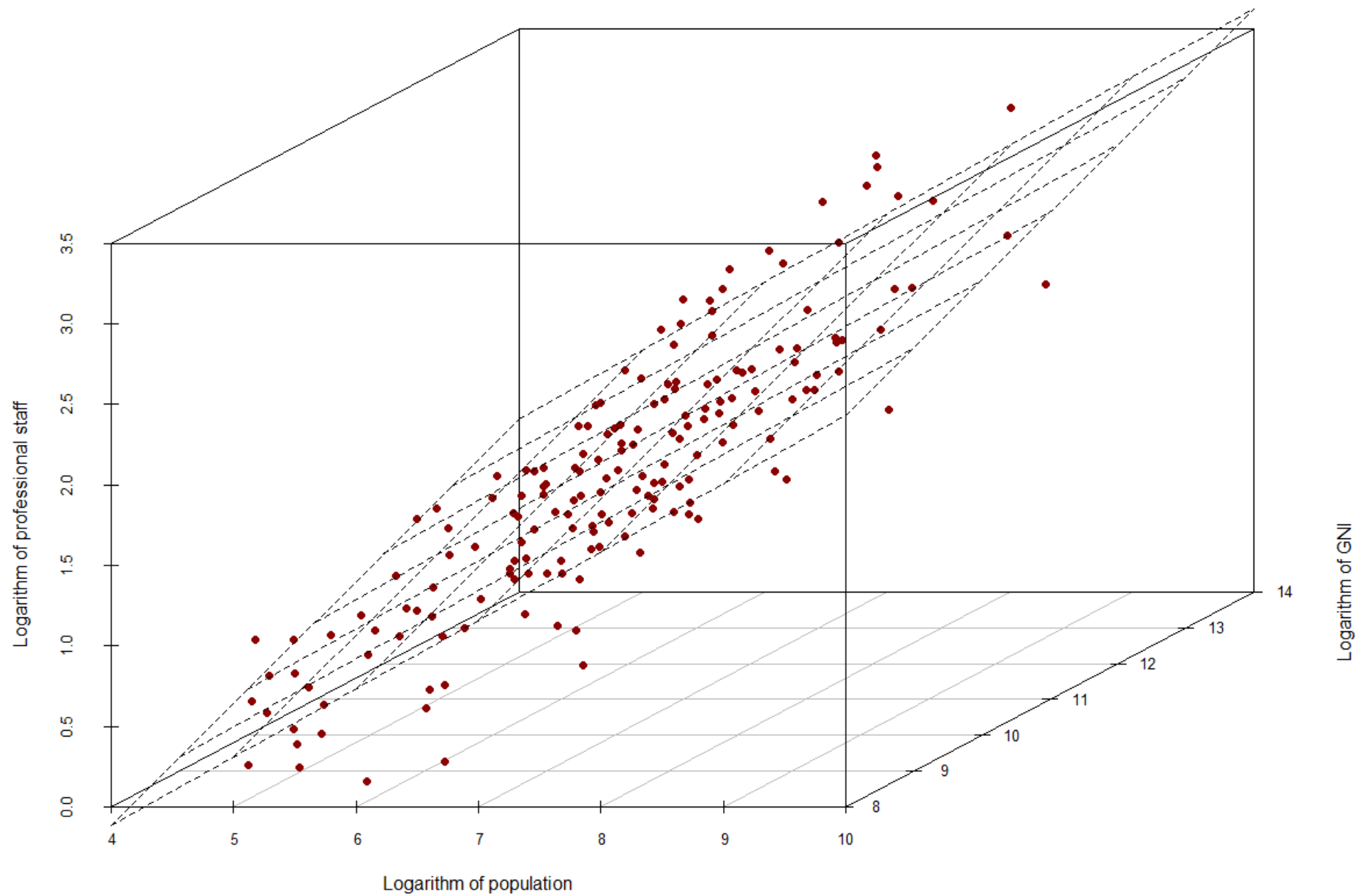
$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

- Y = bezpečnost státu (absence útoku)
- X_1 = síla armády
- X_2 = geografická poloha
- $X_3 = \dots$

Staff, given by GNI and population



Staff, given by GNI and population



Vícenásobná regrese v Excelu

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.854139952					
R Square	0.729555058					
Adjusted R Square	0.725772612					
Standard Error	0.362062564					
Observations	146					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	50.5688	25.2844	192.879136	2.5E-41	
Residual	143	18.7458	0.13109			
Total	145	69.3146				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.260669431	0.03596	90.6872	2.947E-128	3.1896	3.331
governmenteffectiveness	0.612606875	0.03723	16.455	8.36459E-35	0.53902	0.68
polity	0.000163496	0.00568	0.0288	0.977067086	-0.01106	0.011

Je bohatství zemí statisticky spojeno s demokracií a efektivitou vládnutí? (**Data za rok 2000**)

