Jednoduchá lineární regrese

Petr Ocelík

MVZn4003 Úvod do kvantitativních metod

Opakování

- Co měří Pearsonův korelační koeficient?

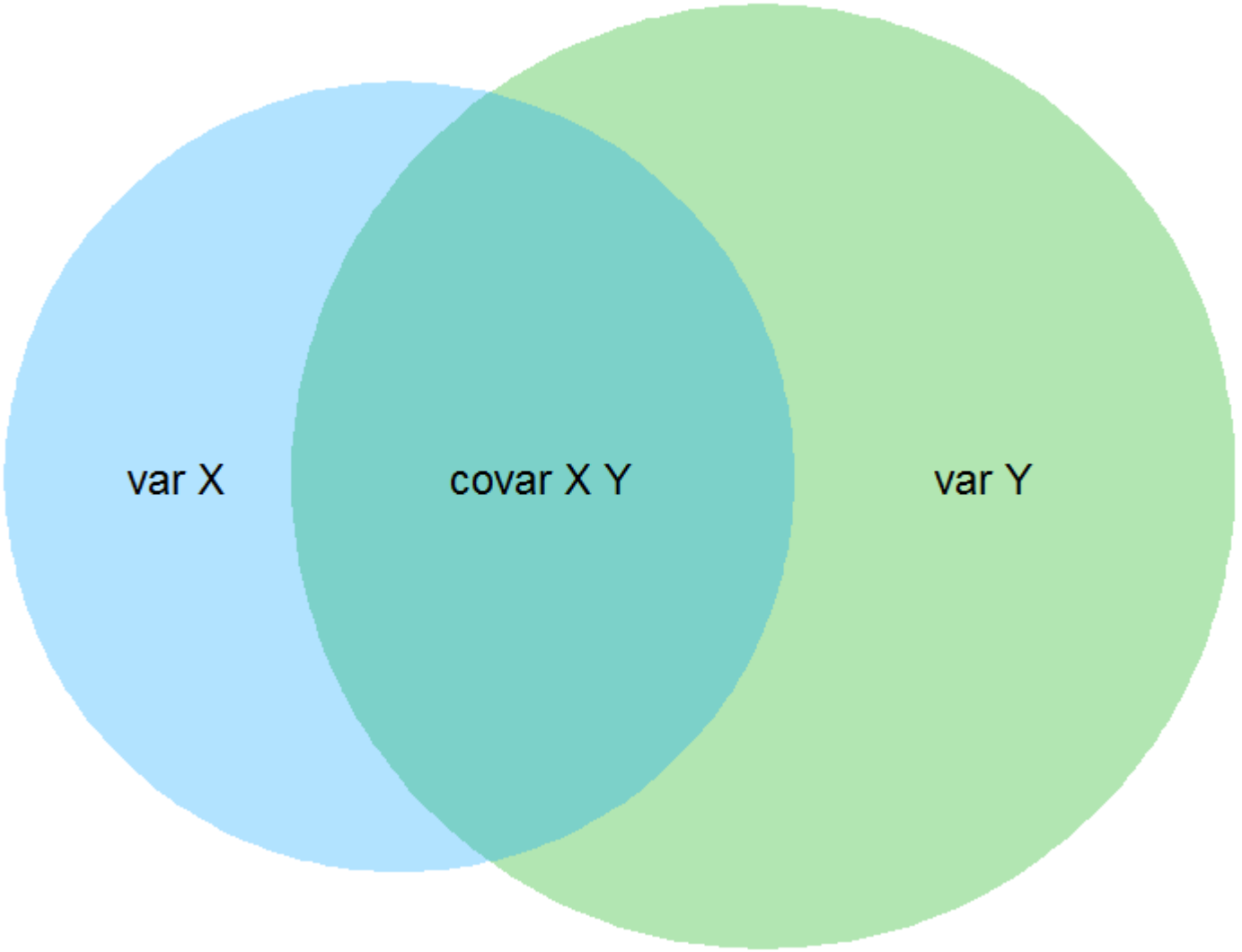
Pearsonův korelační koeficient r

- Pearsonův korelační koeficient (***r***)
- Pearsonovo *r* měří **sílu a směr <u>lineární závislosti</u> mezi dvěma spojitými proměnnými**

- Pearsonovo *r* nabývá hodnot <-1,1>
  - Dokonale pozitivní lineární vztah = 1
  - Dokonale negativní lineární vztah = -1
  - Žádný lineární vztah = 0

- Hodnota *r* není závislá na jednotkách proměnných

- r = kovariance X, Y / kombinovaný rozptyl X, Y



Opakování

- Co nám říká p-hodnota?

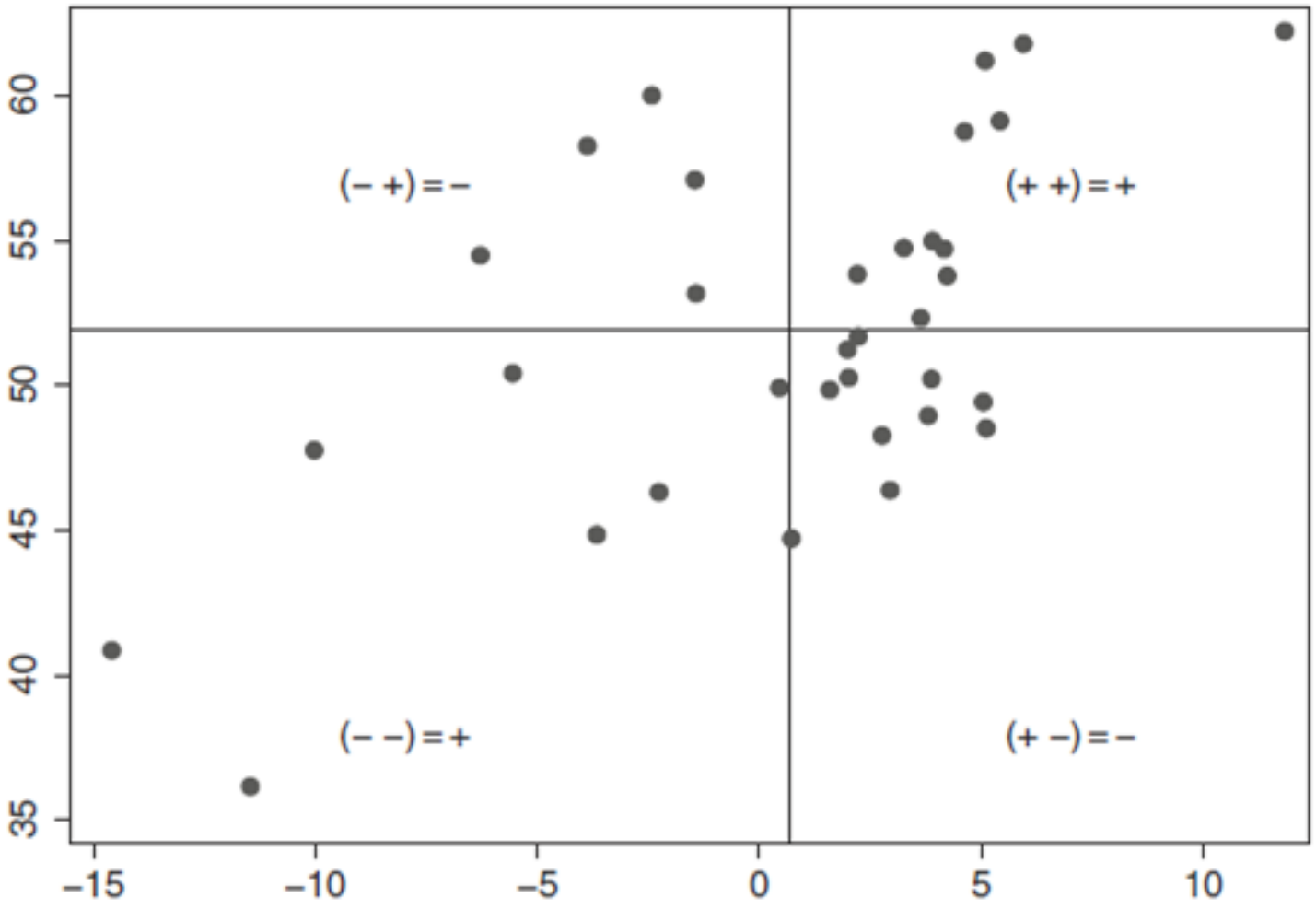
Blok přednášek

1. Jednoduchá lineární regrese
2. Vícečetná lineární regrese
3. Diagnostika a aplikace

Blok přednášek

1. **Jednoduchá lineární regrese**
    1. intuice
    2. regresní model
    3. metoda nejmenších čtverců (OLS)
    4. koeficient determinace
    5. interpretace výsledků
2. Vícečetná lineární regrese
3. Diagnostika a aplikace

- Pearsonovo **r = 0.54** (X = SES, Y = spokojenost s pol. systémem)



upravený Kellstedt & Whitten 2013

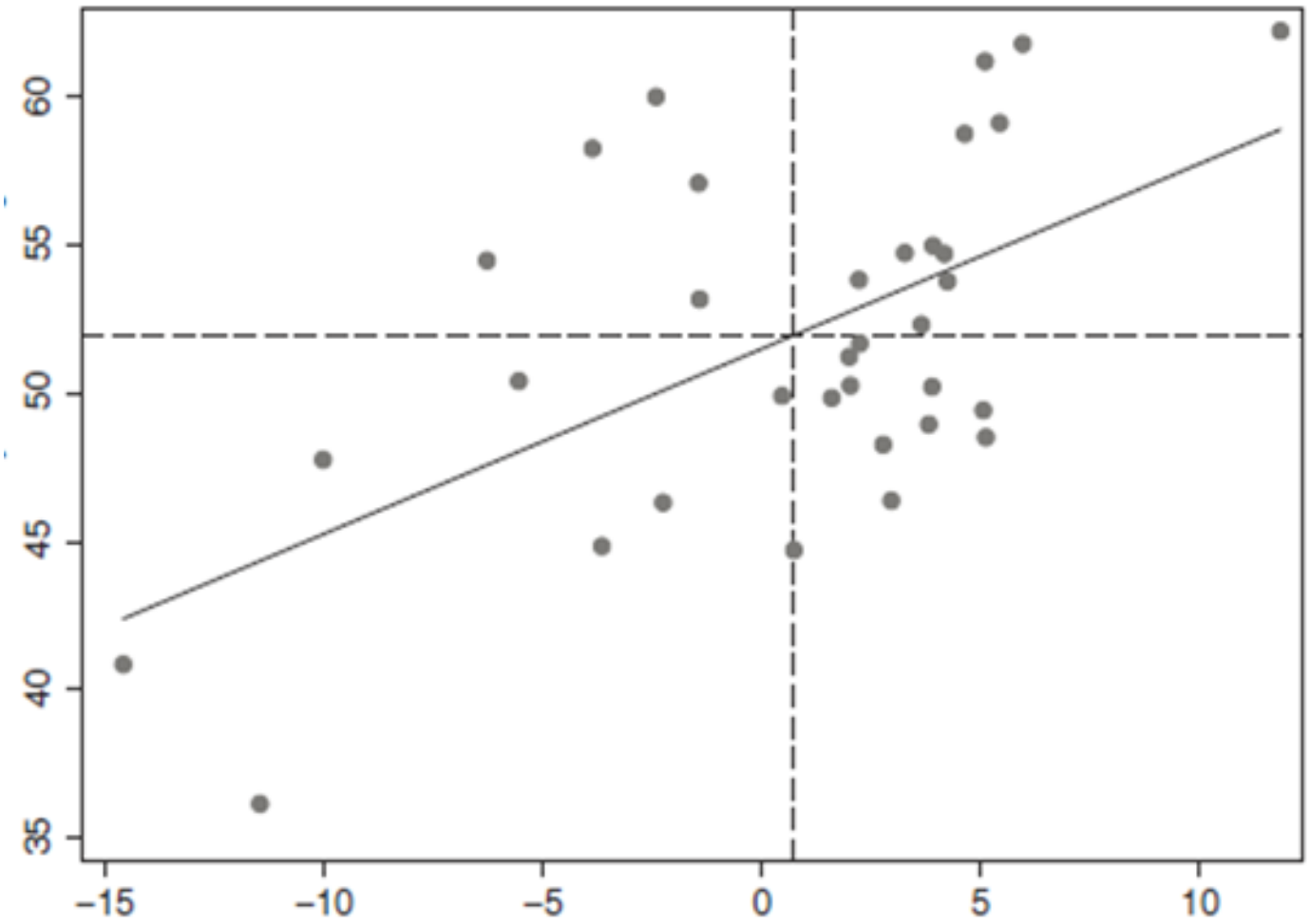
korelace (r)

- Symetrická metoda (nerozlišujeme ZP a prediktor)
- Měří sílu i směr vztahu
- O kolik se v průměru mění hodnoty proměnné v závislosti na změně hodnot druhé?

regrese

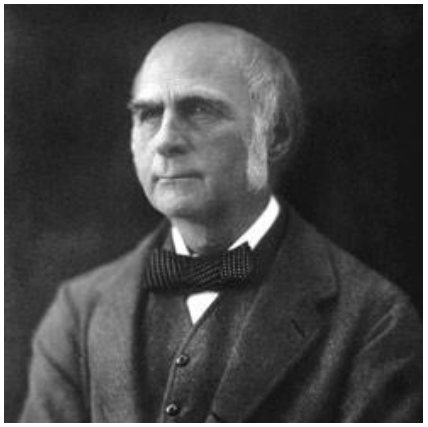
- Asymetrická metoda (rozlišujeme ZP a prediktor/y)
- Měří sílu i směr vztahu a **umožňuje predikci hodnot ZP** v závislosti na hodnotách prediktoru/ů.
- Jak velký je vliv prediktoru/ů na ZP?
- Jaká bude hodnota ZP při určité hodnotě prediktoru/ů?

