Lineární regrese

POLb1139

Peter Spáč

Regresní analýza

- Skupina technik se stejným cílem
- Identifikace efektů jedné nebo vícera NP na ZP
- Co umožňuje:
  - Identifikovat efekt každé nezávislé proměnné
  - Kontrolovat efekty jiných nezávislých/kontrolních proměnných
  - Odhadovat hodnoty ZP na základě konkrétních hodnot NP

Která regrese?

- Výběr závisí od podoby závislé proměnné
- Lineární (OLS) regrese
  - Kardinální proměnná
- Logistická regrese
  - Binární ZP (0/1) – binární logistická regrese
  - Nominální ZP s > 2 hodnotami (0/1/2/3) – multinomiální logistická regrese
- Nejedná se o konečný výčet

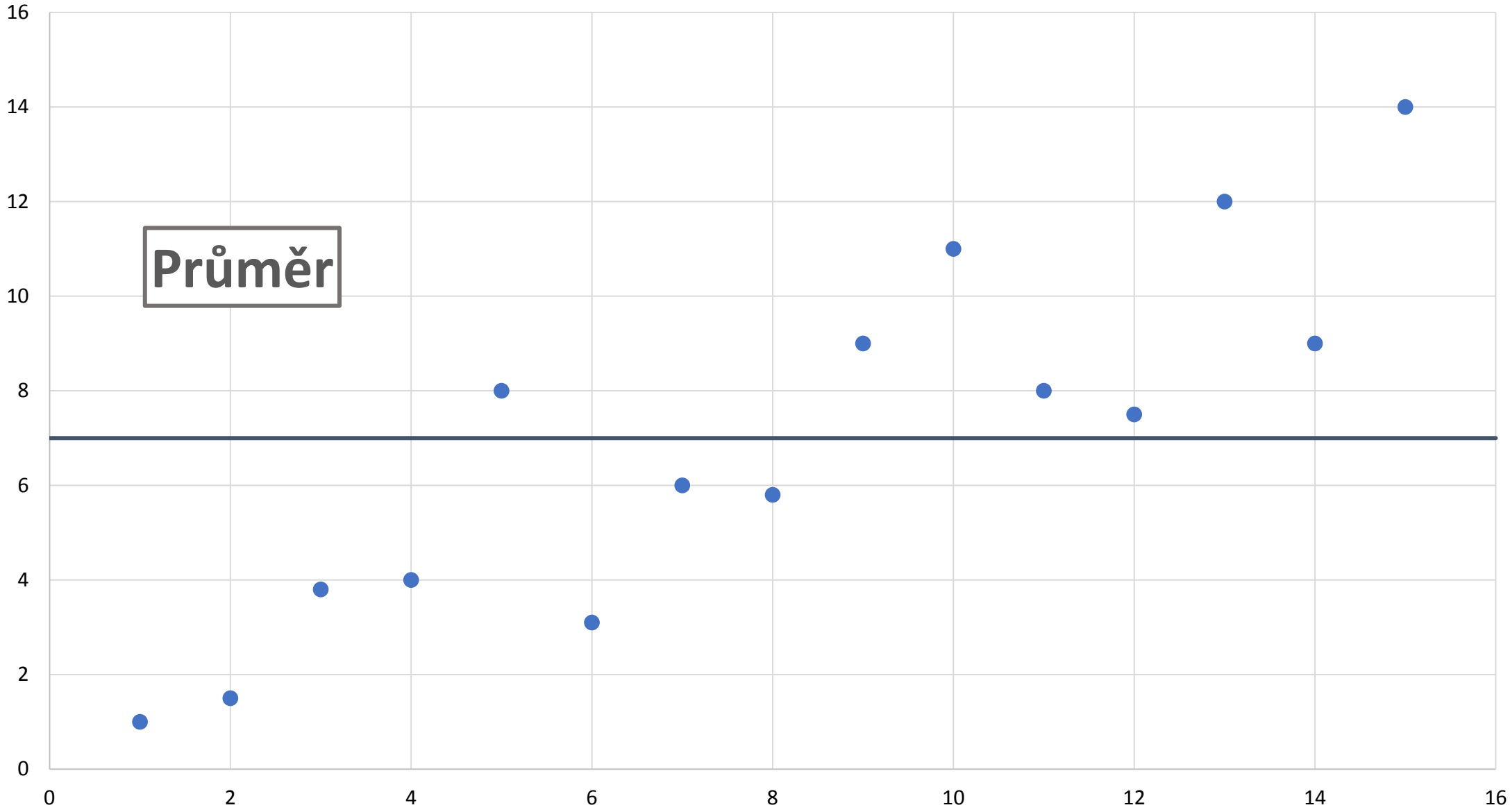
Příklady

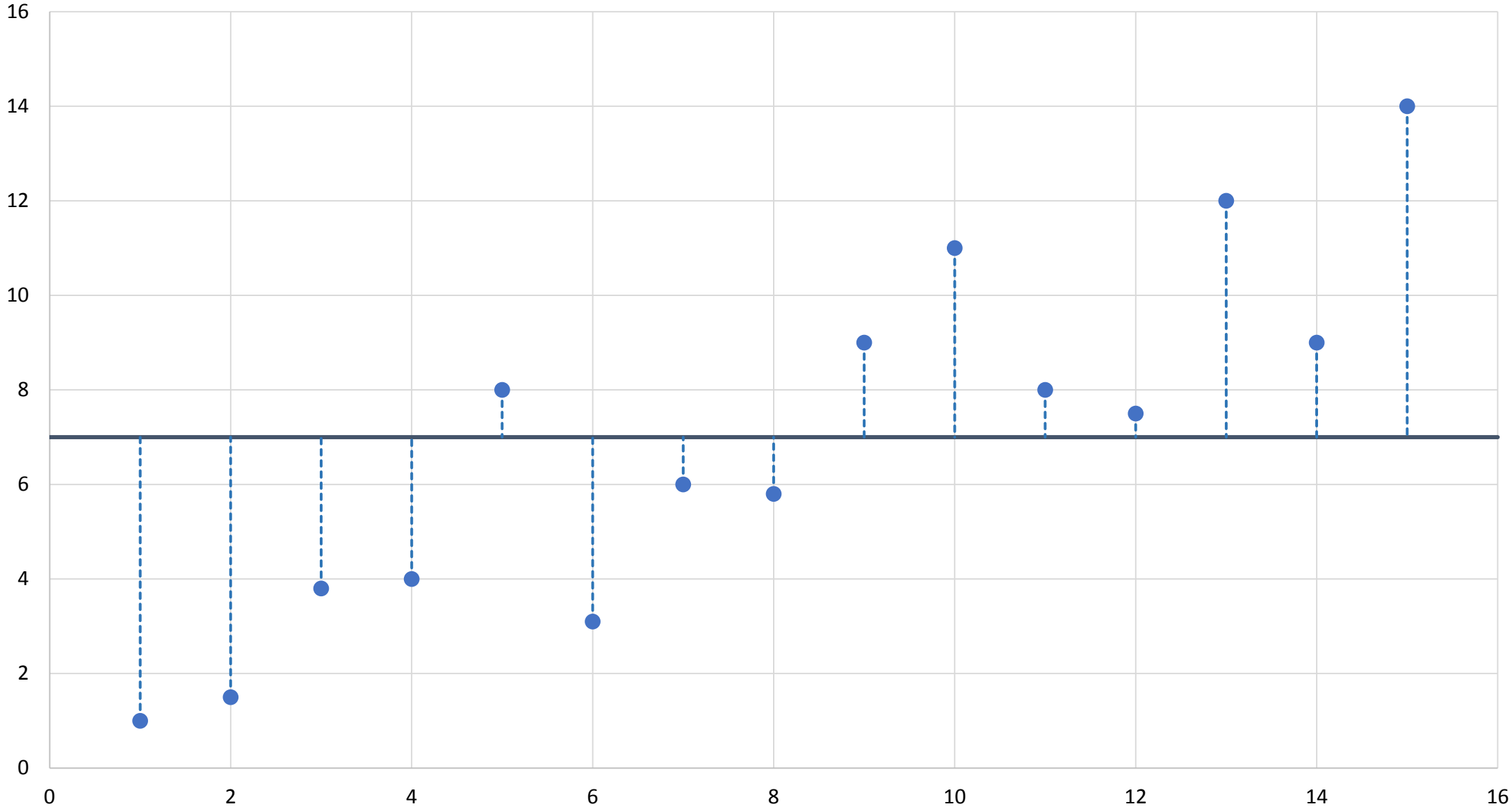
- OLS regrese:
  - Jak věk, pohlaví a vzdělání ovlivňují příjem lidí?
  - Zvyšuje účast na přednáškách % bodů dosažených na zkouškovém testu?