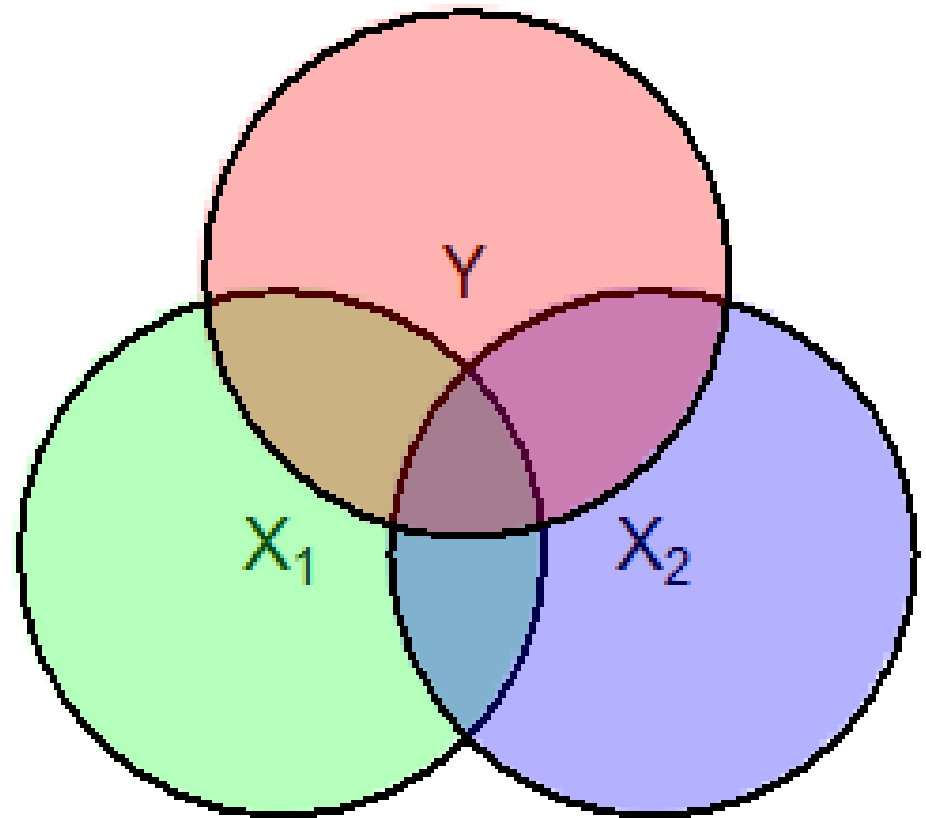
Vícenásobná regrese zahrnuje další prediktory, ale za nějakou cenu...

$$Y = a + bX_1 + e$$

$$Y = a + b_1X_1 + b_2X_2 + e$$

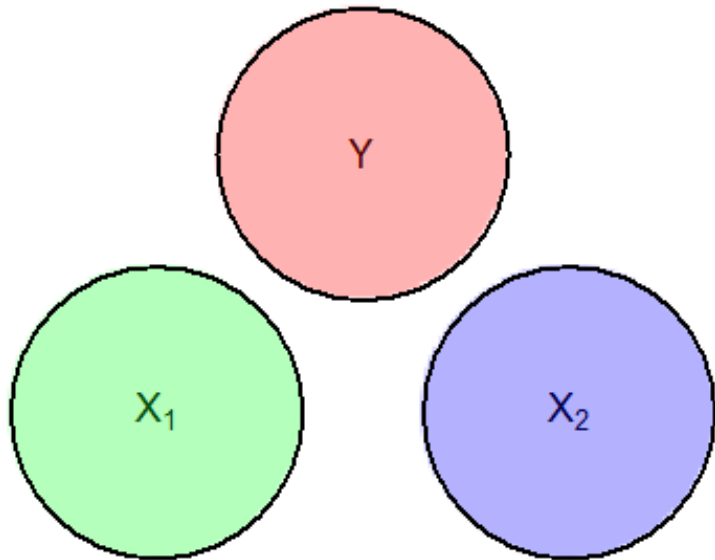
$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

- Y = bezpečnost státu (absence útoku)
- X1 = síla armády
- X2 = geografická poloha
- X3=...

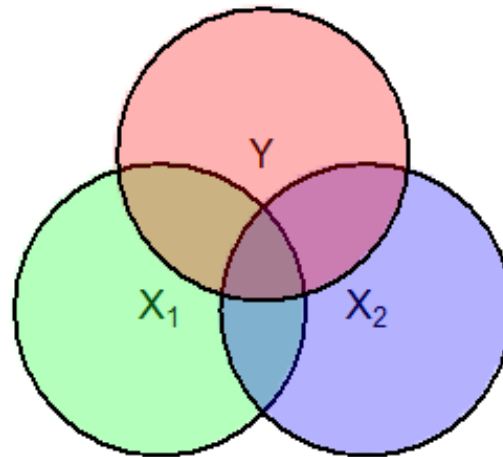


Multikolinearita: další předpoklad, na který si dát pozor (u vícenásobné regrese)

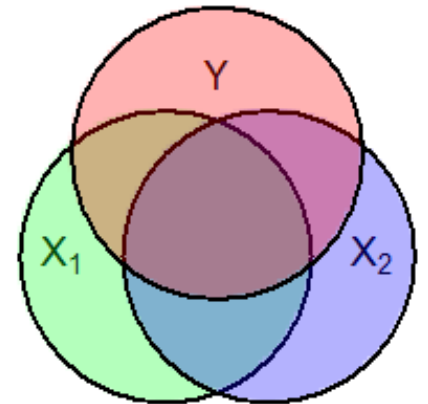
None



Modest



High



Co vysvětluje míru compliance (plnění závazků) transparentnosti států ve WTO?

	Model 1	Model 2
Trade dependence (std)	b1 0.102* (0.053)	0.104* (0.063)
Trade openness (std)	b2 -0.157** (0.064)	-0.135** (0.066)
Administrative capacities (std)	b3 -0.161* (0.085)	0.191** (0.090)
Log (GDPpc) (std)		
Democracy (std)	0.112* (0.066)	0.131** (0.067)
IOs membership (std)	0.179** (0.082)	0.230*** (0.084)
Log (GDP) (std)	0.485 [†] (0.084)	0.446 [†] (0.095)
Membership length (std)		-0.170 (0.110)
No. of obligations (std)		0.075 (0.066)
Constant	0.012 (0.060)	0.027 (0.064)
Observations	111	108
R ²	0.650	0.654
Adjusted R ²	0.630	0.626

