Základy zpracování dat

Peter Spáč

4.5.2016

Analýza dat

- Racionální proces, práce se získanými daty
- Cíle:
  - Najít odpovědi na otázky
  - Potvrdit / zamítnout hypotézy
- Porovnávání proměnných, hledání vzorců a pravidelností
- Vyloučení zdánlivých vztahů

- Panda „Obi-Wan Kenobi“ byla do ZOO dovezena už jako mládě. Postupem času si zvykla na uměle vytvořené prostředí a její zdravotní stav se výrazně zlepšil. V současnosti vyhledává společnost ostatních pand a má kladný vztah i ke svým chovatelům. Pro návštěvníky nepředstavuje žádnou hrozbu, i když přímý kontakt se nedoporučuje.

- Zařazení: živočichové – strunatci – obratlovci – savci – šelmy – medvědovití – ailuropoda

- Velikost: 1,6 – 1,9 m
- Váha: 75 – 160 kg
- Zbarvení: bílá, černá
- Strava: všežravec (bambus, hmyz)
- Délka života: 20 – 30 let

Analýza dat

- Kvantitativní i kvalitativní
- **Společné znaky:**
  - Vytváření závěrů
  - Transparentní procesy a metody
  - Komparace jako ústřední bod
  - Cílem je vyhnout se nesprávným závěrům a zavádějícím vysvětlením

Analýza dat

Kvantitativní	**Kvalitativní**
Standardizovaná, statistické metody	Méně standardizovaná
Analýza začíná až po sběru všech dat	Začíná už před dokončením sběru dat
Převážně zaměřená na testování hypotéz	Může vytvořit nové koncepty a teorie
Nízká úroveň abstrakce	Vysoká úroveň abstrakce

Proměnné

- Charakteristika určité entity
- Reprezentace výseku reality
- Atributy:
  - Označení
  - Hodnoty (minimálně dvě)
- Výsledek procesu operacionalizace

Proměnné - typy

- Nezávislá – předpokládaná příčina
- Závislá – předpokládaný následek

- Přítomnost zprostředkujících proměnných

- Vzájemné postavení proměnných v obou rolích

- Zdánlivé vztahy (na základě „zdravého rozumu“)

Proměnné - typy

- **Nominální:**
  - Jejich hodnoty není možné seřadit ani určit rozdíl mezi nimi
  - Pohlaví, jméno
- **Ordinální (pořadové):**
  - Možné seřazení, ale bez určení rozdílu mezi hodnotami
  - Univerzitní tituly, hodnosti
- **Kardinální (intervalové):**
  - Možné seřazení i určení rozdílu mezi hodnotami
  - Věk, výška, hlasy politické strany, počet teror. útoků

Proměnné - typy

- **Diskrétní:**
  - Limitovaný počet hodnot v rozsahu
  - Počet dětí
  - Dichotomické – pouze dvě hodnoty
- **Spojité:**
  - Nelimitovaný počet hodnot v rozsahu
  - Věk člověka

Případ

- Jednotka analýzy
- Entita, při níž jsou sledovány hodnoty proměnných
- Člověk, region, stát, mezinárodní organizace, extremistická skupina

- Populace (základní soubor) – N
- Vzorek (výběrový soubor) - n

- Volební podpora KDU-ČSL v roce 2013 byla vyšší v oblastech s nadprůměrnou religiozitou, nízkou kupní silou, vysokou nezaměstnaností a početnějším zastoupením obyvatelů se základním a středním vzděláním

- Jak vypadá typický volič KDU-ČSL? (alespoň 3 znaky)

Data

- Informace a údaje získané v rámci sběru
- Nevyhnutelný základ pro fázi analýzy a následného vyhodnocení
- Potřeba „očištění“ dat
- Typy – individuální, agregovaná

Data

- **Individuální:**
  - Mikro-úroveň
  - Práce s výběrovým souborem
  - Údaje za jednotlivce, organizace
  - Simpsonův paradox

- **Agregovaná:**
  - Makro-úroveň
  - Práce s celou populací
  - Údaje za regiony, skupiny populace
  - Možnost **ekologické chyby**

Zabránění teroristickým činům

	2013		2014		Dohromady	
Stát A	39/52	**75 %**	53/136	**38,97 %**	92/188	48,94 %
Stát B	51/ 76	67,11 %	42/112	37,5 %	93/188	**49,47 %**