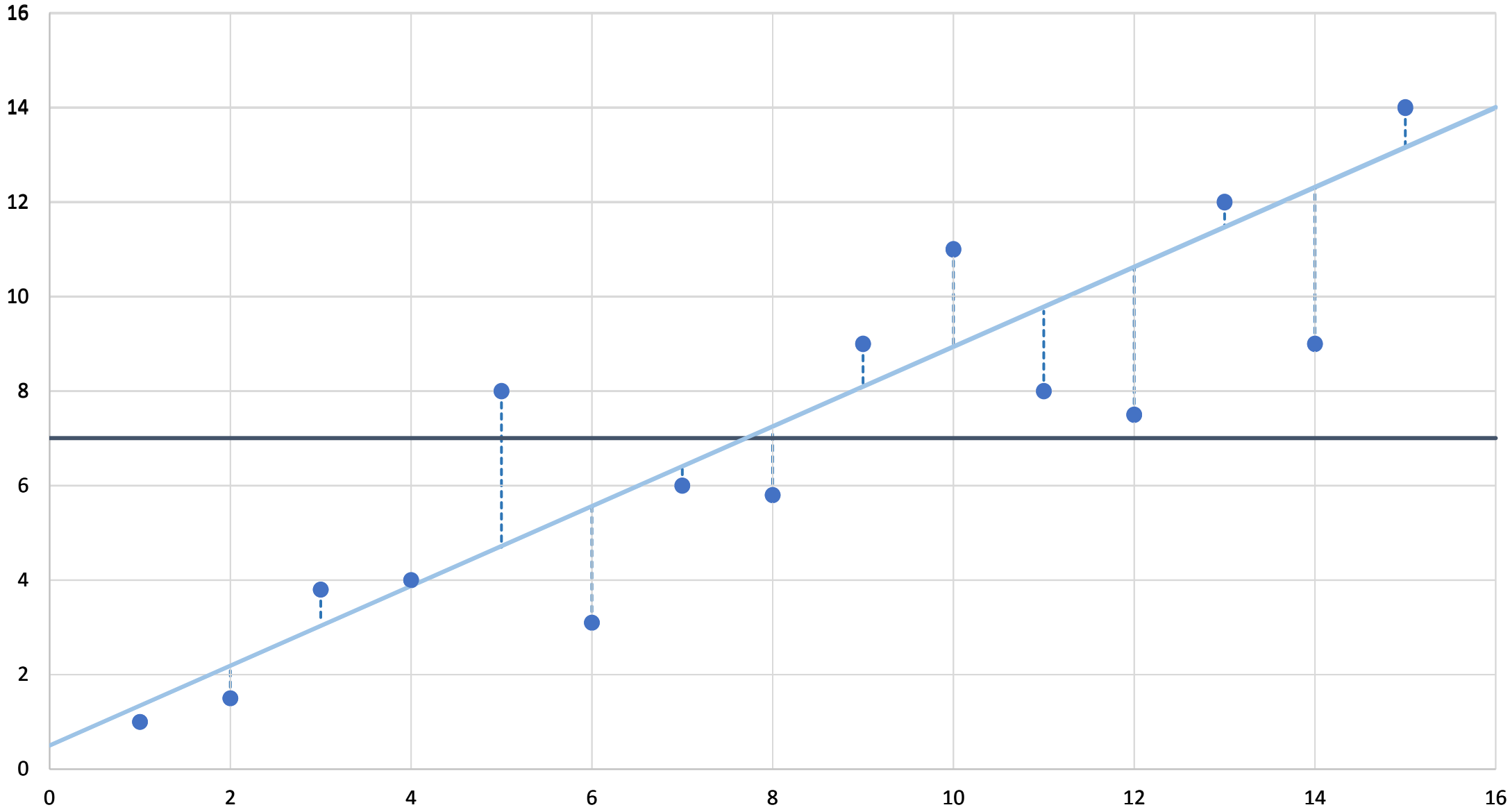
- Logistická regrese:
  - Mají muži vyšší pravděpodobnost dostat se do vězení než ženy?
  - Zvyšuje účast na přednáškách šance vyhnout se hodnocení F v kurzu?