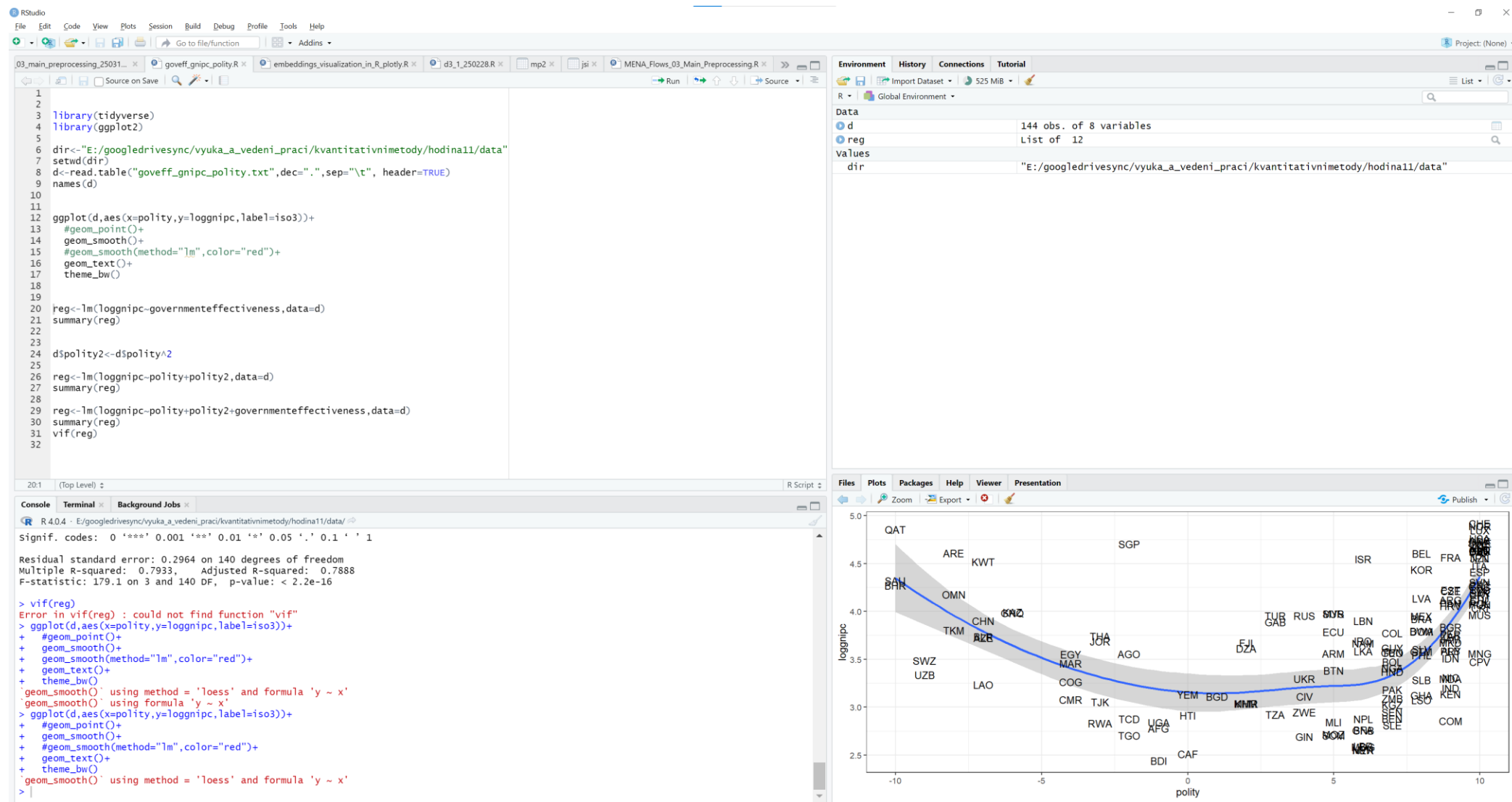
- Dv: compliance
- IVs: administrativní kapacity, závislost na obchodu, demokracie,...

Notes: *p < 0.1; **p < 0.05; ***p < 0.01; [†]p < 0.001; robust standard errors in brackets.