- Y = 51.55 + 0.62 * X (regresní model)
- Y = spokojenost s pol. systémem; X = SES



upravený Kellstedt & Whitten 2013

(Lineární) regrese

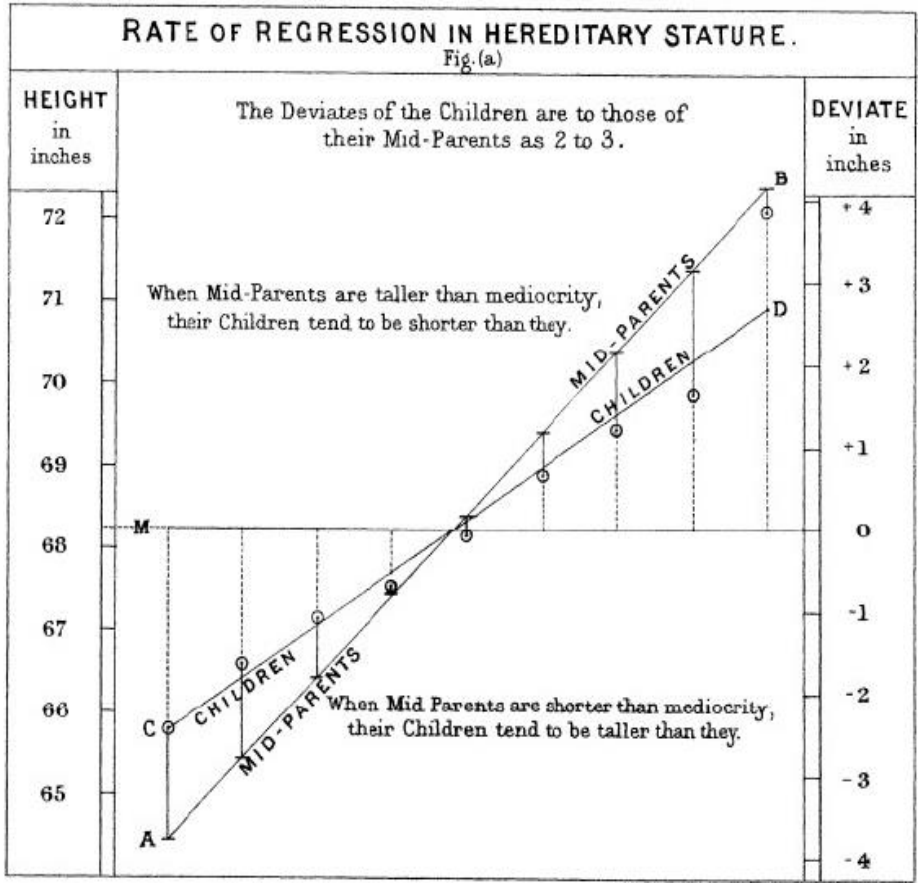


wikimedia

- Regrese je **metoda predikce** (odhadu) hodnot **závislé proměnné** (Y) na základě hodnot jedné (X) či více **nezávislých proměnných** (prediktorů).

- Lineární regrese odhaduje lineární vztah

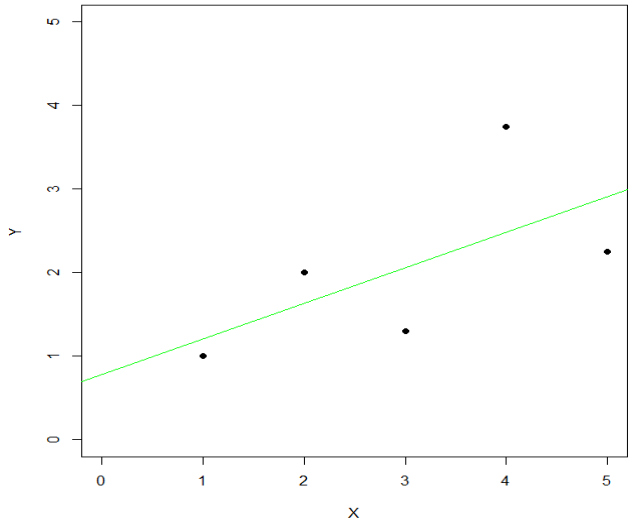
- **Jednoduchá lineární regrese**: pouze jeden prediktor
- Vícečetná lineární regrese: více než jeden prediktor



Galton 1886

Lineární vztah

- **Lineární funkce** je funkce, jejíž hodnota v celém jejím rozsahu *rovnoměrně* (lineárně) roste či klesá
- **Lineární vztah** je možné zobrazit **přímkou** *(linearis)*



$Y = 0.78 + 0.425*X$

- **Lineární <u>regresní</u> funkce** má tvar: $Y = b_0 + b_1*X$
- Y je závislá proměnná, X je nezávislá proměnná (prediktor), b_0, b_1 jsou koeficienty

Regresní model: lineární regresní funkce

- **Lineární regresní funkce** má tvar: $Y = b_0 + b_1 * X$
- Y je závislá proměnná
- X je nezávislá proměnná (prediktor)
- b_0, b_1 jsou koeficienty

protože teoretickou funkci pouze odhadujeme na základě vzorku prostřednictvím empirické funkce, odhad nutně zahrnuje nevysvětlený rozptyl ZP (chyba modelu)

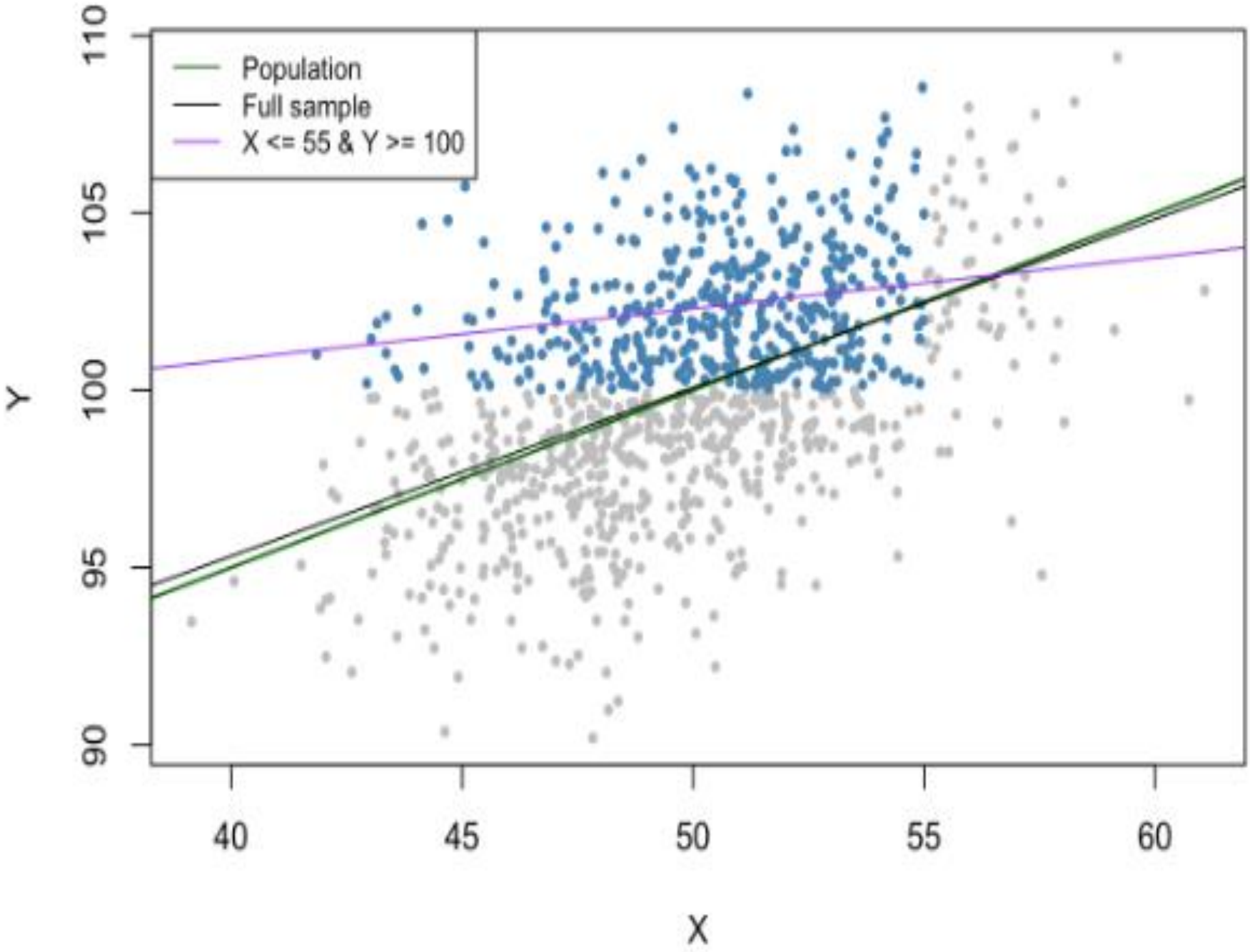
- **Rovněž**:
- $Y = b_0 + b_1 * X$
- závislá proměnná = průsečík + sklon*prediktor

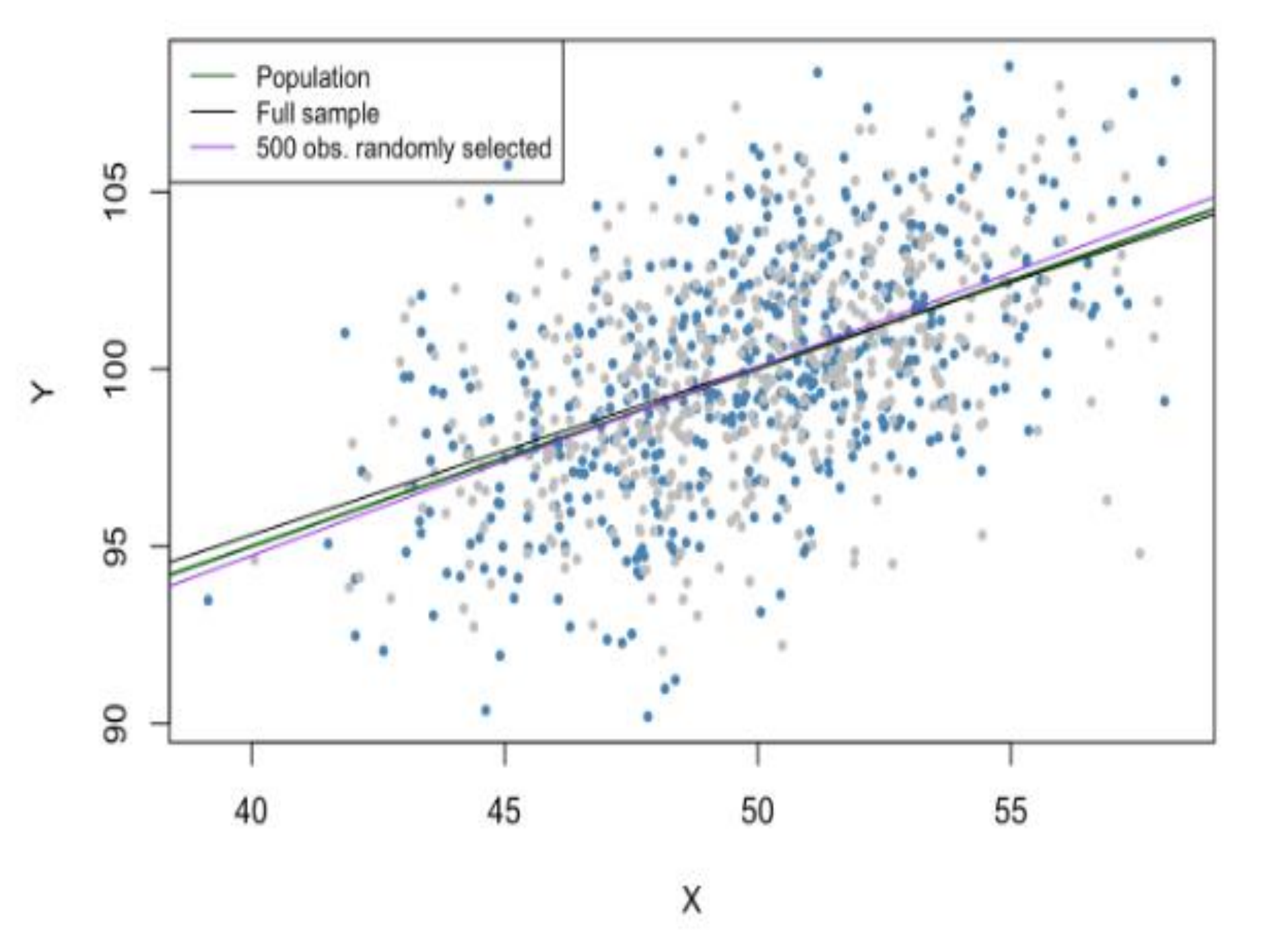
- **Rozlišení pro populaci a vzorek**:

$Y = \beta_0 + \beta_1 X + \varepsilon$; teoretická regresní funkce (populace)

$Y = b_0 + b_1 X + e$; empirická regresní funkce (vzorek)

závislá proměnná = průsečík + sklon*prediktor + chyba


Population
Full sample
X <= 55 & Y >= 100
110
105
100
95
90
Y
40
45
50
55
60
X


Population
Full sample
500 obs. randomly selected
Y
90
95
100
105
40
45
50
55
X

Terminologie a notace

X	Y
independent variable	dependent variable
predictor variable	outcome variable
explanatory variable	response variable

α, a, b, β0, B0, m	β, B, b	ε, e
intercept	slope / gradient	error / residual
constant	coefficient	
alpha	Beta	