OLS regrese - požadavky

- Závislá proměnná (outcome):
  - Přesně jedna proměnná, kardinální
- Nezávislé proměnné (predictors):
  - Jedna nebo více proměnných, přípustné jsou všechny druhy (možnost rekódování)
- Další požadavky:
  - Nezávislost pozorování
  - Absence multikolinearity mezi NP
  - Lineární vztah mezi NP a ZP
  - Normální distribuce chyb
  - Homoskedasticita


Průměr
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16


16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16


16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16

Výstupy lineární regrese

- Parametry:
  - Konstanta
  - Efekt každé NP

- $y = b_0 + b_1*x + b_2*y + b_3*z + \ldots$

- ***y*** – odhadovaná hodnota ZP
- $\boldsymbol{b_0}$ - konstanta
- $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, $\boldsymbol{b_3}$ – nestandardizované beta koeficienty za každou NP
- **x, y, z** – hodnoty NP

R^2 (index determinace)

- Informace o vhodnosti použitého modelu
- Jak dobře náš model (naše NP) vysvětlují změny ZP
- Matematicky jde o srovnání zlepšení regresní přímky oproti průměru
- Rozpětí od 0 po 1 (po úpravě od 0 po 100 %)

- Ukazuje kolik rozptylu ZP jsme schopní vysvětlit pomocí našich NP
- Používejte Adjusted R^2 – kontrola inflace počtu NP

Konstanta (intercept)

- Odhadovaná hodnota ZP za předpokladu, že všechny NP = 0

- $y = \mathbf{b_0} + b_1*x + b_2*y + b_3*z$

- Pokud x, y, z = 0 tak

- $y = \mathbf{b_0} + b_1*0 + b_2*0 + b_3*0$
- $y = \mathbf{b_0}$

Informace o efektech NP

- **Nestandardizovaný B koeficient:**
  - Ukazuje, jak se hodnota ZP změní, když se hodnota NP zvýší o jednotku
  - Např. pokud je NP měřená v hodinách – B ukazuje, jak se změní ZP, pokud se hodnota NP zvýší o jednu hodinu

- $y = b_0 + \mathbf{b_1}*x + \mathbf{b_2}*y + \mathbf{b_3}*z$

Informace o efektech NP

- **Nestandardizovaný B koeficient:**
  - Ukazuje, jak se hodnota ZP změní, když se hodnota NP zvýší o jednotku
  - Např. pokud je NP měřená v hodinách – B ukazuje, jak se změní ZP, pokud se hodnota NP zvýší o jednu hodinu

- **Standardizovaný Beta koeficient:**
  - Srovnává navzájem význam nezávislých proměnných
  - Větší vzdálenost od nuly ukazuje na vyšší význam nezávislé proměnné

- **Signifikantnost:**
  - Ukazuje, jestli se efekt NP dá aplikovat na populaci

Příklad

- Ovlivňuje velikost populace obce účast v lokálních volbách?

- H1: Účast klesá s rostoucí velikostí populace
- H0: Mezi velikostí populace a volební účastí není žádný vztah

- Závislá proměnná:
  - Účast – účast v %

- Nezávislá proměnná:
  - Populace – počet obyvatelů obce v tis.

JASP

- Regression > Linear Regression

- Proměnné:
  - Dependent Variable – Účast
  - Covariates – Populace_tis

Model Summary - Účast

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.265	0.070	0.070	12.589

Note. Null model includes Populace_tis

ANOVA ▼

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	35005.363	1	35005.363	220.867	< .001
	Residual	462316.523	2917	158.490		
	Total	497321.885	2918			

Note. Null model includes Populace_tis

- Model Summary:
  - Náš model vysvětluje 7 procent (0.070 * 100) variability ZP
- ANOVA:
  - Náš model je signifikantním pokrokem v odhadování ZP a pokud to umožňuje povaha dat, je možné naše výsledky aplikovat na populaci

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	60.801	0.244		248.807	< .001
	Populace_tis	−0.591	0.040	−0.265	−14.862	< .001

- Konstanta (Intercept):
  - Odhadovaná hodnota závislé proměnné pokud jsou hodnoty všech nezávislých proměnných = 0
  - V (neexistujícím) městě s populací 0 lidí model odhaduje, že volební účast bude 60,8 procenta

- $y = \mathbf{b_0} + b_1 * x$
- $y = 60.8 + b_1 * 0 = 60.8$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	60.801	0.244		248.807	< .001
	Populace_tis	−0.591	0.040	−0.265	−14.862	< .001