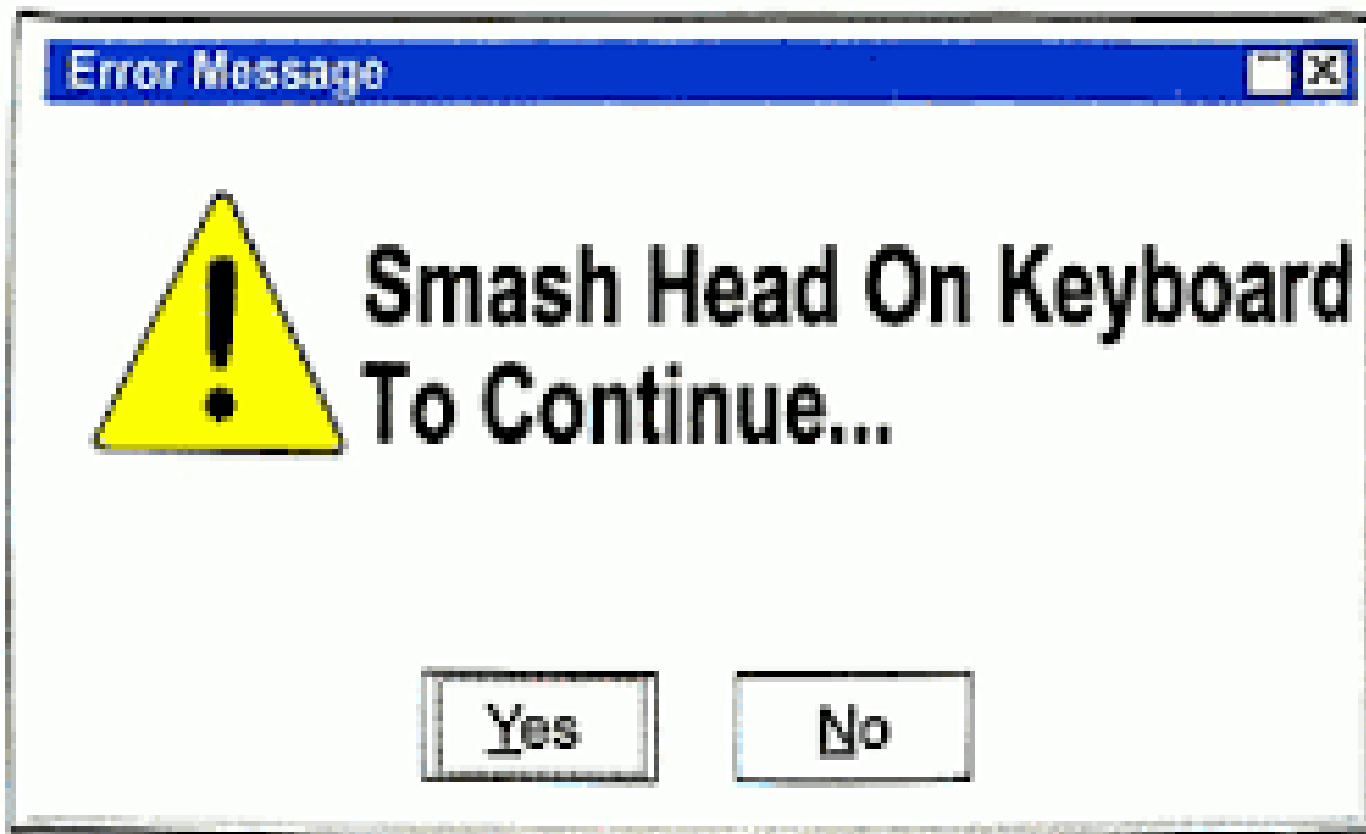
Zdroj: Karlas & Parížek, 2020

Analýza na vyšší úrovni

Od Excelu k R: od klikání ke skriptu



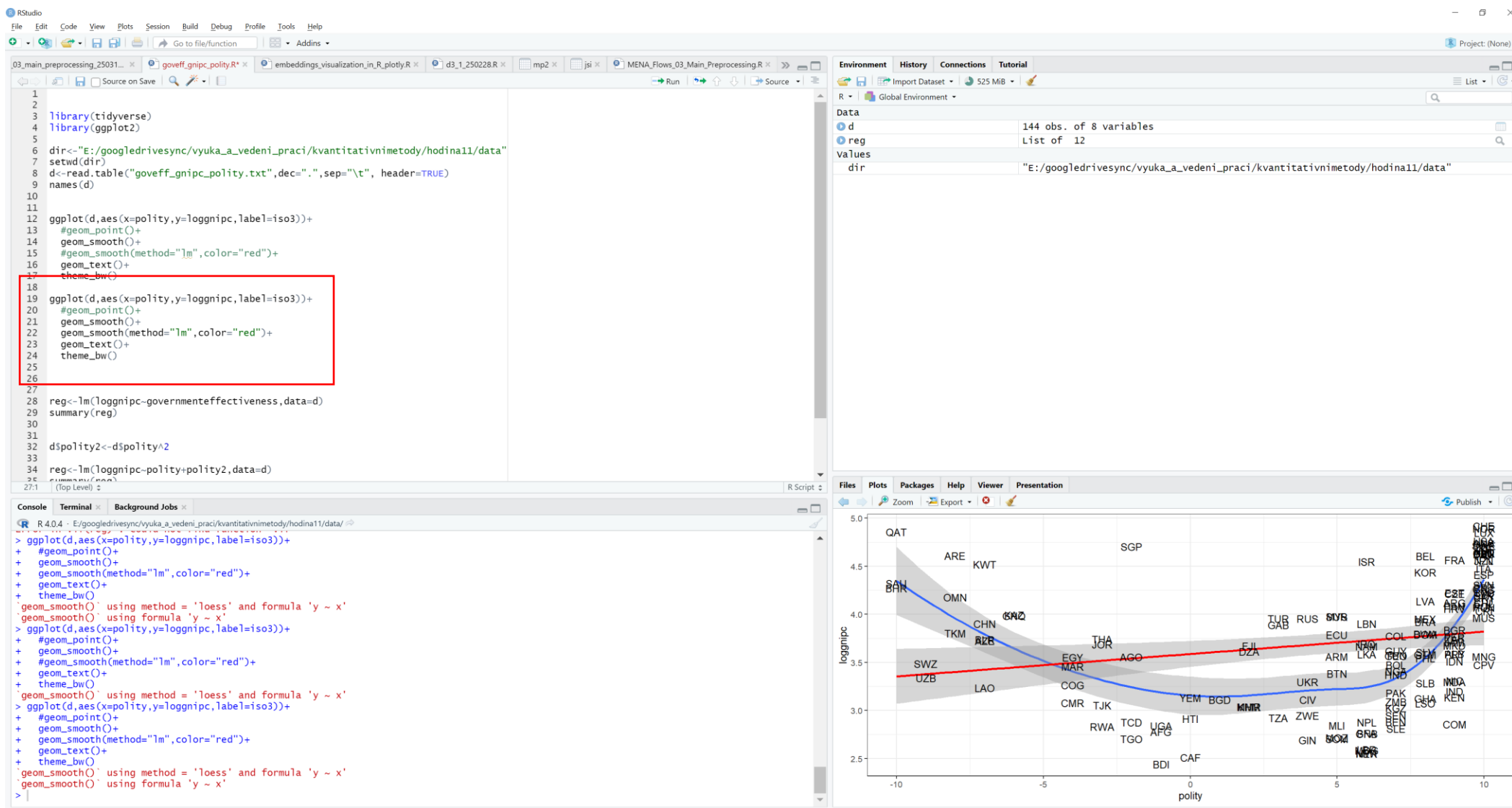
Programovací jazyk a preciznost kódu



©1999-2000 Microsoft Corporation. All rights reserved.

crackmap.com

Efektivita a replikaci/reprodukcce analýzy



Regrese: demokracie (IV) a bohatství (DV)

```
26 reg<-lm(loggnipc~polity,data=d)
27 summary(reg)
28
29
30
31 ggplot(d,aes(x=polity,y=loggnipc,label=iso3))+
32   #geom_point()+
33   geom_smooth()+
34   #geom_smooth(method="lm",color="red")+
35   geom_text()+
36   theme_bw()
37
38
39
40 d$polity2<-d$polity^2
41
42 reg<-lm(loggnipc~polity+polity2,data=d)
43 summary(reg)
```

40:1 (Top Level) ↕

R 4.0.4 · E:/googledrivesync/vyuka_a_vedeni_praci/kvantitativnimetody/hodina11/data/ ↗

residual standard error: 0.319 on 142 degrees of freedom
multiple R-squared: 0.7571, Adjusted R-squared: 0.7554
F-statistic: 442.6 on 1 and 142 DF, p-value: < 2.2e-16

```
· reg<-lm(loggnipc~polity,data=d)
· summary(reg)
```

all:
lm(formula = loggnipc ~ polity, data = d)

residuals:

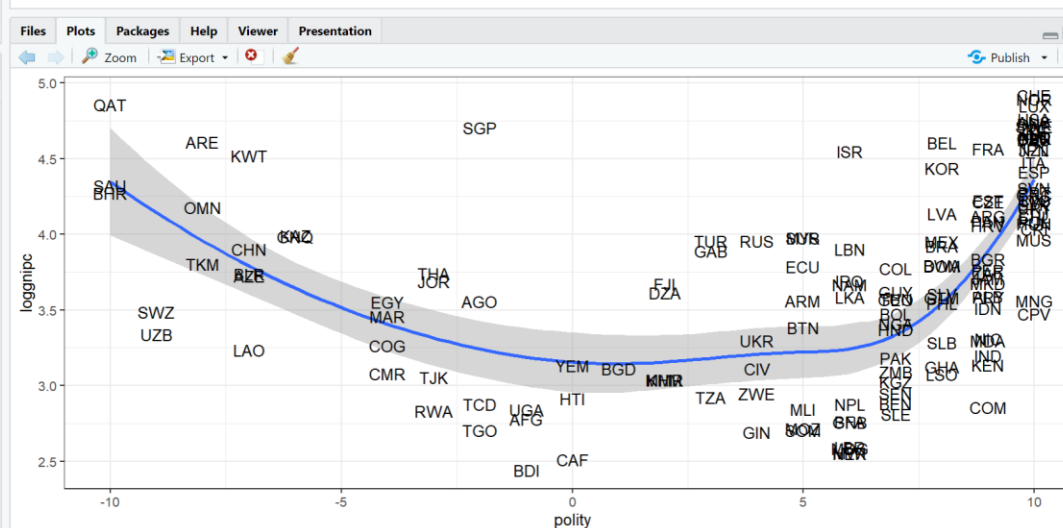
	Min	1Q	Median	3Q	Max
	-1.17593	-0.50014	0.01593	0.42018	1.50323

coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.587351	0.068246	52.565	<2e-16 ***
polity	0.023290	0.009069	2.568	0.0113 *

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual standard error: 0.6327 on 142 degrees of freedom
multiple R-squared: 0.04438, Adjusted R-squared: 0.03765
F-statistic: 6.594 on 1 and 142 DF, p-value: 0.01126



Regrese: demokracie a bohatství nelineárně

```
38
39
40 d$polity2<-d$polity^2
41
42 reg<-lm(loggnipc~polity+polity2,data=d)
43 summary(reg)
44 |
45
```

44:1 (Top Level) ↕

Console

Terminal x

Background Jobs x

R 4.0.4 · E:/googledrivesync/vyuka_a_vedeni_praci/kvantitativnimetody/hodina11/data/ ↗

```
> d$polity2<-d$polity^2
> reg<-lm(loggnipc~polity+polity2,data=d)
> summary(reg)
```

Call:

```
lm(formula = loggnipc ~ polity + polity2, data = d)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.09798	-0.35635	0.00517	0.36829	1.60079

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.013398	0.078051	38.608	<2e-16 ***
polity	-0.017027	0.008072	-2.109	0.0367 *
polity2	0.013538	0.001362	9.940	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4869 on 141 degrees of freedom

Multiple R-squared: 0.4381, Adjusted R-squared: 0.4302

F-statistic: 54.97 on 2 and 141 DF, p-value: < 2.2e-16

Nekonečné možnosti

- Data management (funkce `merge()`, `join()`, `pivot_longer()`, `pivot_wider()`)
- Vizualizace (`ggplot2`, `plotly`)
- Nelineární vztahy, interakční efekty,...
- Všechny nejnovější metody v novém open source knihovnách

Python a další software

The screenshot displays a JupyterLab environment with a notebook titled "Python_Class_8.ipynb". The interface includes a top menu bar with options like File, Edit, View, Insert, Runtime, Tools, and Help. On the right, there are icons for chat, settings, a "Share" button, and a user profile. A left sidebar contains navigation icons for files, search, and code execution. The notebook itself has a title bar with "+ Code" and "+ Text" tabs. The content is organized into sections: "Class 8: Introduction to ML and embeddings", "Your first ML model: a support vector machine (SVM) model to classify news by topic", and "Import usual packages for data handling". The code cells show the installation of scikit-learn and the import of necessary libraries.

Python_Class_8.ipynb ☆ ☁

File Edit View Insert Runtime Tools Help

+ Code + Text

Connect T4 Gemini

↑ ↓ ↻ 🗨️ ✎ 📄 🗑️ ⋮

Class 8: Introduction to ML and embeddings

Your first ML model: a support vector machine (SVM) model to classify news by topic

```
[ ] !pip install scikit-learn # (scientifit kit for ML) is the key package
```

Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (1.5.2)
Requirement already satisfied: numpy>=1.19.5 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.26.4)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (3.5.0)

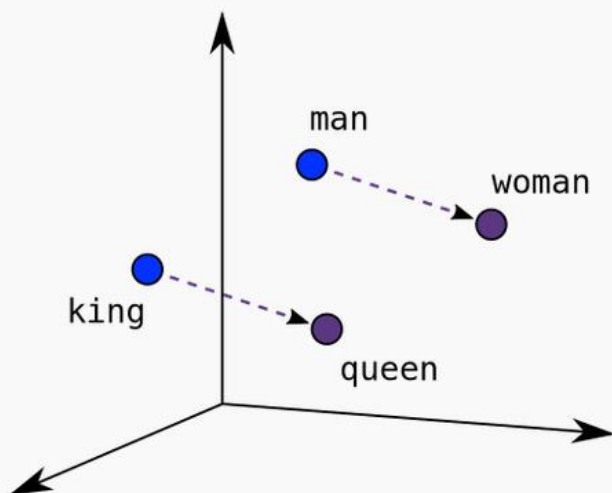
Import usual packages for data handling