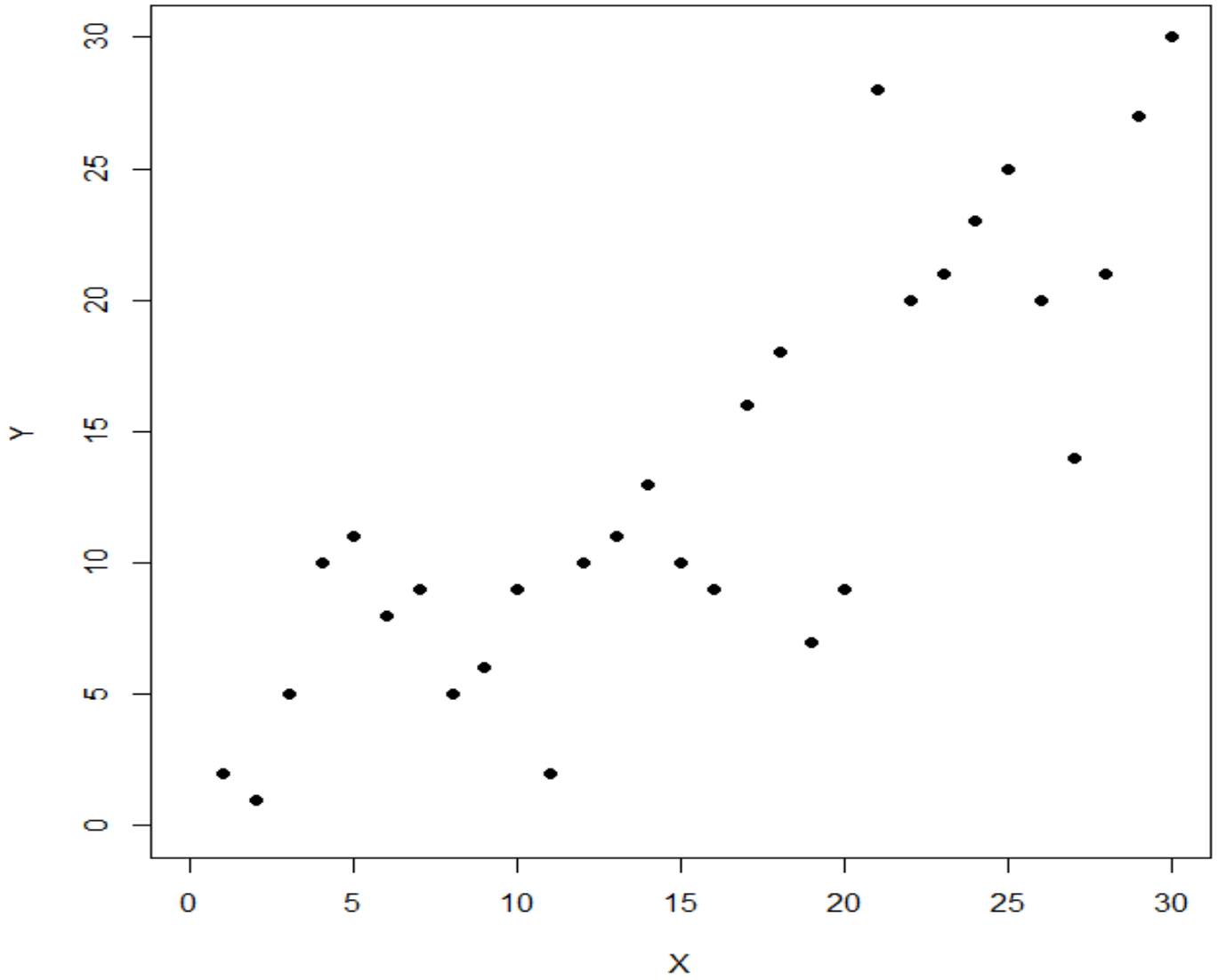
Terminologie a notace

X	Y
nezávislá proměnná	**závislá proměnná**
prediktor	sledovaná proměnná
vysvětlující proměnná	studovaná proměnná

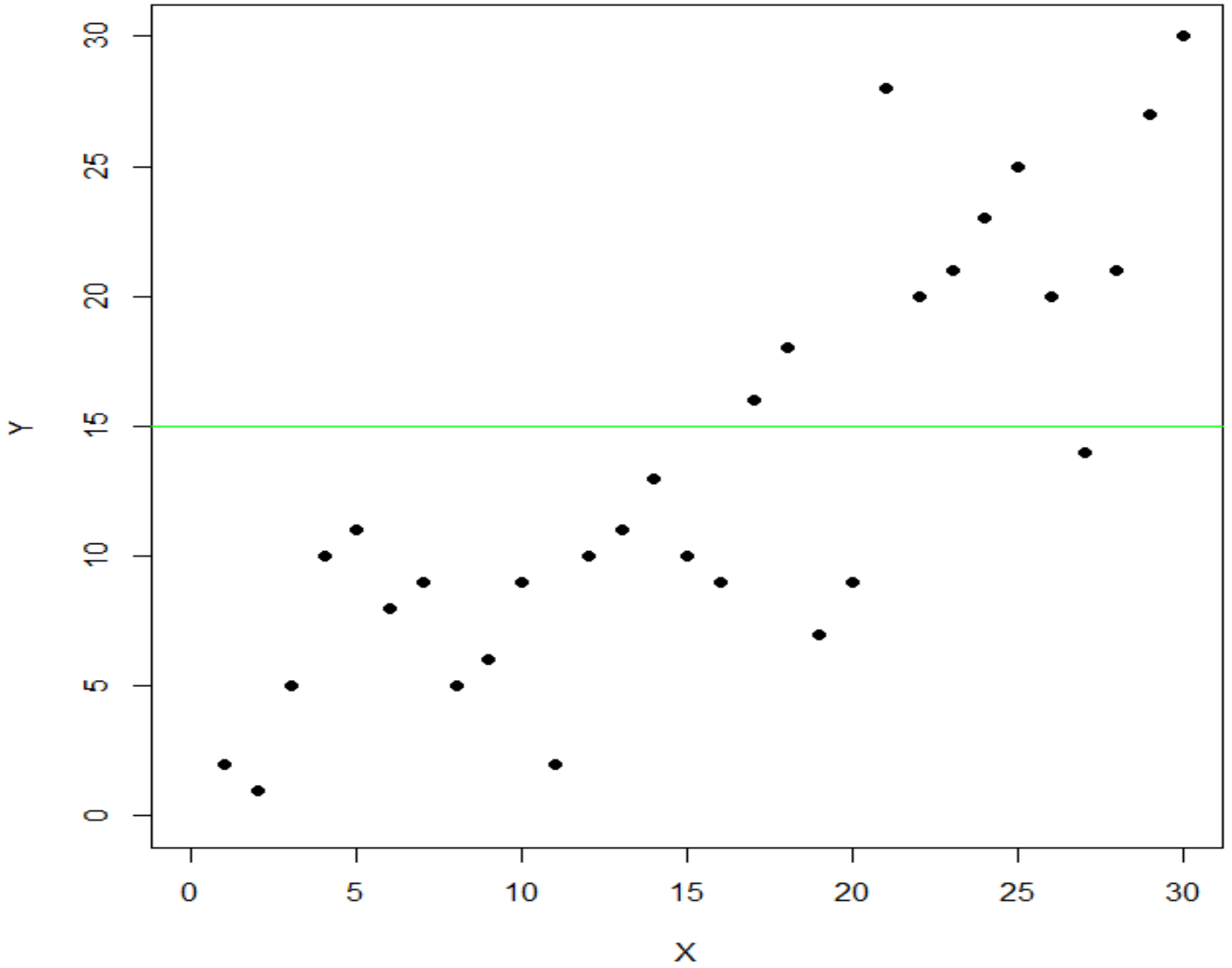
α, a, b0, β0, B0, m	β, B, b	ε, e
průsečík / konstanta	**sklon** / gradient	chyba / **reziduum**
průsečíková konstanta průsečíkový koeficient	konstanta sklonu koeficient sklonu	
alpha	Beta	

Jak odhadujeme regresní model?