- Nestandardizovaný B koeficient:
  - Ukazuje, jak se hodnota ZP změní, pokud se hodnota NP zvýší o jednotku
  - Populace_tis je měřena v tis. lidí
  - Interpretace – za každý nárůst lokální populace o tisíc lidí (NP) se volební účast sníží o 0,591 procentního bodu (ZP)
  - Tento efekt je signifikantní na hladině 99,9 % a můžeme ho aplikovat na populaci (zamítáme nulovou hypotézu o absenci vztahu mezi NP a ZP) – **důležité pouze pokud je možné generalizovat zjištění ze vzorku na populaci!**

- $y = b_0 + \mathbf{b_1}*x$
- $y = 60.8 + (-0.591)*x$
- $y = 60.8 - 0.591*x$

Predikce volební účasti

- $y = b_0 + b_1 * x$
- Účast = 60.8 – 0.591*Populace_tis

	Populace	Populace v tis.	Rovnice	Odhadovaná účast
Město 1	500	0.5	60.8 – 0.591*0.5 = 60.8 – 0.296	**60.5**
Město 2	1,000	1	60.8 – 0.591*1 = 60.8 – 0.591	**60.2**
Město 3	5,000	5	60.8 – 0.591*5 = 60.8 – 2.955	**57.8**
Město 4	10,000	10	60.8 – 0.591*10 = 60.8 – 5.91	**54.9**
Město 5	25,000	25	60.8 – 0.591*25 = 60.8 – 14.775	**46.0**

Příklad 2

- Závisí účast v lokálních volbách na velikosti populace obce, lokální finanční situaci a soutěživosti?

- Závislá proměnná:
  - Účast – účast v %

- Nezávislé proměnné:
  - Populace – počet obyvatelů obce v tis.
  - Finanční situace – index finančního zdraví (1-6; 1 nejhorší, 6 nejlepší)
  - Soutěživost – existence soutěže (alespoň 2 soupeři) – 1 ano, 0 ne (binární proměnná)

JASP

- Regression > Linear Regression

- Proměnné:
  - Dependent Variable – Účast
  - Covariates – Populace_tis, Fin_index, Soutěž

Model Summary - Účast

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.675	0.456	0.455	9.621

Note. Null model includes Populace_tis, Fin_index, Soutěž

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	225581.160	3	75193.720	812.424	< .001
	Residual	269149.210	2908	92.555		
	Total	494730.370	2911			

Note. Null model includes Populace_tis, Fin_index, Soutěž

- Náš model vysvětluje 45,5 procent (0.455 * 100) variability ZP
- Podstatný pokrok v porovnání s modelem s pouze jednou NP

- Náš model je signifikantním pokrokem v odhadování ZP a pokud to umožňuje povaha dat, je možné naše výsledky aplikovat na populaci

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	55.570	1.912		29.069	< .001
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001
	Soutez	17.995	0.397	0.625	45.309	< .001

- Konstanta (Intercept):
  - Odhadovaná hodnota závislé proměnné pokud jsou hodnoty všech nezávislých proměnných = 0
  - V (neexistujícím) městě s populací 0 lidí, finančním indexem 0 a bez soutěže (Soutez = 0) model odhaduje, že volební účast bude 55,57 procenta

- $y = \mathbf{b_0} + b_1*x + b_2*y + b_3*z$
- $y = 55.57 + b_1*0 + b_2*0 + b_3*0 = 55.57$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	55.570	1.912		29.069	< .001
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001
	Soutez	17.995	0.397	0.625	45.309	< .001

- Nestandardizované B koeficienty:
  - Ukazují, jak se hodnota ZP změní, pokud se hodnota NP zvýší o jednotku
  - **Populace_tis** je měřena v tis. lidí
  - Interpretace – za každý nárůst lokální populace o tisíc lidí (NP) se volební účast sníží o 0,77 procentního bodu (ZP)
  - Tento efekt je signifikantní na hladině 99,9 % a můžeme ho aplikovat na populaci (zamítáme nulovou hypotézu o absenci vztahu mezi NP a ZP) – **důležité pouze pokud je možné generalizovat zjištění ze vzorku na populaci!**

- $y = b_0 + \mathbf{b_1}*x + b_2*y + b_3*z$
- $y = 55.57 - \mathbf{0.77}*x + b_2*y + b_3*z$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	55.570	1.912		29.069	< .001
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001
	Soutez	17.995	0.397	0.625	45.309	< .001