```
import os
import pandas as pd
import numpy as np
```

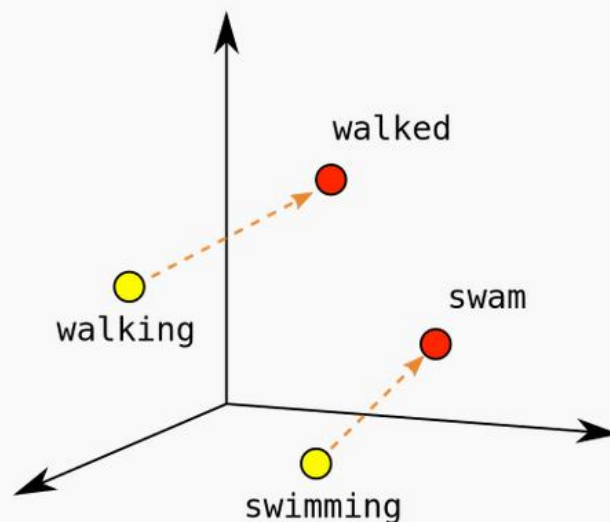
Load specific models and processes from sklearn for data preparation data ML model training

```
[ ] # The model to be used (SVM)
    from sklearn.linear_model import SGDClassifier as SVM # SVM with stochastic gradient descent
```

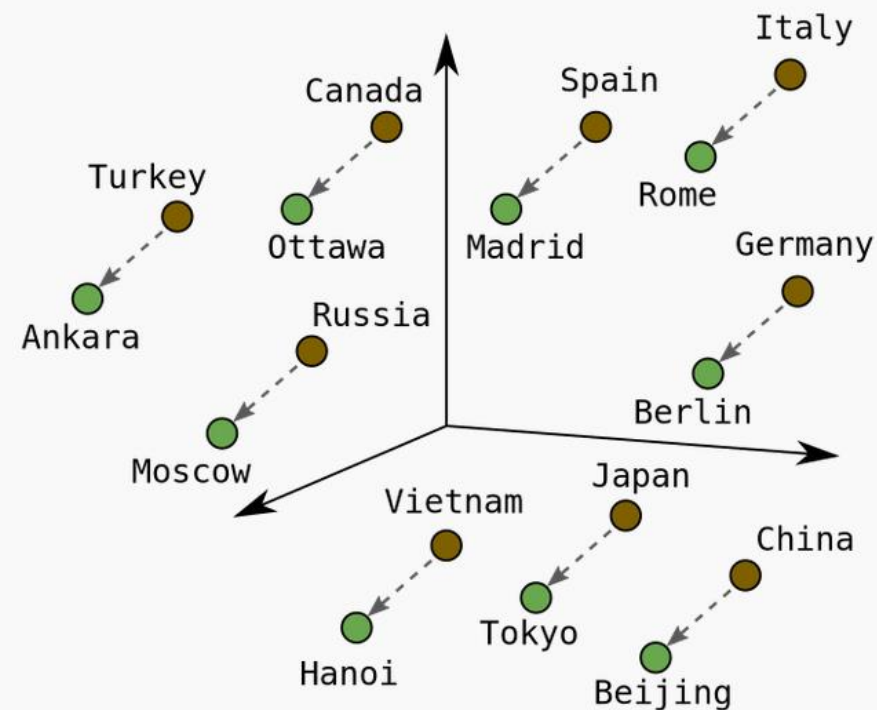
Vektorové vnoření slov



Male-Female

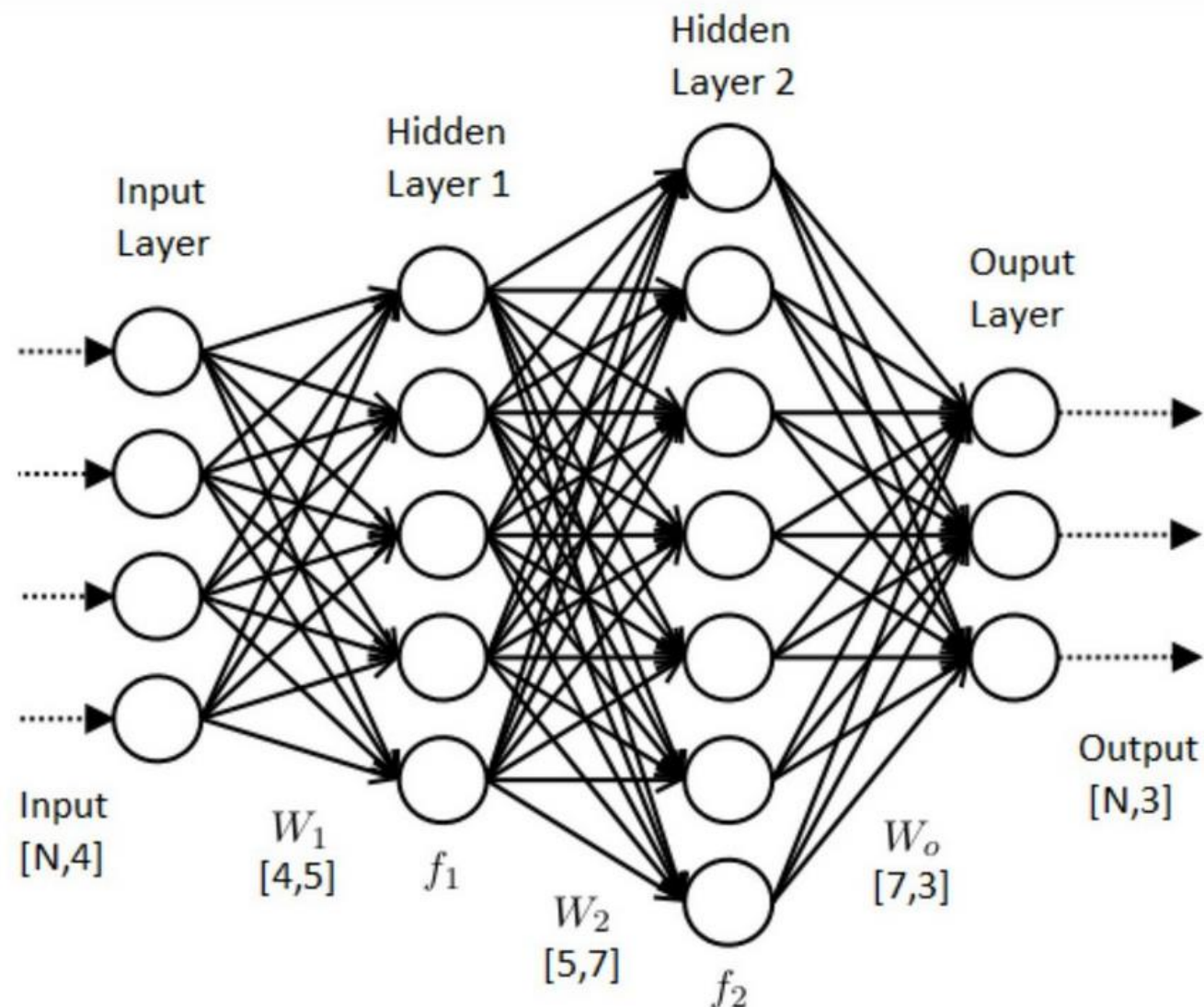


Verb Tense

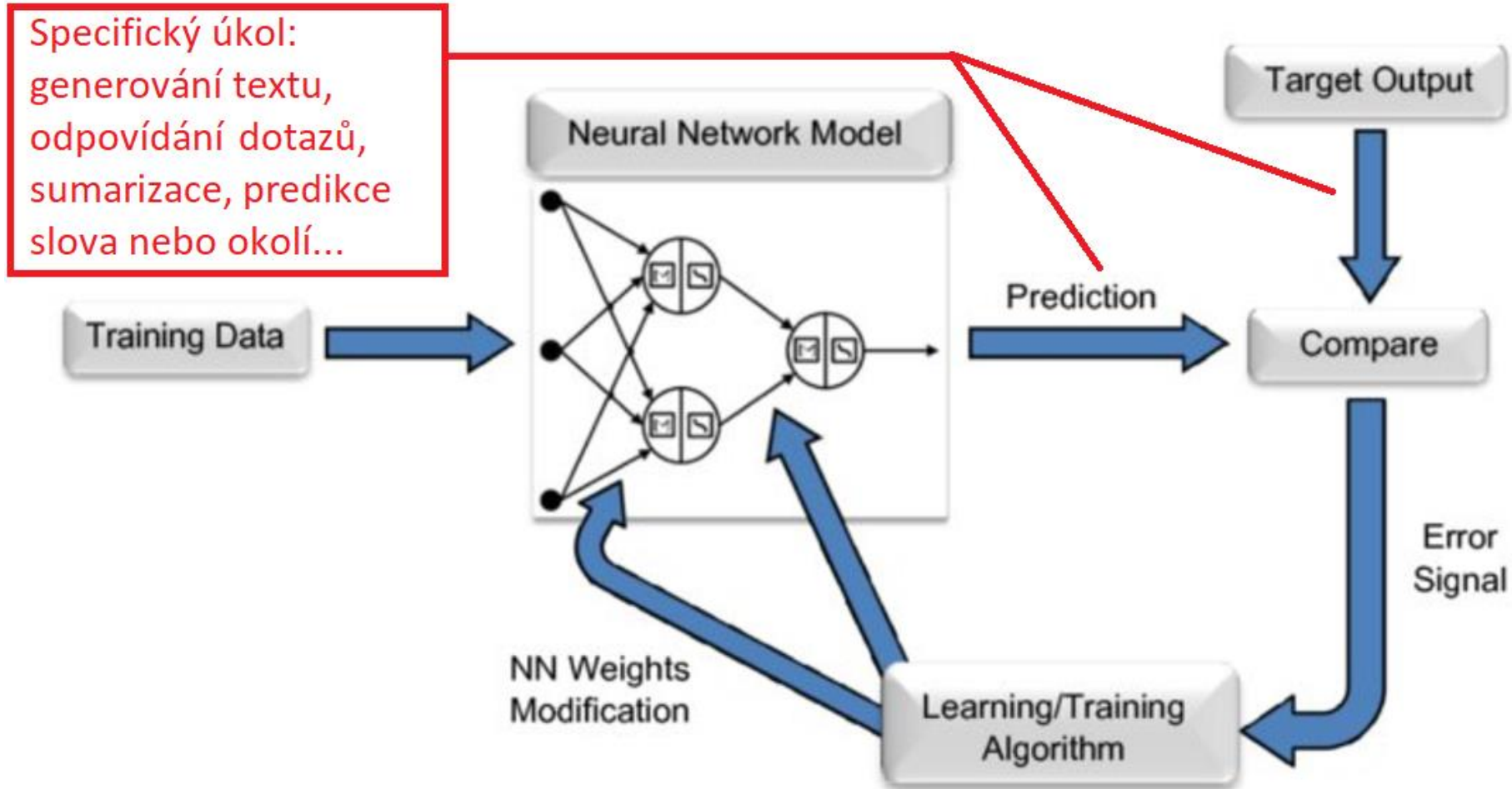


Country-Capital

Umělá neuronová síť:



Trénování umělé neuronové sítě, např. velkého jazykového modelu (LLM)



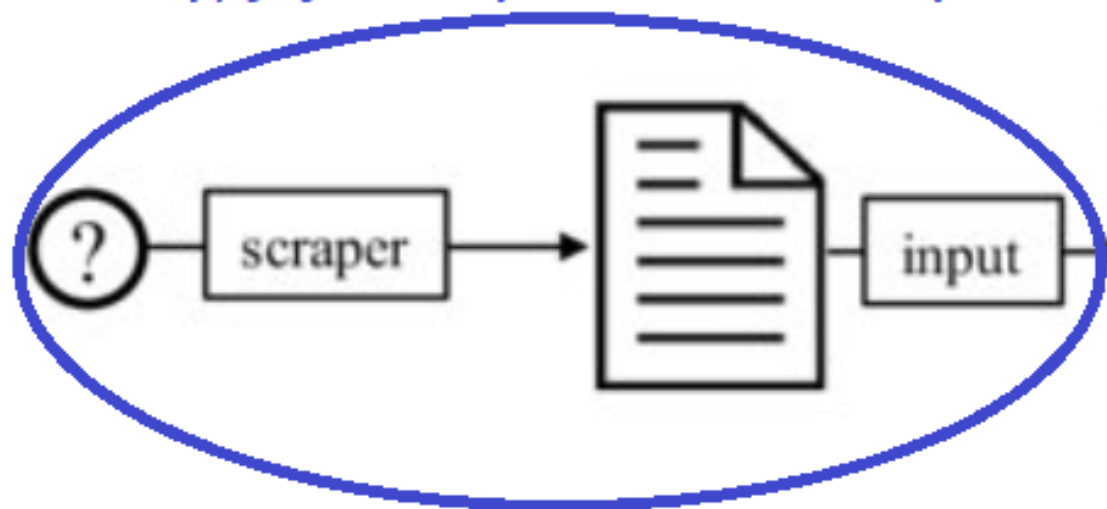
Adaptováno z