Získávání dat

- **Existující datové soubory:**
  - Mezinárodní výzkumné programy (ISSP, ESS, EES)
  - Národní výzkumy
  - Oficiální statistiky (demografie, volby, kriminalita)

- **Sběr vlastních dat:**
  - Obsahová analýza
  - Rozhovory
  - Dotazníky

Získávání dat

- **European Social Survey:**
  - http://www.europeansocialsurvey.org/data/
  - 2002, 2004, 2006, 2008, 2010, 2012, 2014
- **European Value Study:**
  - http://www.europeanvaluesstudy.eu/
  - 1991, 1999/2000, 2008
- **European Election Studies:**
  - http://eeshomepage.net/
  - 1979, 1984, 1989, 1994, 1999, 2004, 2009, 2014

Analýza dat

- Analýza „vyčištěných“ dat

- **Popisná (deskriptivní) statistika:**
  - Univariační – analýza vlastností individuálních proměnných

- **Vysvětlující statistika:**
  - Bivariační, multivariační – souvislosti a vztahy mezi 2 a více proměnnými

Popisné statistiky

- Třídění prvního stupně
- Vztažené k jedné proměnné a jejím hodnotám
- Popis proměnných

- Možnosti:
  - Četnosti
  - Střední hodnoty
  - Distribuce

Četnosti

- Absolutní, relativní, validní, kumulativní relativní

Strana	Abs. počet	Podíl	Validní podíl	Kumulat. procento
ČSSD	432	22,09	34,15	34,15
KSČM	251	12,83	19,84	53,99
TOP09	186	9,51	14,7	68,69
ODS	164	8,38	12,96	81,65
Ostatní	232	11,86	18,34	100
Spolu odp.	1265	64,67	100	
Bez odp.	691	35,33		
Spolu	1956	100		

Střední hodnoty

- **Modus:**
  - Nejčastěji se vyskytující hodnota
  - Nominální, ordinální, kardinální
- **Medián:**
  - Hodnota středního prvku
  - Ordinální, kardinální
- **Průměr:**
  - Součet hodnot vydělený jejich počtem
  - Kardinální

Střední hodnoty

- Paralen
- Strepfen
- Tantum verde
- Strepfen
- Strepsils
- Zodac
- Ibalgin
- Algifen

Střední hodnoty

- 1, 1, 2, 3, 7, 8, 9, 12, 15, 18, 25, 41, 43

- Modus = 1
- Medián = 9
- Průměr = 185 / 13 = 14,23

Střední hodnoty

- Modus – méně časté využití
- Průměr:
  - Tradičně používaný ukazatel
  - Extrémní hodnota může výrazně vychýlit jeho výstupy → průměrná mzda
- Pokud jsou všechny hodnoty ve vzorku stejné, modus, medián a průměr jsou totožné

Mzdy v ČR

	Průměrná mzda	Mediánová mzda
1Q 2015	25 306	21 143
2Q 2015	26 287	22 230
3Q 2015	26 072	22 531
4Q 2015	28 152	23 745
2015	26 467	22 412

Bivariační statistiky

- Třídění druhého stupně
- Posuzuje se hodnota dvou proměnných
- Kontingenční tabulky
- Korelace

Kontingenční tabulky

		EU -	EU +	Spolu
Strana A	Počet	305	570	875
	Podíl	34,9 %	65,1 %	100 %
Strana B	Počet	105	85	190
	Podíl	55,3 %	44,7 %	100 %
Strana C	Počet	544	140	684
	Podíl	79,5 %	20,5 %	100 %
Strana D	Počet	460	56	516
	Podíl	89,1 %	10,9 %	100 %
Spolu	Počet	1 414	851	2 265
	Podíl	62,4 %	37,6 %	100 %

		EU -	EU +	Spolu
Strana A	Počet	305	570	875
	Podíl	34,9 %	65,1 %	100 %
	St. Res.	-21,5	21,5	
Strana B	Počet	105	85	190
	Podíl	55,3 %	44,7 %	100 %
	St. Res.	-2,1	2,1	
Strana C	Počet	544	140	684
	Podíl	79,5 %	20,5 %	100 %
	St. Res.	11,1	-11,1	
Strana D	Počet	460	56	516
	Podíl	89,1 %	10,9 %	100 %
	St. Res.	14,3	-14,3	
Spolu	Počet	1 414	851	2 265
	Podíl	62,4 %	37,6 %	100 %