- Nestandardizované B koeficienty:
  - Ukazují, jak se hodnota ZP změní, pokud se hodnota NP zvýší o jednotku
  - **Fin_index** je měřený na škále 1 až 6
  - Interpretace – za každý nárůst hodnoty finančního indexu o 1 (NP) se volební účast sníží o 1,382 procentního bodu (ZP)
  - Tento efekt je signifikantní na hladině 99,9 % a můžeme ho aplikovat na populaci (zamítáme nulovou hypotézu o absenci vztahu mezi NP a ZP) – **důležité pouze pokud je možné generalizovat zjištění ze vzorku na populaci!**

- $y = b_0 + b_1*x + \mathbf{b_2}*y + b_3*z$
- $y = 55.569 - 0.77*x - \mathbf{1.382}*y + b_3*z$

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	55.570	1.912		29.069	< .001
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001
	Soutez	17.995	0.397	0.625	45.309	< .001

- Nestandardizované B koeficienty:
  - Ukazují, jak se hodnota ZP změní, pokud se hodnota NP zvýší o jednotku
  - **Soutěž** je binární proměnná (0 – žádná soutěž, 1 – alespoň dva kandidáti)
  - Interpretace – pokud v obci existuje volební soutěž (NP), volební účast se zvýší o 17,995 procentních bodů (ZP)
  - Tento efekt je signifikantní na hladině 99,9 % a můžeme ho aplikovat na populaci (zamítáme nulovou hypotézu o absenci vztahu mezi NP a ZP) – **důležité pouze pokud je možné generalizovat zjištění ze vzorku na populaci!**

- $y = b_0 + b_1*x + b_2*y + \mathbf{b_3}*z$
- $y = 55.569 - 0.77*x - 1.382*y + \mathbf{17.995}*z$

Nestandardizovaný B koeficient

- Kardinální vs. binární proměnné

- Stejná definice B pro oba typy proměnných:
  - Ukazuje, jak se změní hodnota ZP, pokud se hodnota NP zvýší o jednotku

ALE

- Binární (dummy) proměnné mají pouze dvě hodnoty – 0 a 1
  - Na rozdíl od kardinálních proměnných, je zde možný pouze jeden nárůst „o jednotku“
  - Odhadovaný efekt je tak kompletně vyčerpán tímto jediným nárustem (z 0 na 1)

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	55.570	1.912		29.069	< .001
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001
	Soutez	17.995	0.397	0.625	45.309	< .001

- **Soutěž:**
  - 0 – žádná soutěž (pouze jeden kandidát)
  - 1 – soutěž (alespoň dva kandidáti)
  - Nárůst z 0 na 1 znamená, že pro města se soutěží model odhaduje o téměř 18 procentních bodů vyšší účast než pro města bez soutěže

- **Populace_tis:**
  - Nárůst populace z 1 na 2 tisíce znamená snížení účasti o 0,77 p.b.
  - Nárůst populace z 1 na 5 tisíc znamená snížení účasti o 3,08 p.b. (4 * -0,77)
  - Nárůst populace z 5 na 12 tisíc znamená snížení účasti o 5,39 p.b. (7 * -0,77)

Standardizovaný Beta koeficient

- Poskytuje informaci o významu jednotlivých NP
- Měřený v standardních odchylkách → umožňuje jednoduché srovnání NP
- Větší vzdálenost od nuly (kladným i záporným směrem) indikuje vyšší význam NP

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	55.570	1.912		29.069	< .001
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001
	Soutez	17.995	0.397	0.625	45.309	< .001

- Výsledky ukazují, že z použitých NP je Soutěž nejvýznamnějším prediktorem volební účasti

