

Regrese – jednoduchá regrese

- Cíl jednoduché (simple) regrese: najít model funkční závislosti (spojité) veličiny Y na jedné (spojité) veličině (na tzv. regresoru) X

- model lineární $Y = \beta_0 + \beta_1 X$

- kvadratický $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

(tvar často napoví bodový graf dat)

- *Příklad: závislost hmotnosti jedince na jeho tělesné výšce*

Regrese – vícenásobná regrese

- Cíl vícenásobné (multiple) regrese:
najít model funkční závislosti
(spojité) veličiny Y na více (spojitých)
veličinách (regresorech)
- např. model lineární se 2 regresory:
$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$
- *Příklad: závislost platu (Y) na věku
zaměstnance (X) a délce jeho praxe
v oboru (Z)*

Jednoduchá regrese

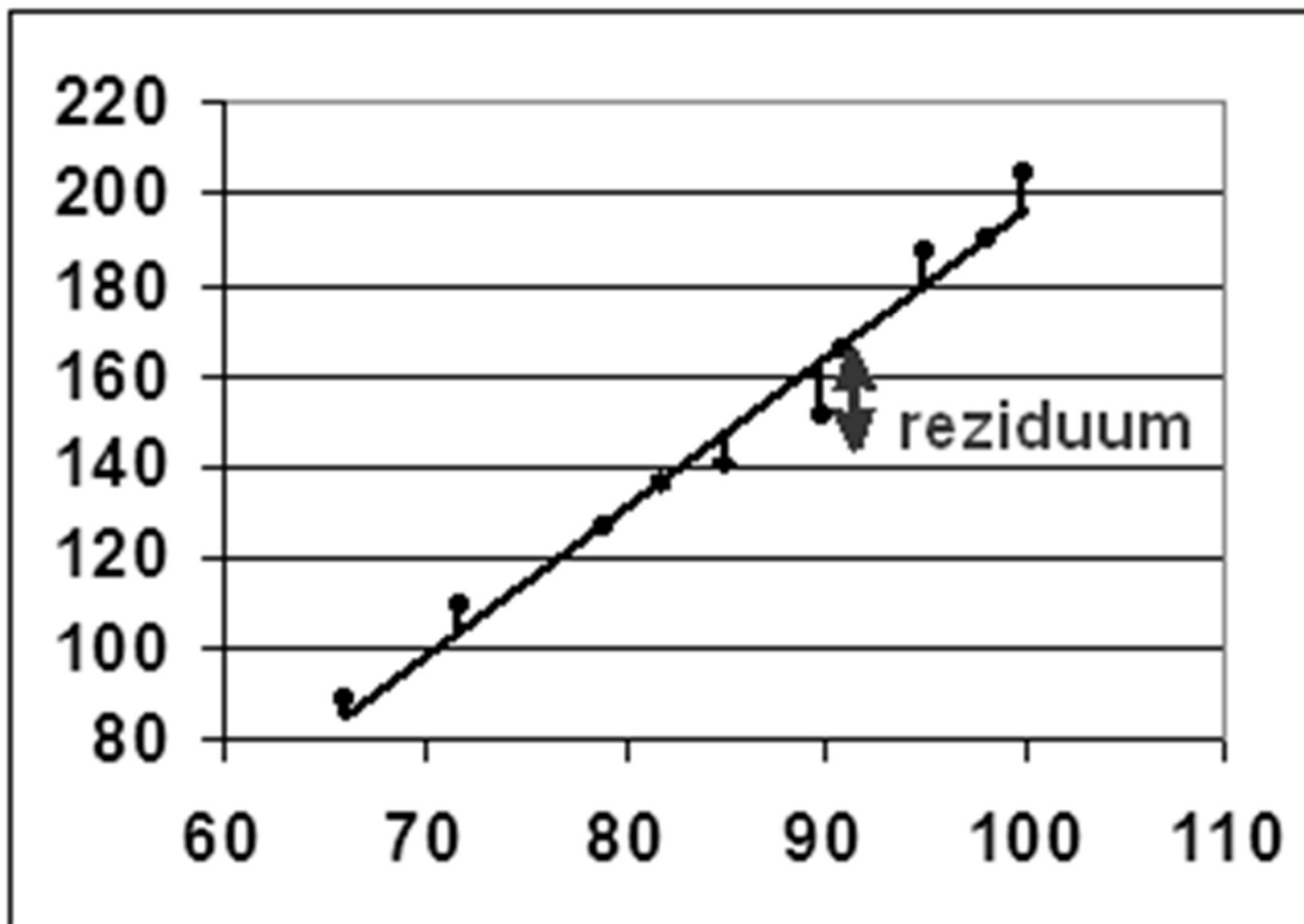
■ Značení:

$(x_i ; y_i)$ $i=1, \dots, n$ data

Y_i $i=1, \dots, n$ model

$e_i = y_i - Y_i$ $i=1, \dots, n$ reziduum

Regrese – bodový graf



Jednoduchá lineární regrese

$$y_1 = Y_1 + e_1 = (\beta_0 + \beta_1 \cdot x_1) + e_1$$

$$y_2 = Y_2 + e_2 = (\beta_0 + \beta_1 \cdot x_2) + e_2$$

...