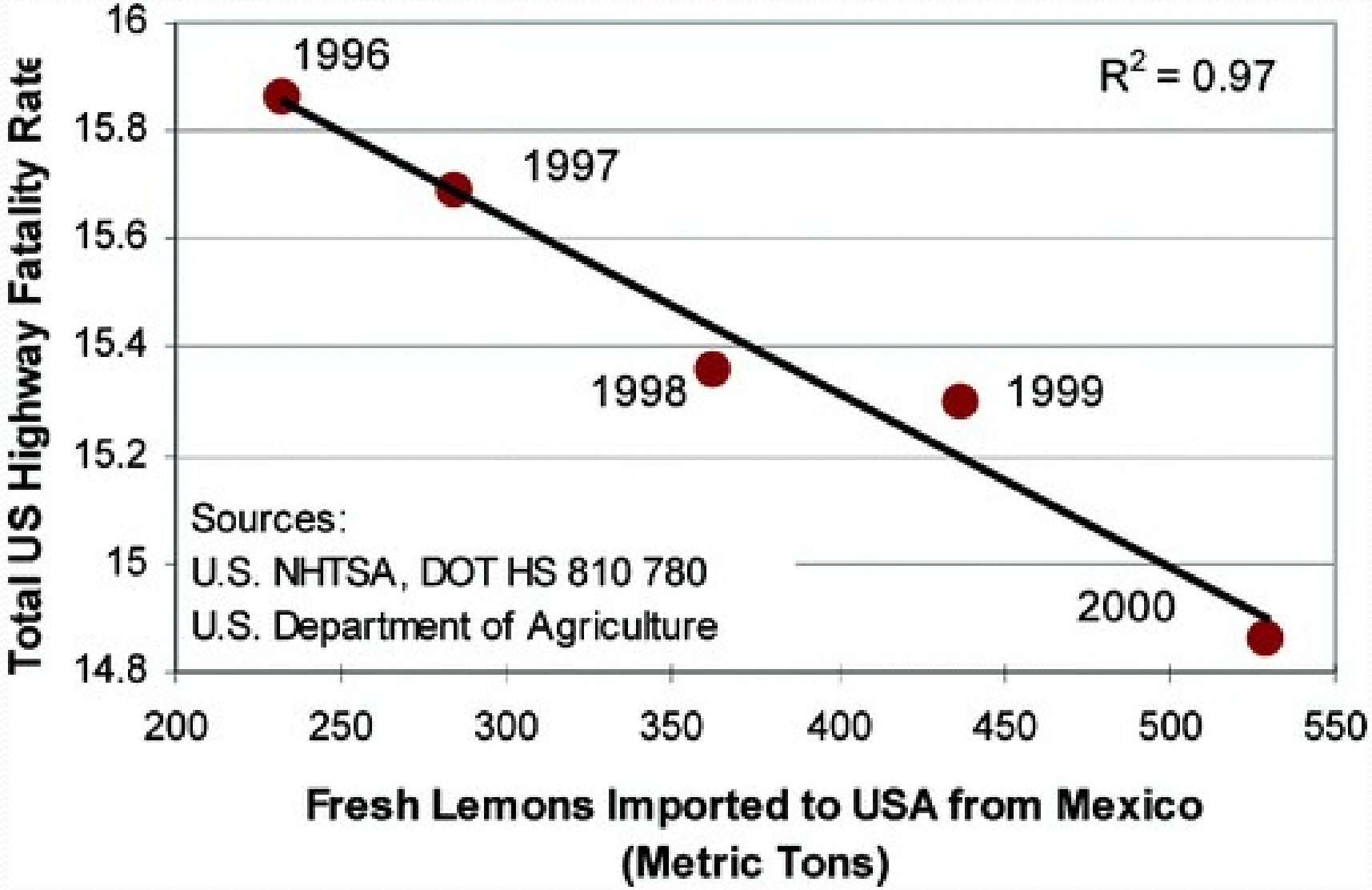
Korelace (korelační koeficient)