- Populace_tis je méně významná a Fin_index je nejméně významná NP

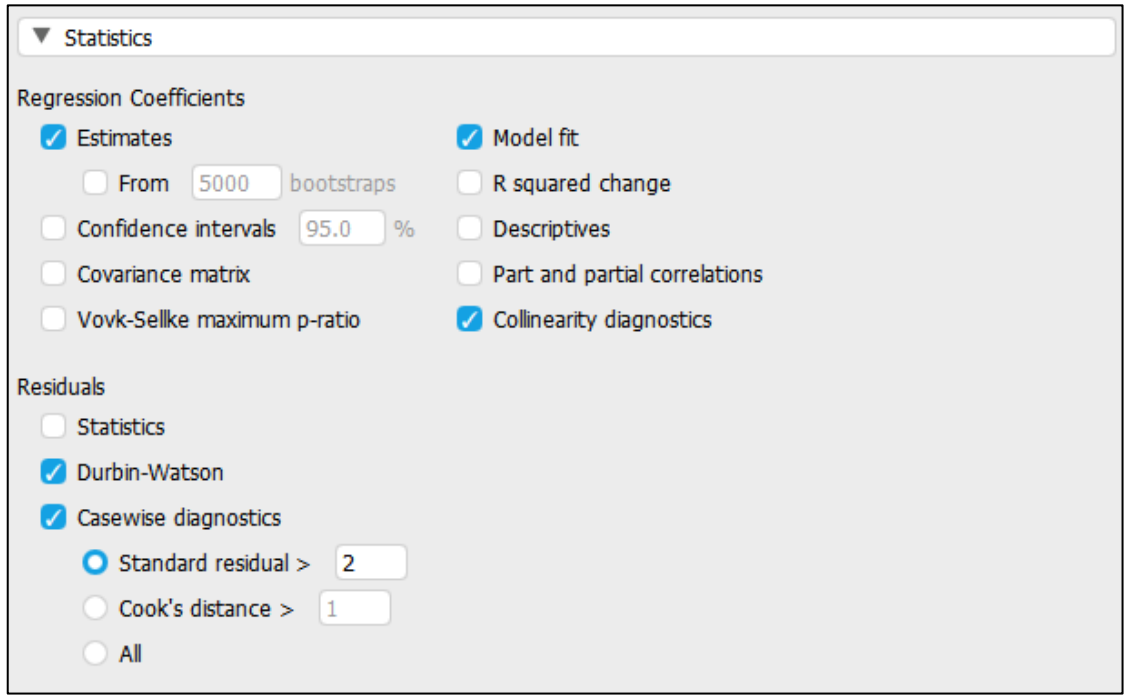
Predikce volební účasti

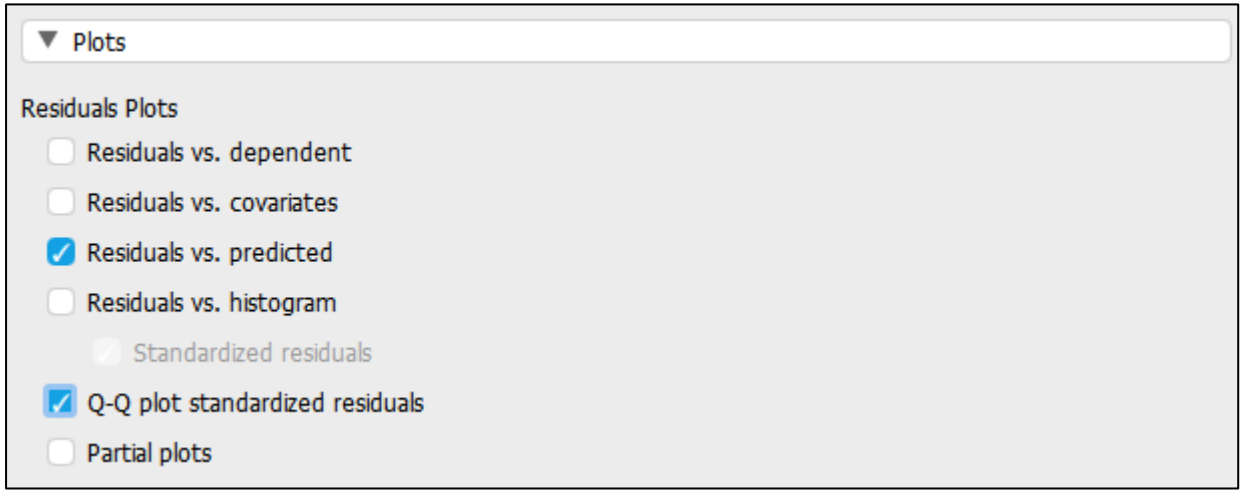
- $y = b_0 + b_1*x + b_2*y + b_3*z$
- Účast = 55.569 – 0.77*Populace_tis – 1.382*Fin_Index + 17.995*Soutěž

	Populace	Fin_Index	Soutěž	Rovnice	Odhadovaná účast
Město 1	1,000	3	0	55.57 – 0.77*1 – 1.382*3 + 17.995*0	**50.7**
Město 2	1,000	3	1	55.57 – 0.77*1 – 1.382*3 + 17.995*1	**68.6**
Město 3	5,000	3	0	55.57 – 0.77*5 – 1.382*3 + 17.995*0	**47.6**
Město 4	10,000	6	1	55.57 – 0.77*10 – 1.382*6 + 17.995*1	**57.6**
Město 5	25,000	6	0	55.57 – 0.77*25 – 1.382*6 + 17.995*0	**28.0**