https://www.researchgate.net/publication/299390844_The_Development_of_Neural_Network-Based_System_Identification_and_Adaptive_Flight_Control_for_an_Autonomous_Helicopter_System/figures?lo=1

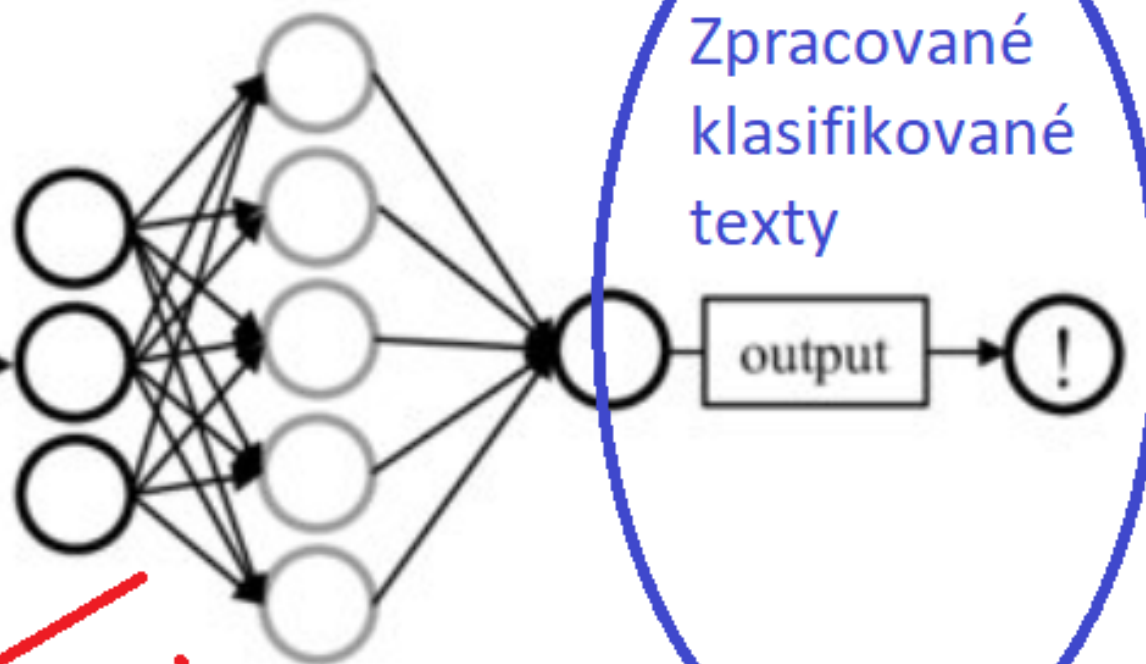
Trénovací data
pro můj úkol



Texty, jejichž význam chci mapovat



původní LLM, který 'dolad'ují



Zpracované
klasifikované
texty

AI na další úrovni



The screenshot displays a Jupyter Notebook interface with a file named `jpm444_class12.ipynb`. The notebook is titled "Chain-of-Thought Prompting". The code is written in Python and demonstrates how to use a language model (llm) to generate text based on a prompt. The prompt is designed to elicit a sentiment analysis response from the model.