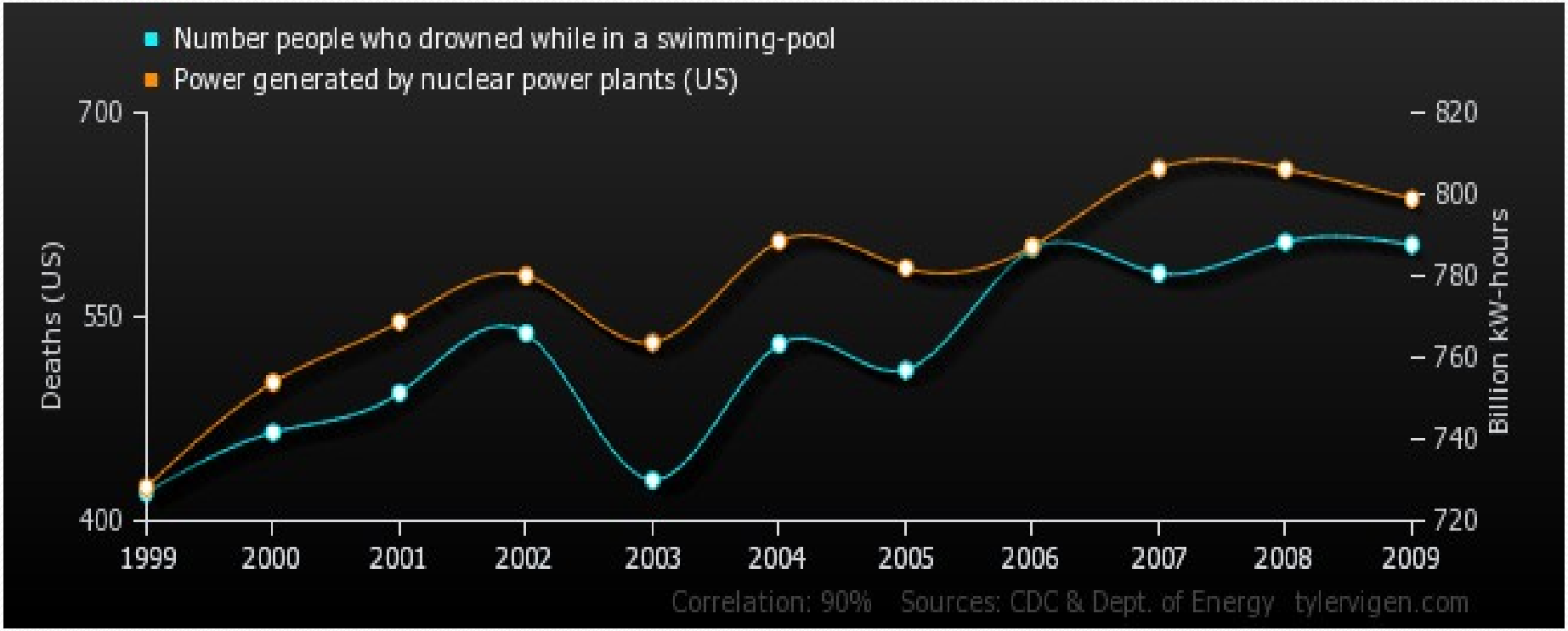
- Odhalení statistické **souvislosti** mezi proměnnými
- Symetričnost – není podstatné, v jakém „pořadí“ jsou proměnné do výpočtu zadávány
- Pozitivní korelace – se vzrůstající hodnotou jedné proměnné vzrůstají hodnoty druhé proměnné a naopak
- Negativní korelace – se vzrůstající hodnotou jedné proměnné klesají hodnoty druhé proměnné a naopak
- Korelace **není důkazem kauzality**

Korelace (korelační koeficient)