$$y_n = Y_n + e_n = (\beta_0 + \beta_1 \cdot x_n) + e_n$$

β_0 parametr – prostý člen
(průsečík grafu přímky s o_Y)

β_1 parametr – lineární člen
(směrnice grafu přímky)

Jednoduchá lineární regrese aneb MATICOVĚ:

$$\mathbf{y} = \mathbf{F} \cdot \boldsymbol{\beta} + \mathbf{e}$$

kde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

$$\mathbf{F} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}$$

Jednoduchá lineární regrese

Odhad neznámých parametrů β_0, β_1 ?

Zjednodušení: $\mathbf{e}$ je „nulový“ vektor;

takto nalezené odhady budeme

značit b_0, b_1 , vektorově jako vektor $\mathbf{b}$;

řešíme tedy maticovou rovnici:

$$\mathbf{y} = \mathbf{F} \cdot \mathbf{b} \quad \text{aneb} \quad \mathbf{F} \cdot \mathbf{b} = \mathbf{y}$$

(kde $\mathbf{y}$, $\mathbf{F}$ známe) s cílem určit $\mathbf{b}$.

Jednoduchá lineární regrese

$$\mathbf{F} \cdot \mathbf{b} = \mathbf{y} \longrightarrow \mathbf{b} = ?$$

Pozor – $\mathbf{F}$ je matice, nelze s ní dělit!

„Trikové“ úpravy (vlastnosti matic):

$$\mathbf{F}^T \mathbf{F} \cdot \mathbf{b} = \mathbf{F}^T \mathbf{y}$$

$$(\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{F} \cdot \mathbf{b} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{y}$$

$$\underline{\underline{\mathbf{b} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{y}}}$$

Jednoduchá lineární regrese

$$\mathbf{b} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{y}$$

u modelu jednoduché lineární regrese:

$$\mathbf{F} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

$$\mathbf{F}^T \mathbf{F} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

Jednoduchá lineární regrese

- Lze odvodit /těžké/ „verzi se sumami“:

$$b_1 = (n \sum x_i y_i - \sum x_i \sum y_i) / [n \sum x_i^2 - (\sum x_i)^2]$$

$$b_0 = (\sum y_i - b_1 \sum x_i) / n$$

Jednoduchá lineární regrese

■ Přehled vzorců (verze s průměry):

$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$	$b_0 = \bar{y} - b_1 \bar{x}$	$\vec{b} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \vec{y}$
$Q_e = \sum (e_i)^2$	$Q_y = \sum (y_i - \bar{y})^2$	$I^2 = 1 - Q_e / Q_y$

k čemu Q_e (součet reziduálních čtverců)?

Regrese (dokonce každý typ)

$$\mathbf{b} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{y}$$

jde o univerzální (pro každý regresní model!) vzorec odhadu parametrů $\mathbf{b}$, modely se liší jen konkr. tvarem $\mathbf{b}$ a $\mathbf{F}$;

jde o tzv. odhad metodou nejmenších čtverců (MNČ): zaručuje $\min \sum (e_i)^2$

Jednoduchá lineární regrese-MNČ

- jiné odvození b_0 , b_1 :

$$Q_e = \sum (e_i)^2 = \sum (y_i - Y_i)^2 = \sum (y_i - \beta_0 - \beta_1 \cdot x_i)^2$$

b_0 , b_1 bude značit řešení problému najít hodnoty pro β_0 , β_1 tak, aby hodnota Q_e byla (při daných x_i , y_i) MINIMÁLNÍ,

aneb řešení soustavy rovnic:

$$\delta Q_e / \delta \beta_0 = 0$$

$$\delta Q_e / \delta \beta_1 = 0$$

Jednoduchá lineární regrese-MNČ

- Derivace mají tvar:

$$\delta Q_e / \delta \beta_0 = 2 \cdot \sum (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (-1)$$

$$\delta Q_e / \delta \beta_1 = 2 \cdot \sum (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (-x_i)$$

tj. po dosazení do soustavy:

$$2 \cdot \sum (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (-1) = 0 \quad |:2$$

$$2 \cdot \sum (y_i - \beta_0 - \beta_1 \cdot x_i) \cdot (-x_i) = 0 \quad |:2; \text{ roznásobit}$$