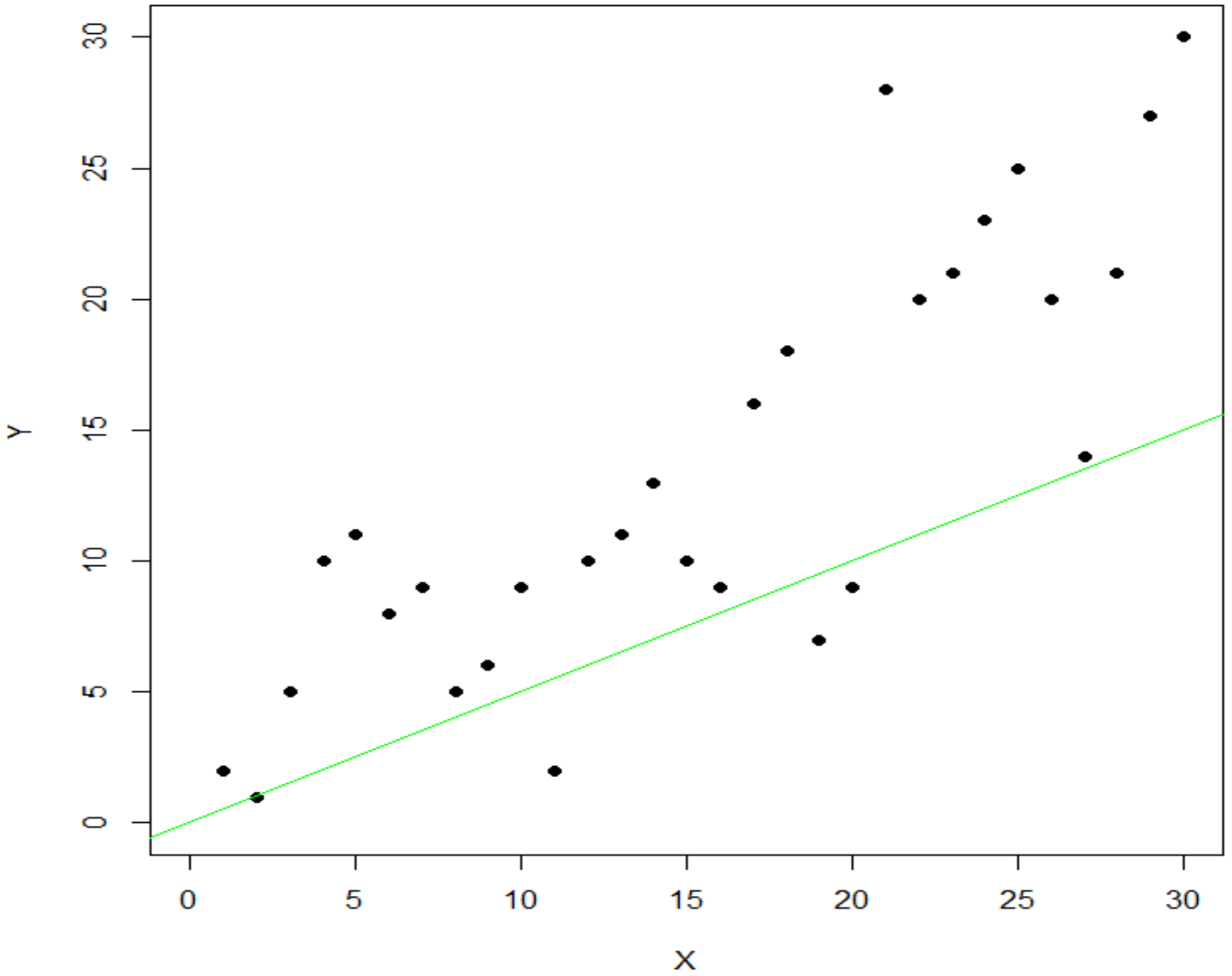
Umístěte regresní přímku



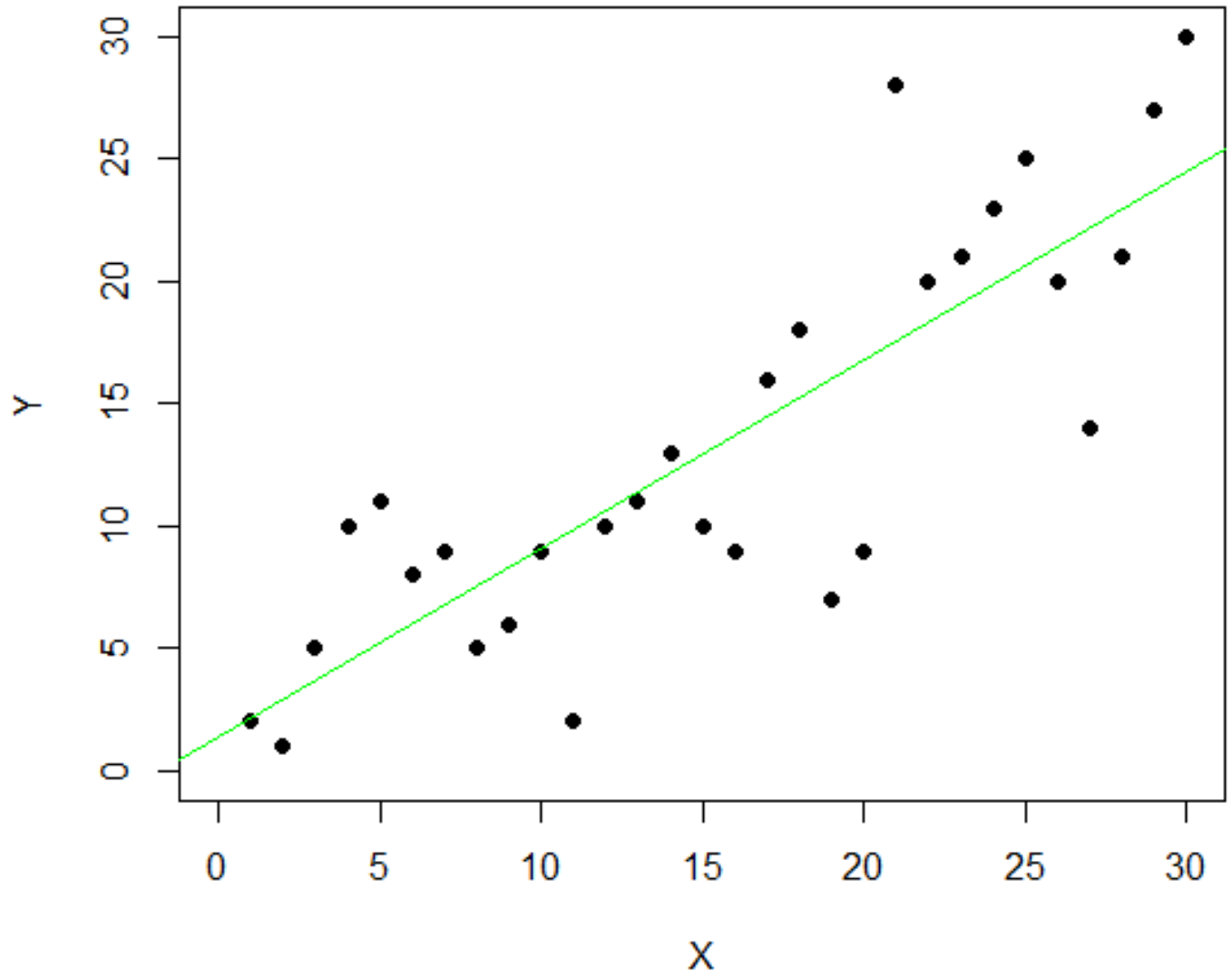
Umístěte regresní přímku



Umístěte regresní přímku



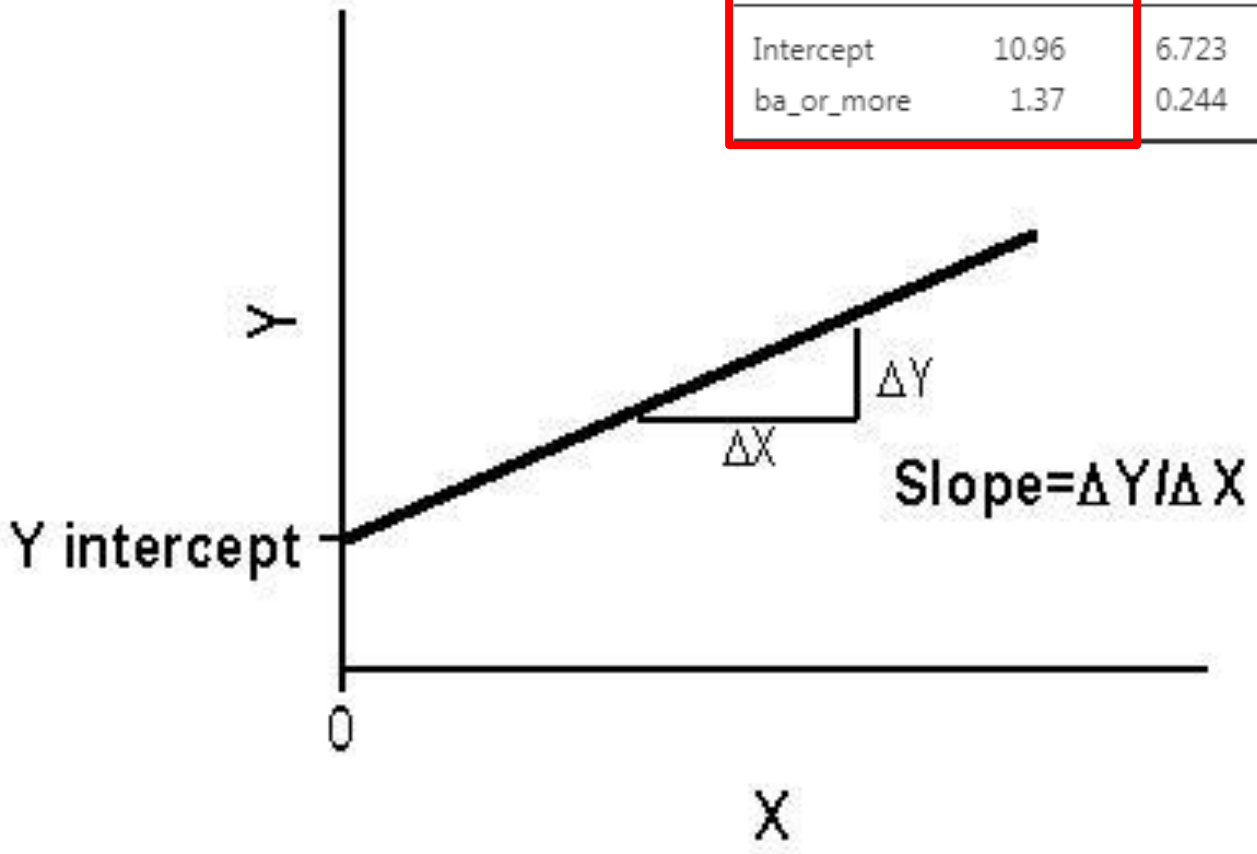
Umístěte regresní přímku



Regresní přímka: průsečík a sklon

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	10.96	6.723	1.63	0.110
ba_or_more	1.37	0.244	5.61	< .001



Metoda nejmenších čtverců

- **Metoda nejmenších čtverců (Ordinary Least Squares):** odhad koeficientů, průsečíku a sklonu, v lineárním regresním modelu

- **Minimalizuje vertikální vzdálenosti** mezi **pozorovanými** hodnotami (**Y**) a regresní přímkou reprezentující **predikované** hodnoty Y (**Y'**).

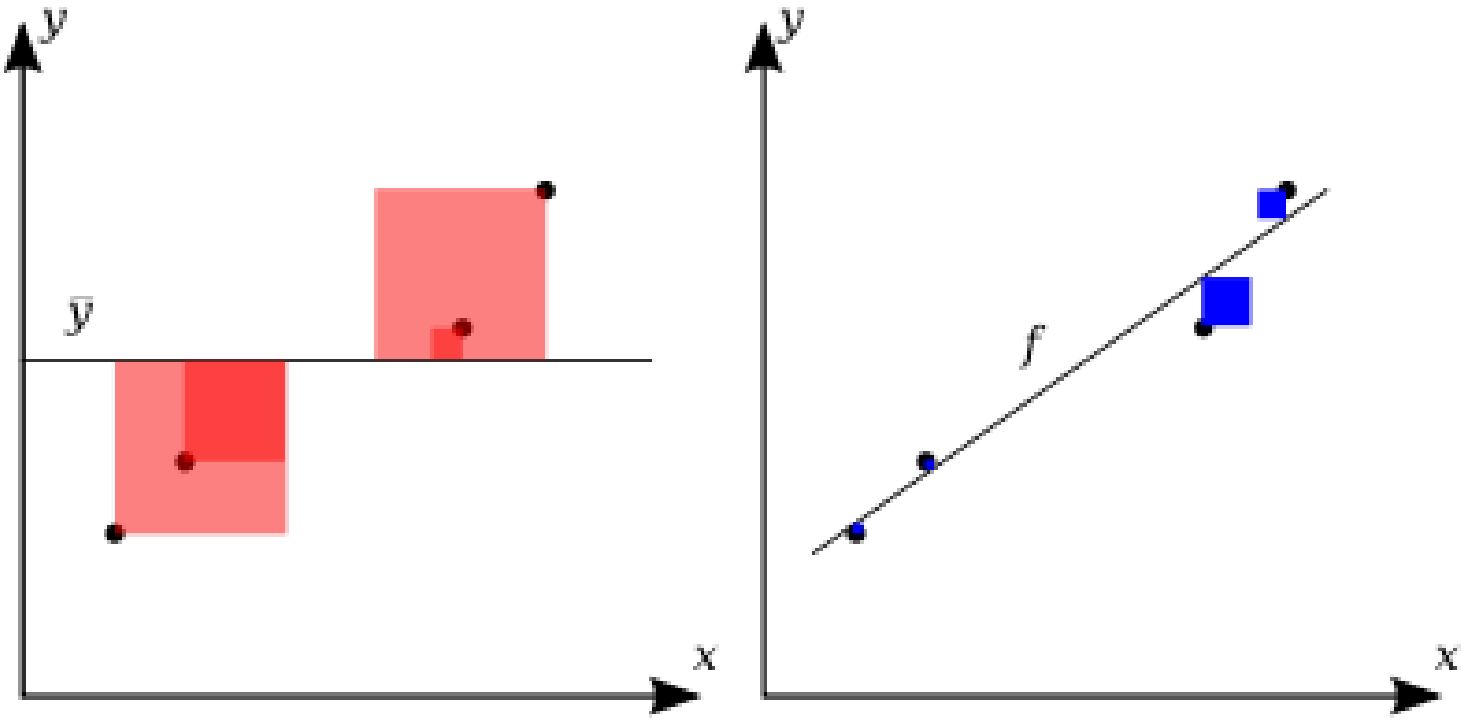
- **Reziduum** je **chyba predikce** daná rozdílem mezi pozorovanou (Y) a predikovanou hodnotou (Y') ZP

- **Reziduum = (Y - Y')** ; součet reziduí ∑(Y - Y') = 0
- Kvadratické reziduum (čtverec rezidua) = (Y - Y')^2
- **součet kvadratických reziduí** ∑(Y - Y ')^2

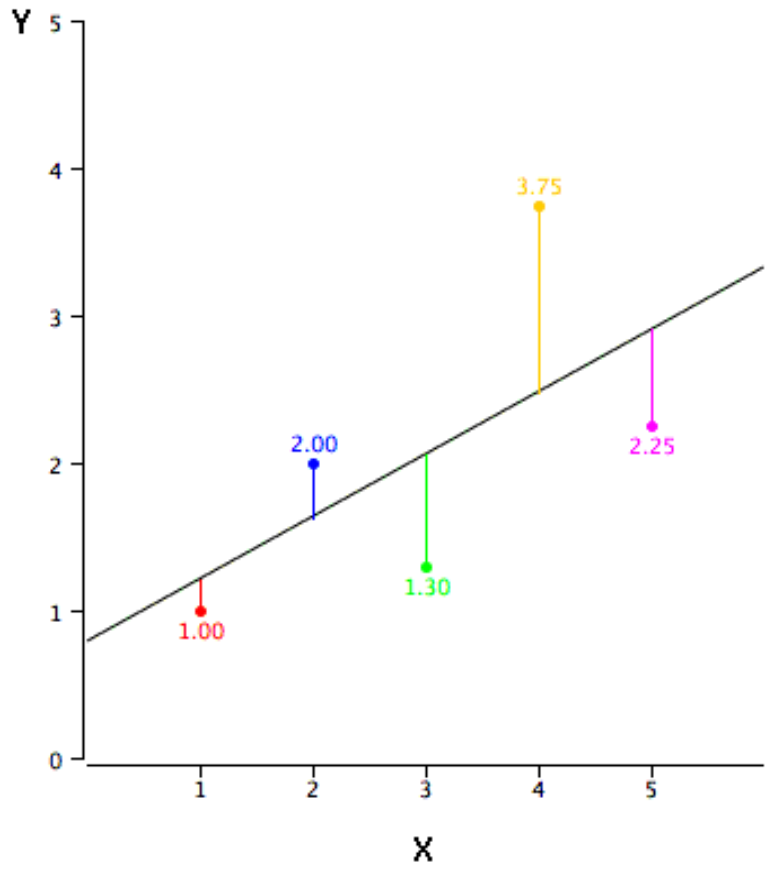
- **OLS minimalizuje součet kvadratických reziduí:** Y' = min∑(Y - Y')^2

Metoda nejmenších čtverců

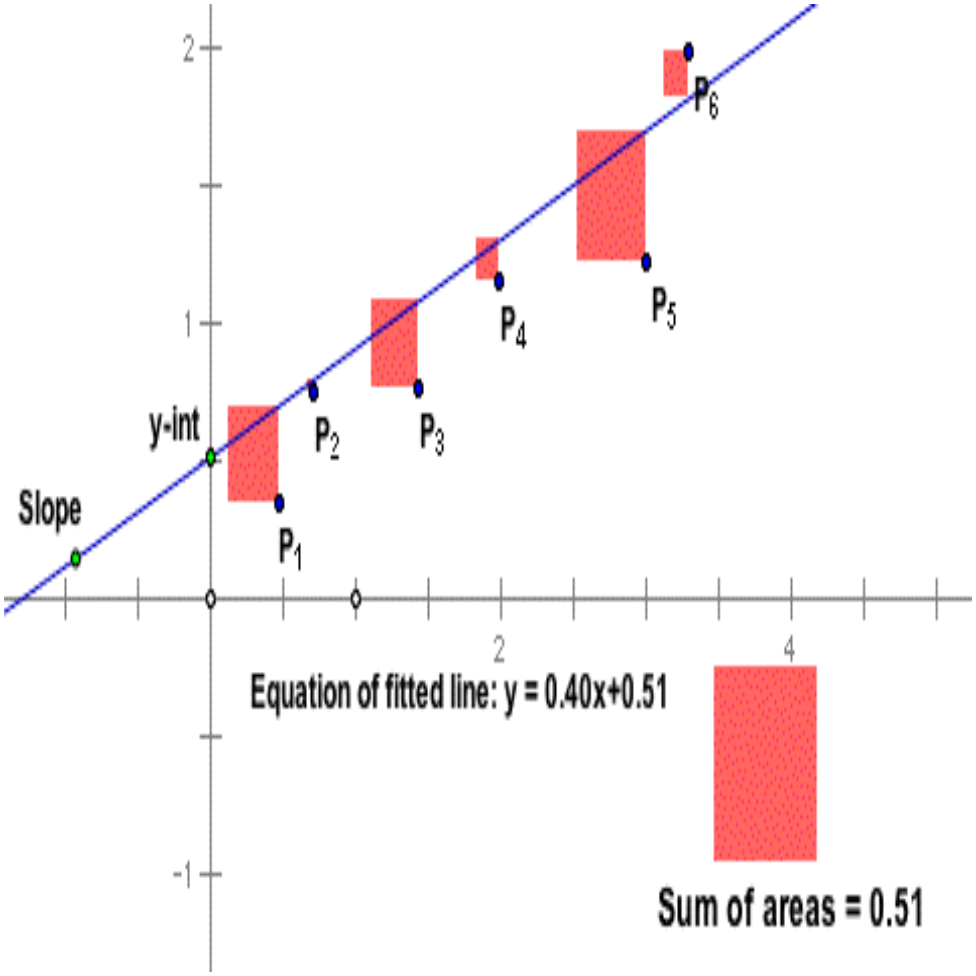
- Srovnání kvadratických (čtverců) reziduí pro průměr Y a regresní model