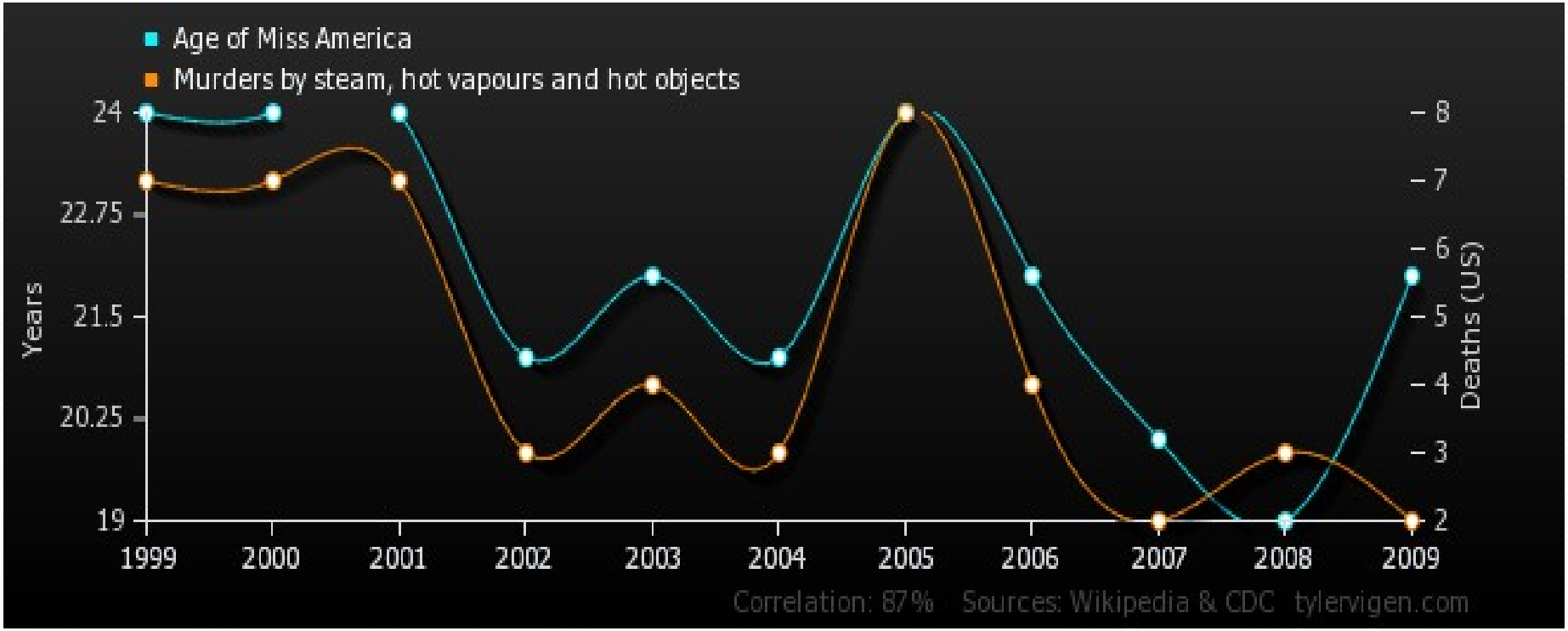
- Rozpětí hodnot -1 až +1
  - - 1 → perfektní negativní korelace
  - 0 → úplná nezávislost
  - + 1 → perfektní pozitivní korelace

- 0,1 – slabá souvislost
- 0,3 – střední souvislost
- 0,5 – silná souvislost


Total US Highway Fatality Rate
16
15.8
15.6
15.4
15.2
15
14.8
1996
1997
1998
1999
2000
$R^2 = 0.97$
Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture
200
250
300
350
400
450
500
550
Fresh Lemons Imported to USA from Mexico
(Metric Tons)



Korelační koeficient: 0,901



Korelační koeficient: 0,870

Multivariační analýza - regrese

- Analýza vztahu více proměnných
- Zjišťuje, jak se při změně jedné nezávislé proměnné za současné neměnnosti ostatních nezávislých proměnných mění hodnota závislé proměnné
- Prokazuje:
  - Vliv nezávislých proměnných na závislou proměnnou
  - Vzájemnou sílu nezávislých proměnných
- Závislá proměnná:
  - Kardinální – lineární regrese
  - Kategorická - logistická regrese

