Korelační a regresní analýza

Lineární regrese

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

doc. Ing. Petr Seďa, Ph.D.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co se dnes dozvíte?

- Asociace a korelace, míry závislosti.
- Modelování statistické závislosti.
- Regresní přímka, metoda nejmenších čtverců.
- Vybrané nelineární závislosti.
- Měření kvality modelu.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Dvourozměrný soubor:

datová tabulka – soubor dvojic $[xi; yi]$

X	Y
x_1	y_1
x_2	y_2
...	...
...	...
x_n	y_n

x_1, x_2 … hodnoty číselné proměnné X

y_1, y_2 … hodnoty číselné proměnné Y

n … počet prvků (rozsah) souboru

Kovariance

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Sdružené charakteristiky souboru:

- popisují vzájemný vztah obou proměnných

kovariance

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x} \cdot \bar{y}$$

výběrová kovariance

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

Kovariance

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vlastnosti kovariance:

- vyjadřuje intenzitu lineární závislosti mezi X a Y

$s_{xy} > 0$ přímá (pozitivní) závislost $X\uparrow\ Y\uparrow$

$s_{xy} < 0$ nepřímá (negativní) závislost $X\uparrow\ Y\downarrow$

$s_{xy} = 0$ lineárně nezávislé veličiny

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Korelační koeficient:

- relativní míra lineární závislosti mezi X a Y

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

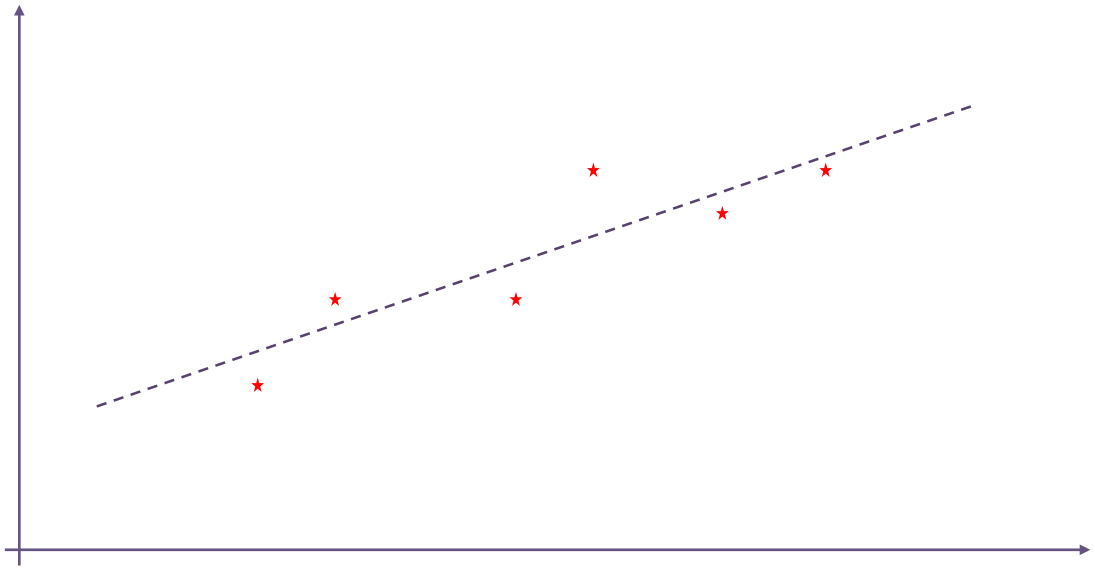
platí: $-1 \leq r_{xy} \leq +1$

$r_{xy} \to \pm 1$ silná (lineární) závislost

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co je to lineární závislost?

Pamatujete si ještě z matematiky, co to znamená, jsou-li dvě veličiny lineárně závislé?

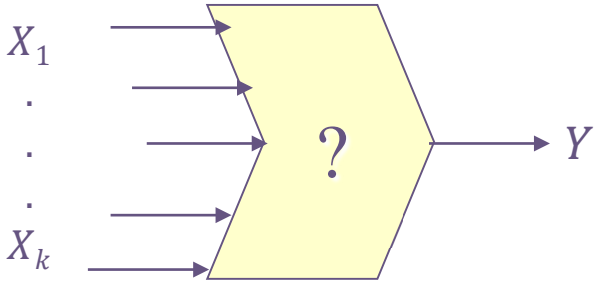


SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co je to regrese?