Metoda nejmenších čtverců



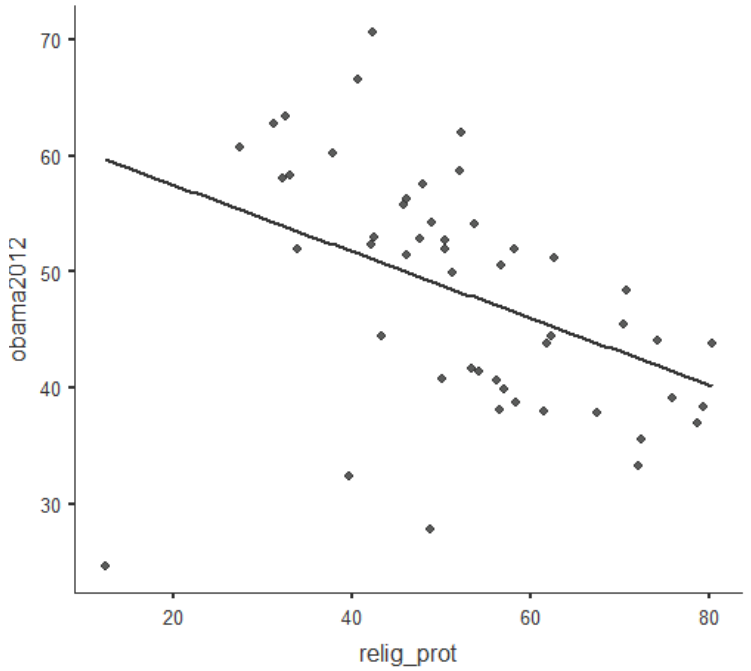
Metoda nejmenších čtverců



příklad

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	63.195	4.9721	12.71	< .001
relig_prot	−0.287	0.0914	−3.14	0.003

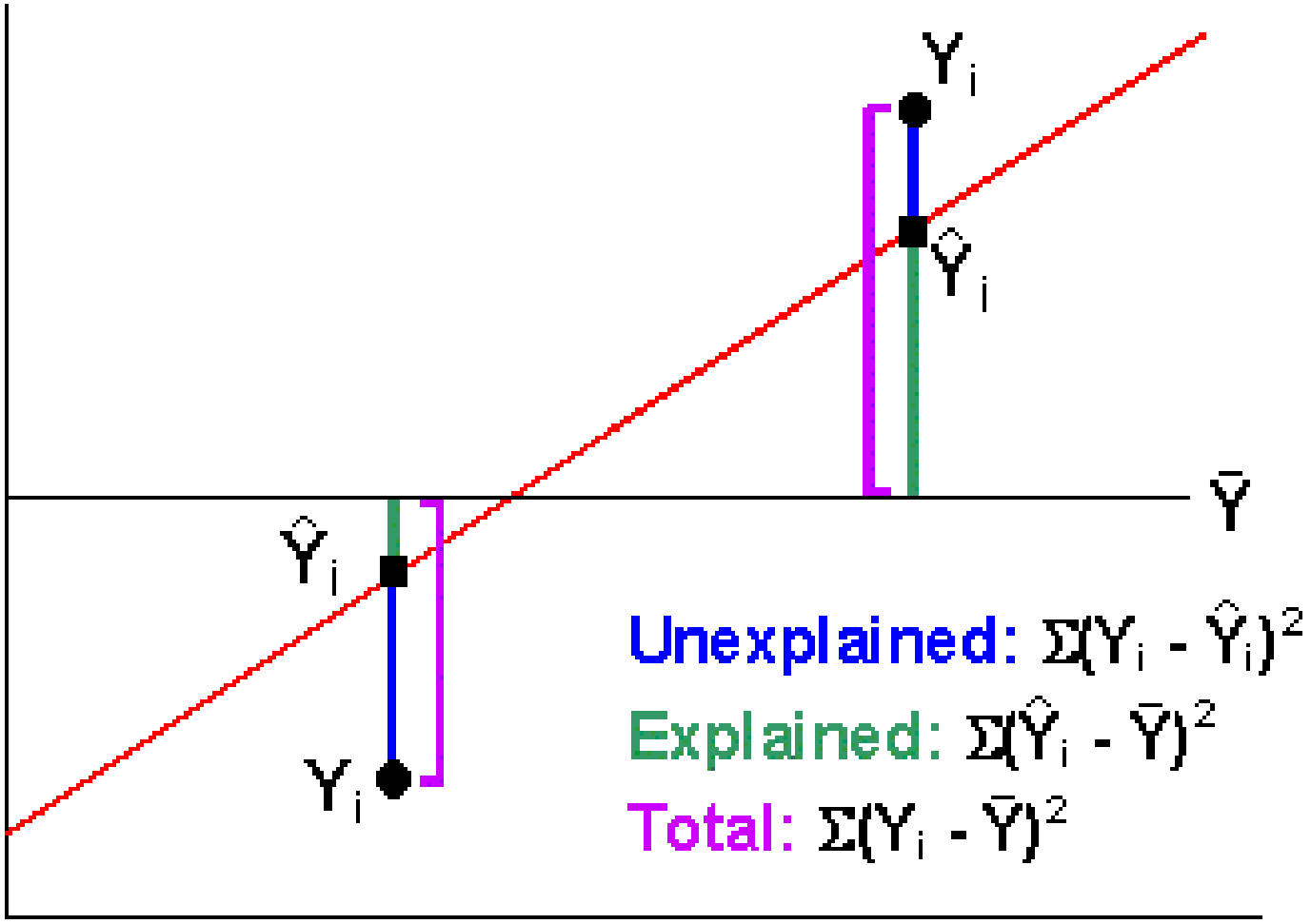


- **ZP (Y)** = % výsledek Obamy ve volbách 2012 (obama2012)
- **Prediktor (X)** = % podíl protestantů (relig_prot)
- **Průsečík (intercept)** = hodnota Y při X = 0
- **Sklon (slope)** = změna hodnoty Y při jednotkové změně X

→ při zvýšení podílu protestantů o 1 p.b. model predikuje snížení volebního zisku Obamy o -0.29 p.b.

Jak dobře regresní model vysvětluje rozptyl závislé proměnné?

Celkový rozptyl = **nevysvětlený rozptyl** + **vysvětlený rozptyl**



- **<u>Rozptyl vysvětlený</u>** regresním model (**SSM**: sum of squares of model)

$$\mathrm{SSM} = \Sigma\big(\mathrm{Y}' - \ \mathrm{mean}(\mathrm{Y})\big)^2$$

- **<u>Nevysvětlený rozptyl:</u>** chyba modelu (**SSR**: sum of squares of residuals):

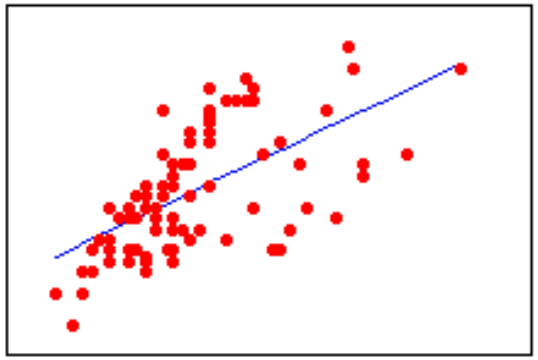
$$\mathrm{SSR} = \Sigma(\mathrm{Y} - \mathrm{Y}')^2$$

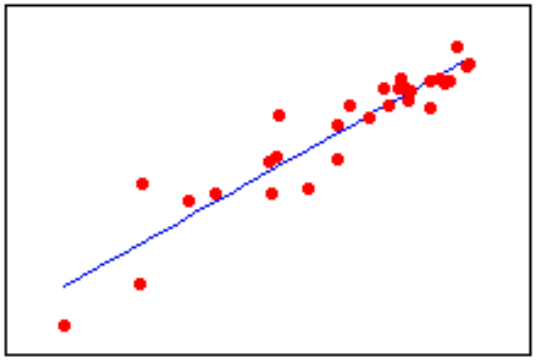
- **<u>Celkový rozptyl</u>** (**SST**: total sum of squares) = SSM + SSR

$$\mathrm{SST} = \Sigma\big(\mathrm{Y} - \mathrm{mean}(\mathrm{Y})\big)^2$$

Y' = predikované hodnoty (přímka)
Y = pozorované hodnoty (body)
mean(Y) = průměrná hodnota ZP

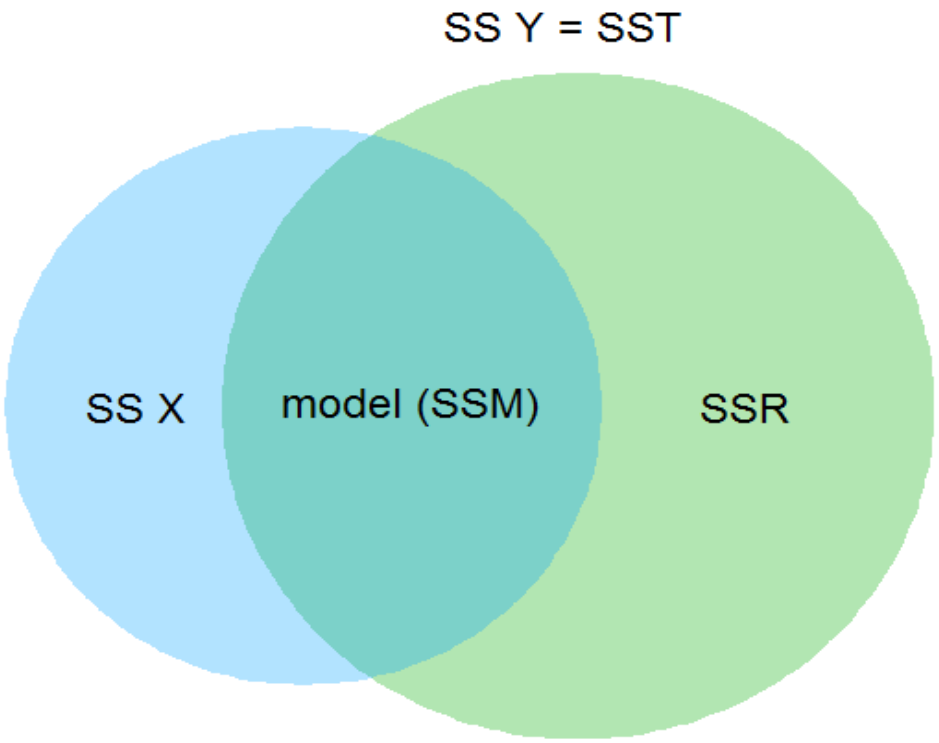
- Podíl **SSM/SST** ukazuje **(1)** těsnost regresního vztahu mezi ZP a prediktory a **(2)** přesnost predikce založené na regresním modelu

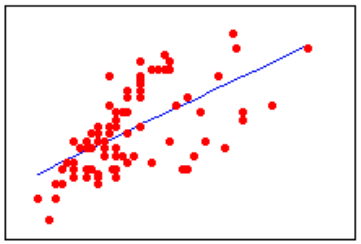


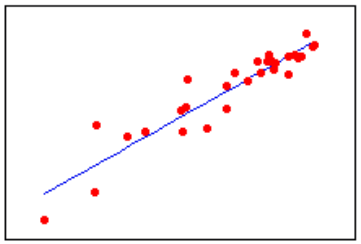


Koeficient determinace

- KD (**R^2**) ukazuje podíl rozptylu Y vysvětleného regresním modelem (SSM) vůči celkovému rozptylu Y (SST) = **SSM / SST**
- SST = SSM (vysvětlený rozptyl) + SSR (nevysvětlený rozptyl)







SSM	sum of squares of model
SSR	sum of squares of residuals
SST	sum of squares of total

Model Fit Measures

Model	R	R^2
1	0.413	0.170

KD R^2 udává, že model vysvětluje 17 % rozptylu ZP (obama2012).

V jednoduché regresi je pro standardizované proměnné (zde %) hodnota R rovna Pearsonovu *r* (nabývá však pouze kladných hodnot).

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	63.195	4.9721	12.71	< .001
relig_prot	−0.287	0.0914	−3.14	0.003

- **ZP (Y)** = % výsledek Obamy ve volbách 2012 (obama2012)
- **Prediktor (X)** = % podíl protestantů (relig_prot)
- **Průsečík (intercept)** = hodnota Y při X = 0
- **Sklon (slope)** = změna hodnoty Y při jednotkové změně X

→ při zvýšení podílu protestantů o 1 p.b. model predikuje snížení volebního zisku Obamy o -0.29 p.b.

Jsou odhady statisticky významné?

Celý model: testovací statistika F

- Abychom zjistili, zda se regresní model jako *celek* významně liší od nuly, užijeme **F-test**
- **F-test:** závisí hodnota ZP na hodnotě prediktoru?
- H0: předpoklad, že všechny koeficienty $\beta = 0$
- HA: $\beta \neq 0$

wikimedia

- $F = \frac{signal}{noise}$; $F = \frac{vysvětlený\ rozpyl}{nevysvětlený\ rozptyl}$; $F = \frac{\mathrm{SSM}}{\frac{\mathrm{SSR}}{(\mathrm{n}-2)}}$

SSM = vysvětlený rozptyl
SSR = nevysvětlený rozptyl
n = velikost výběru

- F rozdělení s 1 (čitatel) a n – 2 s.v. (jmenovatel)
- F-test pro mnohorozměrnou regresi zohledňuje počet prediktorů

F-hodnota regresního modelu je **testovací statistikou**

Fischerovo rozdělení

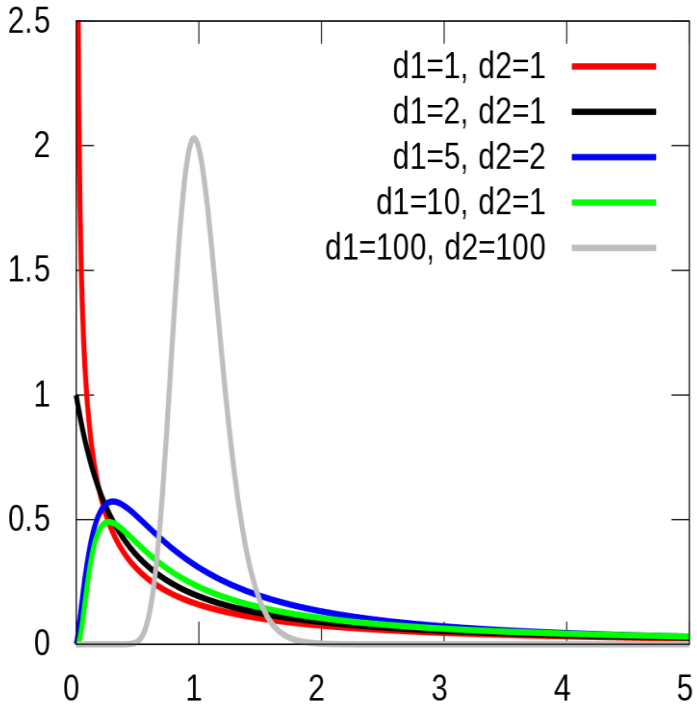


TABLE E

F critical values (continued)