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,405	,164	,163	5,50964

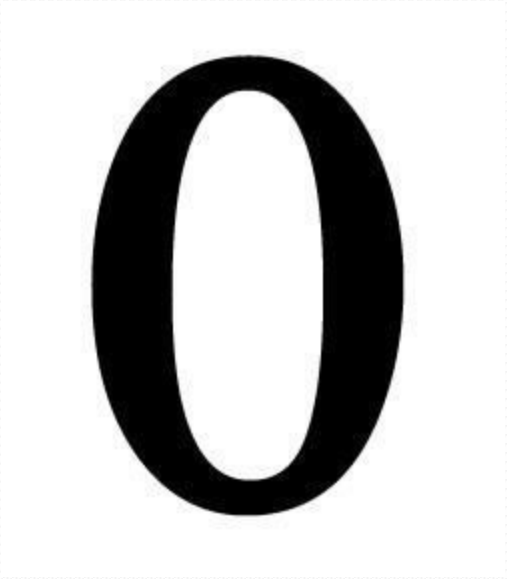
		B	**Beta**	**Sig.**
	(Constant)	3,213		,000
Pohlaví	Žena	-0,259	-0,017	,224
Titul	Mgr.	-0,039	-0,003	,855
	Doktor	0,657	0,045	,010
	Profesor	3,594	0,106	,000
Povolání	Politik	4,911	0,226	,000
	Lok. politik	0,192	0,008	,574
	Podnikatel	-0,12	-0,006	,676
Pořadí		-0,031	-0,223	,000

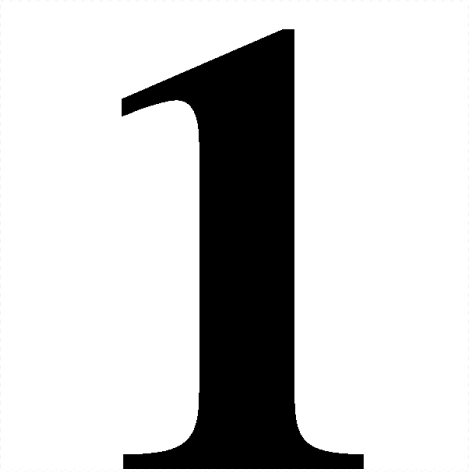
Lineární regrese

- **Index determinace (R^2):**
  - Uvádí, jaký podíl výskytu závislé proměnné je vysvětlený námi použitými nezávislými proměnnými
- **Nestandardizovaný koeficient (b):**
  - Uvádí o kolik se změní hodnota závislé proměnné při změně nezávislé proměnné (při zachování ostatních nezávislých) o jednotku
- **Standardizovaný koeficient (Beta):**
  - Uvádí sílu nezávislé proměnné vůči ostatním nezávislým proměnným

Multivariační analýza - regrese

- Dummy proměnné:
  - Pouze hodnoty **nula a jedna**
  - Typický výsledek úpravy proměnných, které by jinak nemohly být zapracovány (např. povolání, titul)





Dummy proměnné

- Univerzitní tituly – bez, Bc., Mgr., Ing., Ph.D., …
- Pro každou kategorii se vytvoří samostatná dummy proměnná (např. Mgr.)
- Hodnoty těchto proměnných jsou pouze nula anebo jedna (Mgr. = 1, všechno ostatní = 0)
- Do výpočtu regrese se zařadí všechny mimo jedné (němá proměnná)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,405	,164	,163	5,50964

		B	Beta	Sig.
	(Constant)	3,213		,000
Pohlaví	Žena	-0,259	-0,017	,224
Titul	Mgr.	-0,039	-0,003	,855
	Doktor	0,657	0,045	,010
	Profesor	3,594	0,106	,000
Povolání	Politik	4,911	0,226	,000
	Lok. politik	0,192	0,008	,574
	Podnikatel	-0,12	-0,006	,676
Pořadí		-0,031	-0,223	,000

Kvalitativní analýza

- Osobní interpretace dat (neměla by být příliš osobní)
- Potřeba kritické rozvahy
- Tvorba kategorií, porovnávání, hledání podobností a kontrastů
- Otevřenost alternativním vysvětlením
- Není jediná „správná“ cesta

Kvalitativní analýza

- Tipy:
  - Ujistit se, že kategorie dat jsou konzistentní s cíli práce
  - Cíle práce jako kritérium pro tvorbu kategorií
  - Opatrné kódování dat a jejich řazení do kategorií
  - Nedávat data násilím pouze do jedné kategorie

Kde zpracovat data

- Výrazná pomoc výpočetní techniky
- Software zaměřený na kvantitativní i kvalitativní analýzu
- Excel, SPSS, STATA, R, ATLAS/ti,…
- Interpretace je vždy záležitostí výzkumníka