korelace – vzájemný vztah dvou proměnných

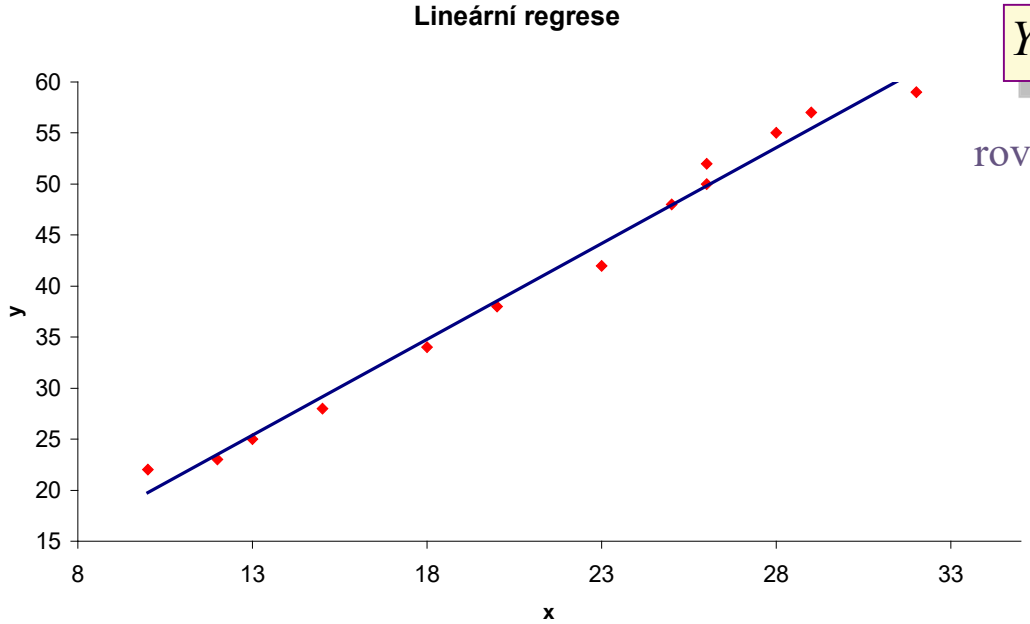
regrese – hledání matematického vztahu mezi dvěma či více proměnnými



$$Y = f(X_1, X_2, \cdots, X_k)$$ regresní funkce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jednoduchá lineární regrese:



$$Y_i = b_0 + b_1 \cdot x_i$$

rovnice regresní přímky

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Metoda nejmenších čtverců:

Jak zvolit hodnoty b_0 a b_1 ?

minimalizace součtu čtverců odchylek

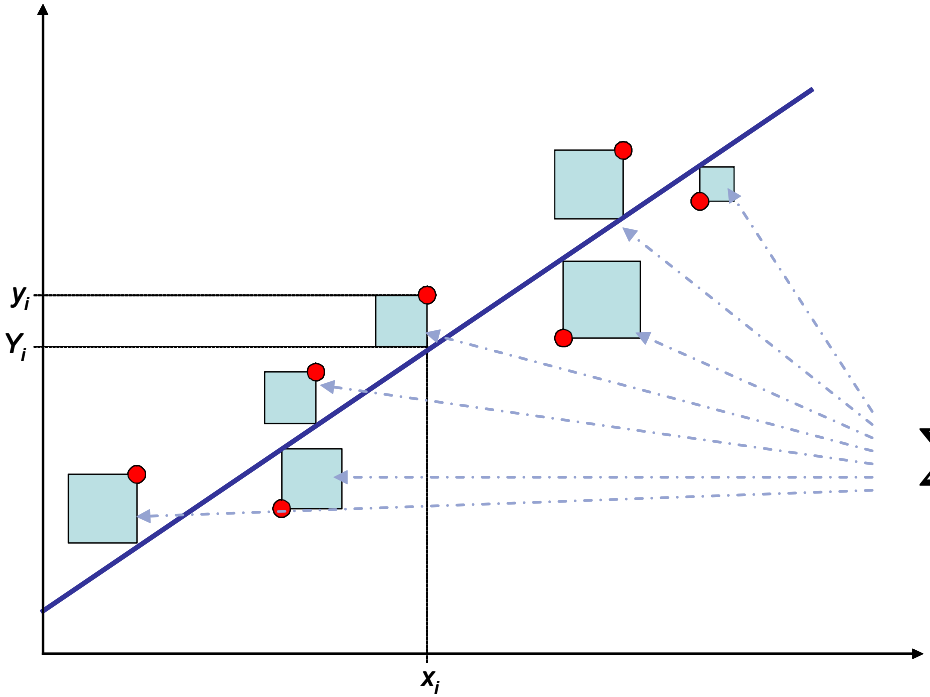
$$S_R = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

x_i … hodnota nezávislé (vstupní) proměnné
y_i … hodnota závislé (výstupní) proměnné

Y_i … teoretická (vypočtená) hodnota

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Proč metoda nejmenších čtverců?



$$\sum_i (Y_i - y_i)^2 \rightarrow \min$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jak na metodu nejmenších čtverců?

Výpočet koeficientů b_0 a b_1:

$$\frac{\partial S_R}{\partial b_0} = 0 \quad \wedge \quad \frac{\partial S_R}{\partial b_1} = 0$$

Řešení:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vlastnosti regresní přímky?