Kontrola předpokladů

- Kontrola reziduí
- Homoskedasticita
- Multikolinearita – asociace mezi NP





Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
H_0	(Intercept)	55.570	1.912		29.069	< .001		
	Populace_tis	−0.770	0.031	−0.347	−25.078	< .001	0.980	1.020
	Fin_index	−1.382	0.361	−0.053	−3.831	< .001	0.994	1.006
	Soutez	17.995	0.397	0.625	45.309	< .001	0.984	1.016

Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Intercept)	Populace_tis	Fin_index	Soutez
H_0	1	2.933	1.000	0.001	0.021	0.001	0.028
	2	0.858	1.849	0.000	0.964	0.000	0.004
	3	0.205	3.786	0.006	0.009	0.007	0.967
	4	0.004	25.770	0.993	0.006	0.992	0.001

- VIF nad 5 (10) nebo Tolerance nižší než 0,2 (0,1) indikují problém
- Podobně, více vyšších hodnot na stejné dimenzi indikuje multikolinearitu
- Řešení – více modelů / odstranění proměnné

Rezidua

- Data by měla obsahovat maximálně:
  - 5 % případů, které mají rezidua > 2 (< -2)
  - 1 % případů, které mají rezidua > 2,5 (< -2,5)
- Pokud identifikujeme odlehlé případy, je možné spočítat model bez těchto případů a porovnat výsledky, zda došlo ke změně

Homoskedasticita

- Grafické zobrazení nemá znázorňovat viditelný vzorec
- Heteroskedasticita znamená problém, pokud je naší ambicí generalizovat výsledky na populaci

Rekódování na více dummy proměnných

- Fin_index – kardinální proměnná

- Rekódujeme na 3 dummy proměnné (1/0):
  - Fin_low – finanční zdraví < 4,5
  - Fin_mid – finanční zdraví 4,5 – 5,5
  - Fin_high – finanční zdraví > 5,5

- Každý případ má hodnotu 1 v jedné dummy proměnné a 0 ve zbylých dvou
- Do modelu vstupují pouze 2 z 3 dummy proměnných – zbylá dummy proměnná slouží jako **referenční kategorie**

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	50.903	0.753		67.606	< .001
	Populace_tis	−0.767	0.031	−0.345	−24.991	< .001
	Soutez	17.993	0.397	0.624	45.302	< .001
	Fin_mid	−2.539	0.737	−0.094	−3.443	< .001
	Fin_high	−3.146	0.772	−0.112	−4.077	< .001

- Interpretace nestandardizovaných B koeficientů:
  - Srovnáváme s vyloučenou (referenční) kategorii

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	50.903	0.753		67.606	< .001
	Populace_tis	−0.767	0.031	−0.345	−24.991	< .001
	Soutez	17.993	0.397	0.624	45.302	< .001
	Fin_mid	−2.539	0.737	−0.094	−3.443	< .001
	Fin_high	−3.146	0.772	−0.112	−4.077	< .001

- Interpretace nestandardizovaných B koeficientů:
  - Srovnáváme s vyloučenou (referenční) kategorii

  - V obcích se středním finančním zdravím (Fin_mid = 1) model odhaduje, že volební účast bude o 2,54 procentních bodů nižší **než v referenční kategorii**, tedy oproti obcím s nízkým finančním zdravím (Fin_low = 1)
  - Tento rozdíl je signifikantní na hladině 99,9 % (důležité, pokud generalizujeme naše výstupy na populaci)

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	50.903	0.753		67.606	< .001
	Populace_tis	−0.767	0.031	−0.345	−24.991	< .001
	Soutez	17.993	0.397	0.624	45.302	< .001
	Fin_mid	−2.539	0.737	−0.094	−3.443	< .001
	Fin_high	−3.146	0.772	−0.112	−4.077	< .001

- Interpretace nestandardizovaných B koeficientů:
  - Srovnáváme s vyloučenou (referenční) kategorii

  - V obcích s vysokým finančním zdravím (Fin_high = 1) model odhaduje, že volební účast bude o 3,15 procentních bodů nižší **než v referenční kategorii**, tedy oproti obcím s nízkým finančním zdravím (Fin_low = 1)
  - Tento rozdíl je signifikantní na hladině 99,9 % (důležité, pokud generalizujeme naše výstupy na populaci)