Degrees of freedom in the denominator	p	Degrees of freedom in the numerator: 1	2	3	4	5	6	7	8	9
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57

Model Fit Measures

Model	R	R^2	Overall Model Test F	df1	df2	p
1	0.413	0.170	9.86	1	48	0.003

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	63.195	4.9721	12.71	< .001
relig_prot	−0.287	0.0914	−3.14	0.003

Model se statisticky významně odlišuje od nuly; při dané úrovni testu tedy odmítáme H0: $\beta_1 = B_k = 0$

Koeficienty: testovací statistika *t*

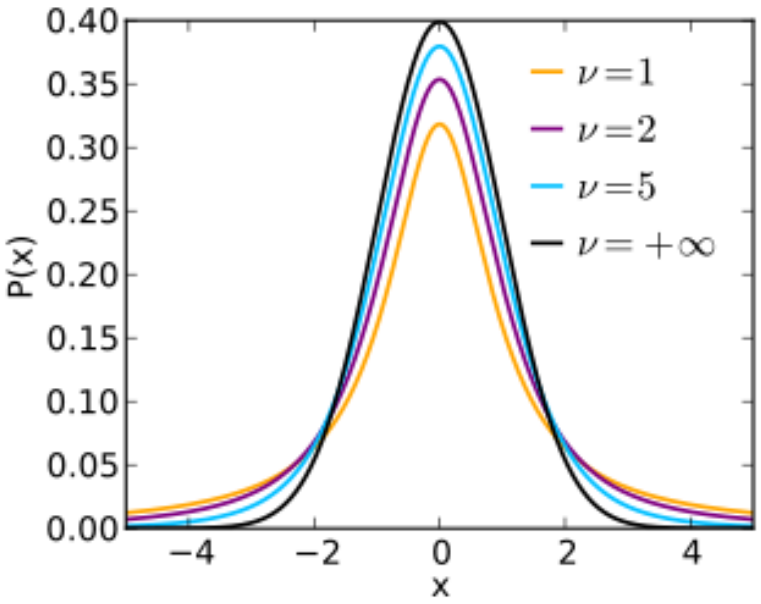
- Abychom zjistili, zda se koeficient *b* statisticky významně liší od nuly, užijeme dílčí **t-test** pro regresní koeficient
- *Je hodnota b významně odlišná od předpokládané populační hodnoty koeficientu* (0)? H0: β = 0; HA: $\beta \neq 0$

- $t = \frac{signal}{noise}$; $t = \frac{\mathrm{b} - \beta}{\mathrm{SE(b)}}$; *t* rozdělení s n – 2 stupni volnosti

  - b = koeficient sklonu (spočteno na výběrových datech)
  - β = předpokládaný populační koeficient (β = 0)
  - SE(b) = std. chyba koeficientu (viz Handl 2009)

- **t-hodnota** regresního koeficientu *b* je **testovací statistikou**

t rozdělení



	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
1	**63.657**	31.821	**12.706**	6.314	3.078
2	**9.925**	6.965	**4.303**	2.920	1.886
3	**5.841**	4.541	**3.182**	2.353	1.638
4	**4.604**	3.747	**2.776**	2.132	1.533
5	**4.032**	3.365	**2.571**	2.015	1.476
6	**3.707**	3.143	**2.447**	1.943	1.440
7	**3.499**	2.998	**2.365**	1.895	1.415
8	**3.355**	2.896	**2.306**	1.860	1.397
9	**3.250**	2.821	**2.262**	1.833	1.383
10	**3.169**	2.764	**2.228**	1.812	1.372
11	**3.106**	2.718	**2.201**	1.796	1.363
12	**3.055**	2.681	**2.179**	1.782	1.356
13	**3.012**	2.650	**2.160**	1.771	1.350
14	**2.977**	2.624	**2.145**	1.761	1.345
15	**2.947**	2.602	**2.131**	1.753	1.341
16	**2.921**	2.583	**2.120**	1.746	1.337
17	**2.898**	2.567	**2.110**	1.740	1.333
18	**2.878**	2.552	**2.101**	1.734	1.330
19	**2.861**	2.539	**2.093**	1.729	1.328
20	**2.845**	2.528	**2.086**	1.725	1.325
21	**2.831**	2.518	**2.080**	1.721	1.323
22	**2.819**	2.508	**2.074**	1.717	1.321
23	**2.807**	2.500	**2.069**	1.714	1.319

Model Fit Measures

Model	R	R^2	Overall Model Test F	df1	df2	p
1	0.413	0.170	9.86	1	48	0.003

Model se statisticky významně odlišuje od nuly; při dané úrovni testu tedy odmítáme H0: $\beta_1 = B_k = 0$

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	63.195	4.9721	12.71	< .001
relig_prot	−0.287	0.0914	−3.14	0.003

Koeficient průsečíku i koeficient sklonu jsou statisticky významné; při dané úrovni testu tedy odmítáme H0: $\beta = 0$

- **Koeficient** (Estimate) = odhady hodnot koeficientů průsečíku a sklonu
- **Std. chyba** (SE) = přesnost odhadu koeficientů (viz Handl 2009)
- **t-hodnota** (t) = testovací statistika koeficientů; t = koeficient / SE
- **p-hodnota** (p) = pravděpodobnost, že pozorujeme danou, nebo extrémnější, hodnotu testovací statistiky *t* při platnosti H0: $\beta = 0$

Shrnutí

- Lineární regresi využíváme pro **predikci hodnot spojité proměnné** (ZP) v závislosti na hodnotách jednoho či více prediktorů

- Zpravidla nás zajímá, do jaké míry **specifický prediktor ovlivňuje hodnotu ZP** – v závislosti na tom formulujeme H0 a HA (o koeficientu sklonu)

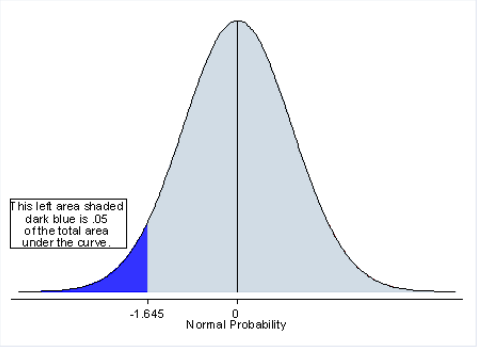
- **Prokázaný regresní vztah neimplikuje kauzalitu!**

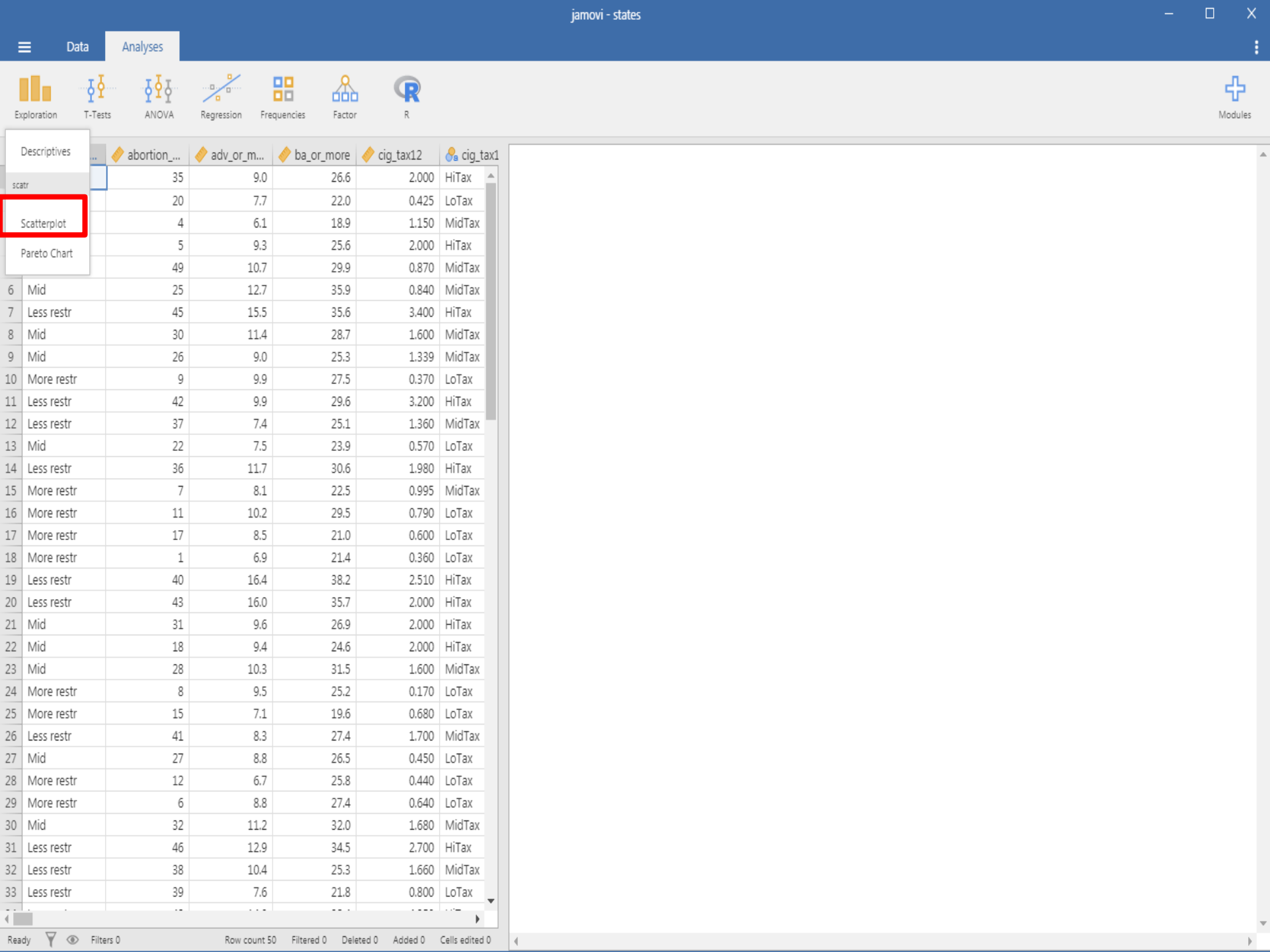
- Jako každá metoda, lineární regrese spočívá na řadě předpokladů (následující přednášky)

Jednoduchá lineární regrese v Jamovi

- **Teoretický rámec:** volební chování je do určité míry určeno sociokulturními štěpnými liniemi (Norris & Inglehart 2019; Lipset & Rokkan 1967)

- **H0:** Podíl protestantů **relig_prot** (prediktor X) neovlivňoval volební podporu Obamy **obama2012** (závislá proměnná Y); $\beta_1 >= 0$
- **HA:** Podíl protestantů **relig_prot** (prediktor X) snižoval volební podporu Obamy **obama2012** (závislá proměnná Y); $\beta_1 < 0$




jamovi - states
Data
Analyses
Exploration
T-Tests
ANOVA
Regression
Frequencies
Factor
R
Modules
Descriptives
scatr
Scatterplot
Pareto Chart
abortion_...
adv_or_m...
ba_or_more
cig_tax12
cig_tax1
1 35 9.0 26.6 2.000 HiTax
2 20 7.7 22.0 0.425 LoTax
3 4 6.1 18.9 1.150 MidTax
4 5 9.3 25.6 2.000 HiTax
5 49 10.7 29.9 0.870 MidTax
6 Mid 25 12.7 35.9 0.840 MidTax
7 Less restr 45 15.5 35.6 3.400 HiTax
8 Mid 30 11.4 28.7 1.600 MidTax
9 Mid 26 9.0 25.3 1.339 MidTax
10 More restr 9 9.9 27.5 0.370 LoTax
11 Less restr 42 9.9 29.6 3.200 HiTax
12 Less restr 37 7.4 25.1 1.360 MidTax
13 Mid 22 7.5 23.9 0.570 LoTax
14 Less restr 36 11.7 30.6 1.980 HiTax
15 More restr 7 8.1 22.5 0.995 MidTax
16 More restr 11 10.2 29.5 0.790 LoTax
17 More restr 17 8.5 21.0 0.600 LoTax
18 More restr 1 6.9 21.4 0.360 LoTax
19 Less restr 40 16.4 38.2 2.510 HiTax
20 Less restr 43 16.0 35.7 2.000 HiTax
21 Mid 31 9.6 26.9 2.000 HiTax
22 Mid 18 9.4 24.6 2.000 HiTax
23 Mid 28 10.3 31.5 1.600 MidTax
24 More restr 8 9.5 25.2 0.170 LoTax
25 More restr 15 7.1 19.6 0.680 LoTax
26 Less restr 41 8.3 27.4 1.700 MidTax
27 Mid 27 8.8 26.5 0.450 LoTax
28 More restr 12 6.7 25.8 0.440 LoTax
29 More restr 6 8.8 27.4 0.640 LoTax
30 Mid 32 11.2 32.0 1.680 MidTax
31 Less restr 46 12.9 34.5 2.700 HiTax
32 Less restr 38 10.4 25.3 1.660 MidTax
33 Less restr 39 7.6 21.8 0.800 LoTax
Ready
Filters 0
Row count 50
Filtered 0
Deleted 0
Added 0
Cells edited 0

jamovi - states

Data | Analyses

Exploration | T-Tests | ANOVA | Regression | Frequencies | Factor | R | Modules

Scatterplot

pot_policy
prochoice
prolife
relig_cath
relig_high
relig_low
religiosity3
romney2012

X-Axis: relig_prot
Y-Axis: obama2012
Group:

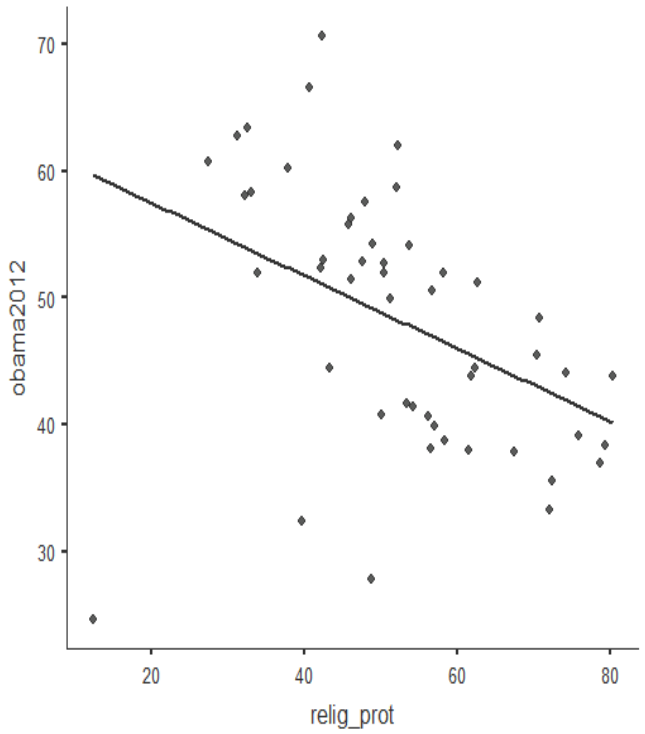
Regression Line

- None
- Linear
- Smooth
- Standard error

Marginals

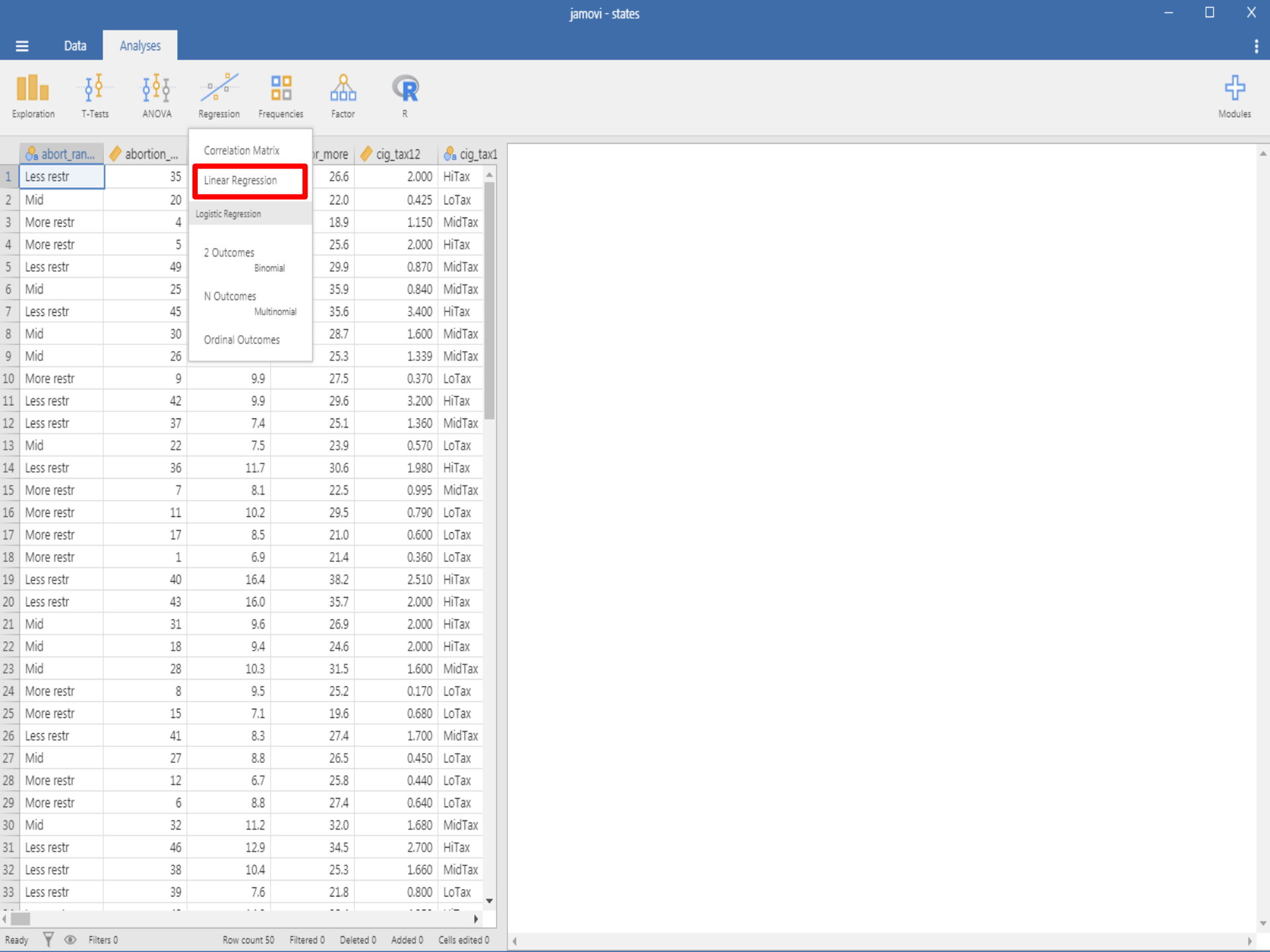
- None
- Densities
- Boxplots

Scatterplot



References

[1] The jamovi project (2019). *jamovi*. (Version 1.0) [Computer Software]. Retrieved from https://www.jamovi.org.


jamovi - states
Data
Analyses
Exploration
T-Tests
ANOVA
Regression
Frequencies
Factor
R
Modules
Correlation Matrix
Linear Regression
Logistic Regression
2 Outcomes
Binomial
N Outcomes
Multinomial
Ordinal Outcomes
abort_ran...
abortion_...
cig_tax12
cig_tax1
Less restr 35 26.6 2.000 HiTax
Mid 20 22.0 0.425 LoTax
More restr 4 18.9 1.150 MidTax
More restr 5 25.6 2.000 HiTax
Less restr 49 29.9 0.870 MidTax
Mid 25 35.9 0.840 MidTax
Less restr 45 35.6 3.400 HiTax
Mid 30 28.7 1.600 MidTax
Mid 26 25.3 1.339 MidTax
More restr 9 9.9 27.5 0.370 LoTax
Less restr 42 9.9 29.6 3.200 HiTax
Less restr 37 7.4 25.1 1.360 MidTax
Mid 22 7.5 23.9 0.570 LoTax
Less restr 36 11.7 30.6 1.980 HiTax
More restr 7 8.1 22.5 0.995 MidTax
More restr 11 10.2 29.5 0.790 LoTax
More restr 17 8.5 21.0 0.600 LoTax
More restr 1 6.9 21.4 0.360 LoTax
Less restr 40 16.4 38.2 2.510 HiTax
Less restr 43 16.0 35.7 2.000 HiTax
Mid 31 9.6 26.9 2.000 HiTax
Mid 18 9.4 24.6 2.000 HiTax
Mid 28 10.3 31.5 1.600 MidTax
More restr 8 9.5 25.2 0.170 LoTax
More restr 15 7.1 19.6 0.680 LoTax
Less restr 41 8.3 27.4 1.700 MidTax
Mid 27 8.8 26.5 0.450 LoTax
More restr 12 6.7 25.8 0.440 LoTax
More restr 6 8.8 27.4 0.640 LoTax
Mid 32 11.2 32.0 1.680 MidTax
Less restr 46 12.9 34.5 2.700 HiTax
Less restr 38 10.4 25.3 1.660 MidTax
Less restr 39 7.6 21.8 0.800 LoTax
Ready
Filters 0
Row count 50
Filtered 0
Deleted 0
Added 0
Cells edited 0

Linear Regression

obama_win12
pop2000
pop2010
pop2010_hun_thou
popchng0010
popchngpct
pot_policy
prochoice
prolife
relig_cath
relig_high
relig_low
religiosity3

Dependent Variable: obama2012

Covariates: relig_prot

Factors:

Model Builder
Reference Levels
Assumption Checks
Model Fit
Model Coefficients
Estimated Marginal Means

Linear Regression

Model Fit Measures

Model	R	R^2
1	0.413	0.170

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	63.195	4.9721	12.71	< .001
relig_prot	−0.287	0.0914	−3.14	0.003

References

[1] The jamovi project (2019). *jamovi*. (Version 1.0) [Computer Software]. Retrieved from https://www.jamovi.org.

[2] R Core Team (2018). *R: A Language and envionment for statistical computing*. [Computer software]. Retrieved from https://cran.r-project.org/.

Linear Regression

obama_win12
pop2000
pop2010
pop2010_hun_thou
popchng0010
popchngpct
pot_policy
prochoice
prolife
relig_cath
relig_high
relig_low
religiosity3

Dependent Variable: obama2012

Covariates: relig_prot

Factors:

Model Builder

Reference Levels

Assumption Checks

Model Fit

Fit Measures

- [x] R
- [x] R^2
- [ ] Adjusted R^2
- [ ] AIC
- [ ] BIC
- [ ] RMSE

Overall Model Test

- [x] F test

Model Coefficients

Linear Regression

Model Fit Measures

Model	R	R^2	Overall Model Test F	df1	df2	p
1	0.413	0.170	9.86	1	48	0.003

Model Coefficients - obama2012

Predictor	Estimate	SE	t	p
Intercept	63.195	4.9721	12.71	< .001
relig_prot	−0.287	0.0914	−3.14	0.003

References

[1] The jamovi project (2019). *jamovi*. (Version 1.0) [Computer Software]. Retrieved from https://www.jamovi.org.

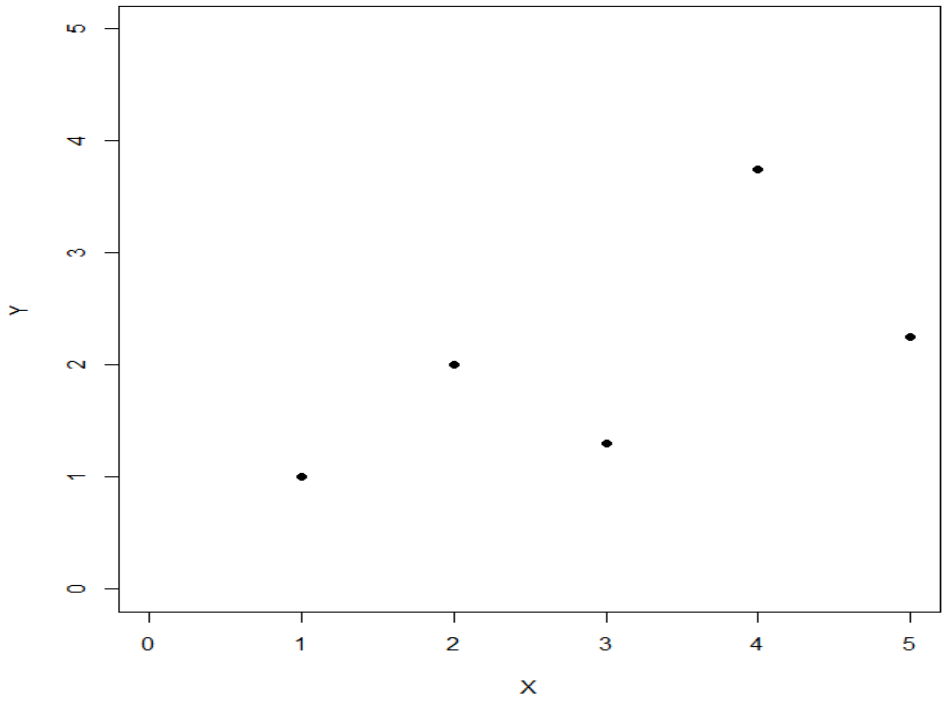
[2] R Core Team (2018). *R: A Language and envionment for statistical computing*. [Computer software]. Retrieved from https://cran.r-project.org/.

Seminář

Výpočet regresního modelu

- Máme dvě proměnné X (prediktor) a Y (ZP)

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25



- Do jaké míry X predikuje Y?

Based on Lane 2016

Výpočet regresního modelu

- Statistiky pro výpočet:

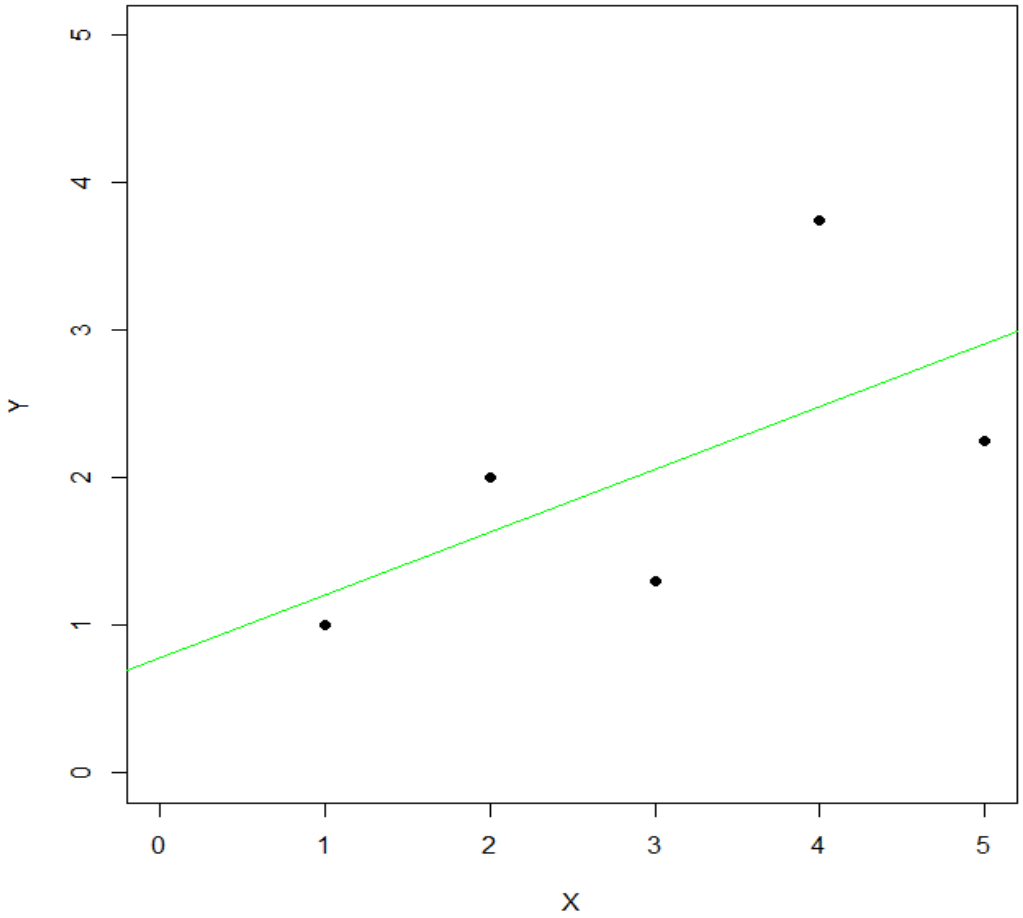
m(X)	**m(Y)**	**s(X)**	**s(Y)**	**r(X, Y)**
3	2.06	1.581	1.072	0.627