- regresní koeficient b_1
  směrnice regresní přímky
  mezní přírůstek závisle proměnné Y

$$\Delta X = 1 \quad \rightarrow \quad \Delta Y = b_1$$

- koeficient b_0
  průsečík regresní přímky s osou y

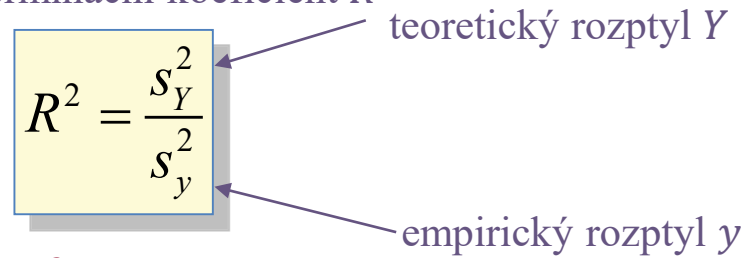
přímka prochází těžištěm $[\bar{x}\,;\,\bar{y}]$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jednoduchá lineární regrese:

- kvalita regresního modelu

determinační koeficient R^2



$0 \leq R^2 \leq 1$

- jakou část variability závislé proměnné Y lze vysvětlit vlivem nezávislé proměnné X

- pro lineární modely je determinační koeficient druhou mocninou koeficientu korelace

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Metoda nejmenších čtverců a analýza rozptylu (ANOVA):

reziduální součet čtverců S_R

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

nevysvětlená variabilita

teoretický součet čtverců S_T (regresní)

$$S_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

vysvětlená variabilita

celkový součet čtverců S_y

$$S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_y = S_T + S_R$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Tabulka ANOVA:

	součet čtverců	stupeň volnosti	rozptyl	F poměr
teoretický T	S_T	1	$s_T{}^2 = S_T$	$F = s_T{}^2/s_R^2$
reziduální R	S_R	$n - 2$	$s_R^2 = S_R/(n-2)$	
celkový y	S_y	$n - 1$		

reziduální rozptyl:

$$s_R^2 = \frac{S_R}{n-2}$$

determinační koeficient:

$$R^2 = \frac{S_T}{S_y}$$

Příklad 2 – obrat prodejny

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vedení firmy věří, že na roční obrat prodejny má vliv počet obyvatel v místě. Proto provedlo průzkum v 10 vybraných lokalitách:

počet obyvatel (tis.)	5	7	8	14	17	23	23	28	29	35
roční obrat (mil. Kč)	3,8	2,1	3,6	5,2	7,6	7,0	9,2	7,5	10,1	11,4

a) Co je nezávislou proměnnou a co je závislou proměnnou?

b) Určete rovnici regresní přímky, víte-li, že kovariance mezi znaky $s_{xy} = 26{,}325$, průměrná hodnota počtu obyvatel ve všech lokalitách je 18,9 tis., průměrný výše obratu na pobočku je 6,75 mil. Kč a rozptyl počtu obyvatel je 95,89 tis^2, interpretujte vypočtené koeficienty regresní přímky.

c) Vypočtěte determinační koeficient.

d) Odhadněte obrat prodeje v obci s 12 tis. obyvateli.

Příklad 2 – obrat prodejny

Ad a) Nezávisle proměnnou je počet obyvatel ve městě a závislou výše obratu pobočky.

Ad b) Koeficienty regresní přímky:

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{26,325}{95,89} = 0,2745$$

$$b_0 = \bar{y} - b_1\bar{x} = 6,75 - 0,2745 \cdot 18,9 = 1,562$$

Regresní přímka:

$$y = 1,562 + 0,2745x$$

Při zvýšení počtu obyvatel o 1 tisíc vzroste obrat o 0,2745 mil. Kč.

Průměrná výše obratu na pobočce bez ohledu na počet obyvatel.

Příklad 2 – obrat prodejny

Ad c) K výpočtu reziduí a součtů čtverců potřebujeme spočítat „modelové“ hodnoty Y_i dosazením počtu obyvatel do vypočtené regresní přímky:

počet obyvatel (tis.)	5	7	8	14	17	23	23	28	29	35
roční obrat (mil. Kč)	3,8	2,1	3,6	5,2	7,6	7,0	9,2	7,5	10,1	11,4
obrat podle modelu	2,93	3,48	3,76	5,41	6,23	7,88	7,88	9,25	9,52	11,17

$$S_R = \sum (y_i - \hat{y}_i)^2 = 10{,}57$$

$$S_y = \sum (y_i - \bar{y})^2 = 82{,}82$$

$$S_T = S_y - S_R = 72{,}25$$

$$R^2 = \frac{S_T}{S_y} = \frac{72{,}25}{82{,}82} = 0{,}872$$

Přímka vysvětluje variabilitu y z 87,2% jako vliv proměnné x.

Příklad 2 – obrat prodejny

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Ad d) Bodový odhad pro $x = 12$:

$$y = b_0 + b_1 x = 1{,}562 + 0{,}2745 \cdot 12 = 4{,}856$$

V obci s 12 tis. obyvateli lze očekávat roční obrat ve výši 4,86 mil. Kč.

Literatura

1. Janáček J. *Statistika jednoduše*. Praha: Grada, 2022. **(kapitola 4, 6 a 7).**

Děkuji za pozornost.