```
[ ] initial_messages = [
    {
        "role": "system",
        "content": "You assign texts into categories. "
    },
    {
        "role": "user",
        "content": "text: the pizza tastes terrible\\nWhat sentiment (positive, neutral, or negative) would you assign? Provide some thoughts."
    }
]

sampling_params = SamplingParams(temperature=0.2,
                                max_tokens=256,
                                seed=42)

[ ] outputs = llm.chat(initial_messages, sampling_params)

for output in outputs:
    prompt = output.prompt
    generated_text = output.outputs[0].text
    print(f"Generated text: {generated_text!r}")
```

The output of the notebook shows the progress of the prompt processing and the generated text:

```
Processed prompts: 100%|██████████| 1/1 [00:02<00:00, 2.14s/it, est. speed input: 20.08 toks/s, output: 29.89 toks/s]Generated text: ' The sentiment expressed in the text "the pizza ta
```




Odkazy

- Gerring, John, and Dino Christenson. *Applied Social Science Methodology: An Introductory Guide*. Cambridge, United Kingdom ; New York: Cambridge University Press, 2017.
- Karlas, Jan, and Michal Parizek. “Supply of Policy Information in the World Trade Organization: Cross-National Compliance with One-Time and Regular Notification Obligations, 1995–2014.” *World Trade Review* 19, no. 1 (2020): 30–50. <https://doi.org/10.1017/S1474745618000393>.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. ‘Distributed Representations of Words and Phrases and Their Compositionality’. *arXiv:1310.4546 [Cs, Stat]*, 16 October 2013. <http://arxiv.org/abs/1310.4546>.
- Parizek, Michal, and Jakub Stauber. ‘The Narratives of the War in Ukraine in News Media Worldwide’. *6th Annual COMPTExT Conference 2024, Amsterdam, the Netherlands, May 2–4, 2024*, 2024.