- Koeficient sklonu **(b1): r(X, Y) * s(Y) / s(X)**
- Koeficient průsečíku **(b0): m(Y) – b1*m(X)**

- **b1** = 0.627 * 1.072 / 1.581 = **0.425**
- **b2** = 2.06 – 0.425 * 3 = **0.785**

Výpočet regresního modelu

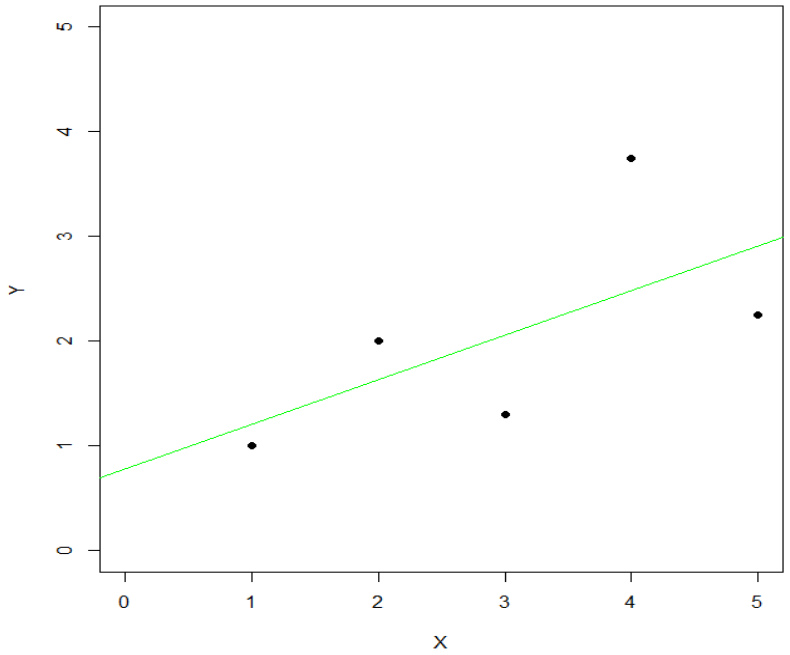
- Výsledná regresní přímka



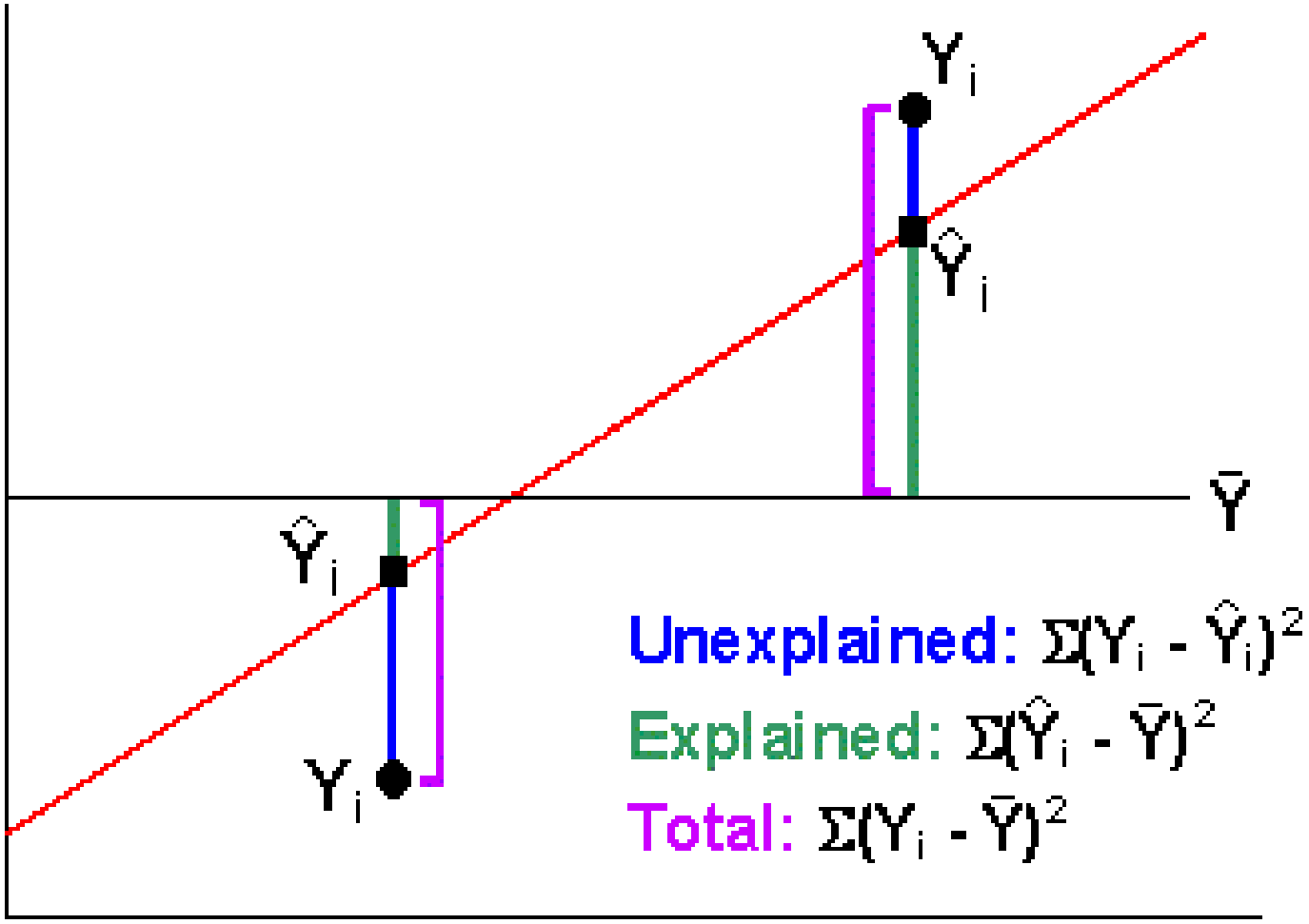
Výpočet regresního modelu

- Regresní model: Y = 0.78 + 0.425*X
- **Průsečík:** hodnota Y při hodnotě X = 0
- **Sklon:** změna hodnoty Y při jednotkové změně X
- **Součet čtverců reziduí:** chyba modelu

- **Jaká je hodnota Y' pro X = 2?**
- Y' = 0.785 + (0.425)*2
- Y' = 0.785 + 0.850 = 1.635
- **Jaká je hodnota Y' pro X = 6?**
- Y' = 0.785 + (0.425)*6
- Y' = 0.785 + 2.55 = 3.335



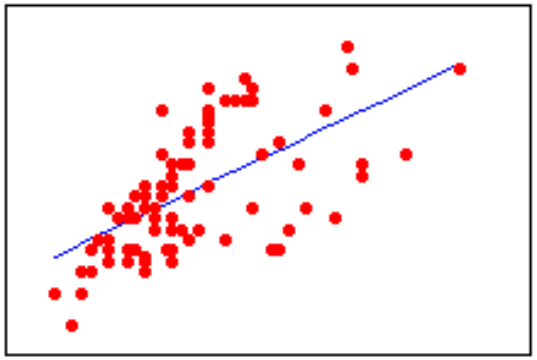
Celkový rozptyl = nevysvětlený rozptyl + vysvětlený rozptyl

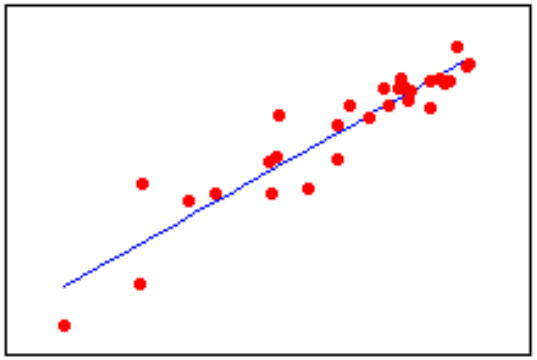


- **<u>Rozptyl vysvětlený</u>** regresním model (**SSM**: sum of squares of model)
- $\mathrm{SSM} = \Sigma\left(\mathrm{Y}' - \ \mathrm{mean}(\mathrm{Y})\right)^2$
- **<u>Nevysvětlený rozptyl</u>** – chyba regresního modelu (**SSR**: sum of squares of residuals):
- $\mathrm{SSR} = \Sigma(\mathrm{Y} - \mathrm{Y}')^2$

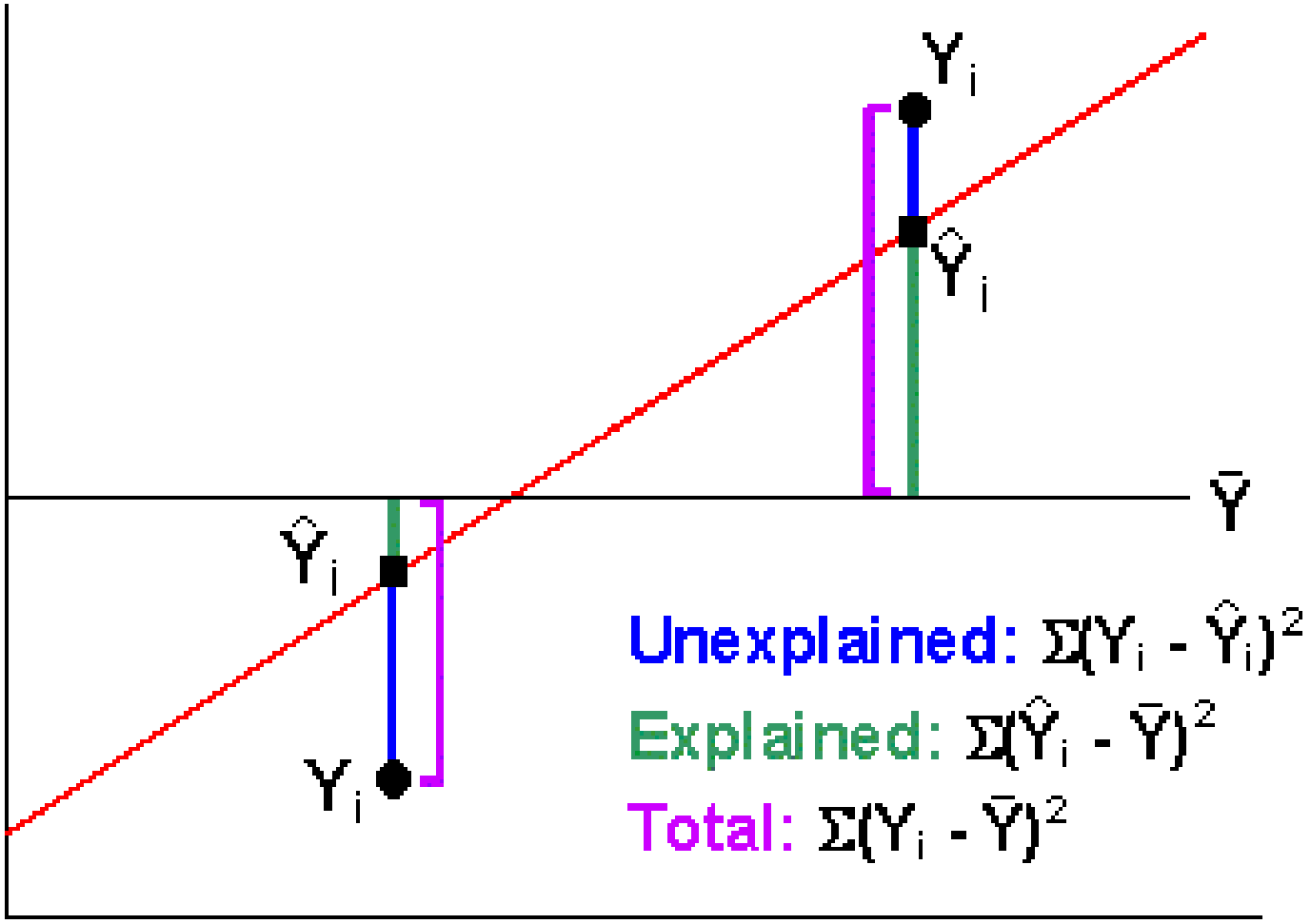
Y' = predikované hodnoty (přímka)
Y = pozorované hodnoty (body)
mean(Y) = průměrná hodnota ZP

- **<u>Celkový rozptyl</u>** (**SST**: total sum of squares) = SSM + SSR
- $\mathrm{SST} = \Sigma\left(\mathrm{Y} - \mathrm{mean}(\mathrm{Y})\right)^2$
- Podíl **SSM/SST** ukazuje **(1)** sílu regresního vztahu mezi ZP a prediktory a **(2)** přesnost predikce založené na regresním modelu (funkci)





Celkový rozptyl = **nevysvětlený rozptyl** + **vysvětlený rozptyl**



Koeficient determinace

SSM: součet čtverců rozdílů mezi predikovanými hodnotami Y’ a průměrem Y

Y’	**mean Y**	**(Y’ – mY)**	**(Y’ – mY)^2**
1.210	2.06	-0.850	0.72
1.653	2.06	-0.425	0.18
2.060	2.06	0	0
2.485	2.06	0.425	0.18
2.910	2.06	0.850	0.72
sum (SSM)			**1.81**

SSR: součet čtverců rozdílů mezi pozorovanými hodnotami Y a predikovanými hodnotami Y’

Y	**Y’**	**Y – Y’**	**(Y – Y’)^2**
1	1.210	-0.210	0.044
2	1.653	0.365	0.133
1.3	2.060	-0.760	0.578
3.75	2.485	1.265	1.600
2.25	2.910	-0.660	0.436
sum (SSR)			**2.791**

- SST = SSM + SSR = 1.81 + 2.791 = 4.59
- **R^2** = SSM / SST = 1.81 / 4.59 = 0.39 = **39 %**