Jednoduchá lineární regrese-MNČ

$$\sum (-y_i + \beta_0 + \beta_1 \cdot x_i) = 0$$

$$\sum (-x_i y_i + \beta_0 \cdot x_i + \beta_1 \cdot x_i^2) = 0$$

aneb po rozdělení Σ na 3 sumy, využití,
že $\Sigma \beta_0 = n \cdot \beta_0$ a po převedení vždy první
sumy (se znaménkem $-$) zleva doprava:

$$n \cdot \beta_0 + \beta_1 \cdot \Sigma x_i = \Sigma y_i$$

$$\beta_0 \cdot \Sigma x_i + \beta_1 \cdot \Sigma x_i^2 = \Sigma x_i y_i$$

Jednoduchá lineární regrese-MNČ

Řešením této soustavy je opět:

$$b_1 = (n\sum x_i y_i - \sum x_i \sum y_i) / [n\sum x_i^2 - (\sum x_i)^2]$$

$$b_0 = (\sum y_i - b_1 \sum x_i) / n$$

Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

(Neter-Wasserman-Kutner, USA, 1990)

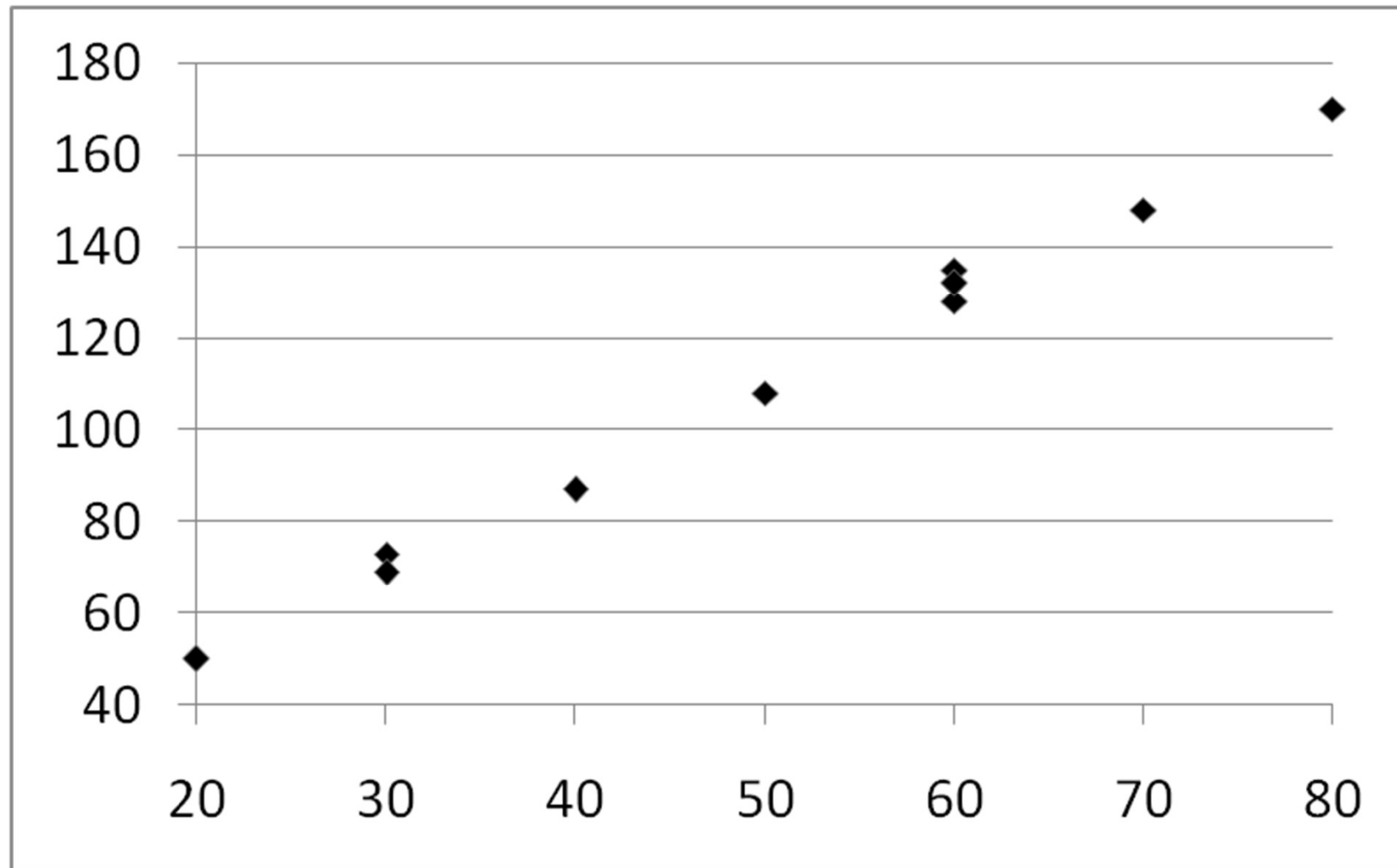
X...velikost staveníště

Y...počet hodin, odpracovaných dělníky

x_i	30	20	60	80	40	50	60	30	70	60
y_i	73	50	128	170	87	108	135	69	148	132

Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company



Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

$$\bar{x}=50, \bar{y}=110, \overline{x^2}=2840, \overline{y^2}=13466,$$

$$\overline{xy}=(30 \cdot 73 + \dots + 60 \cdot 132)/10=6180$$

$$b_1=(6180-50 \cdot 110)/(2840-50^2)=\underline{\underline{2,0}}$$

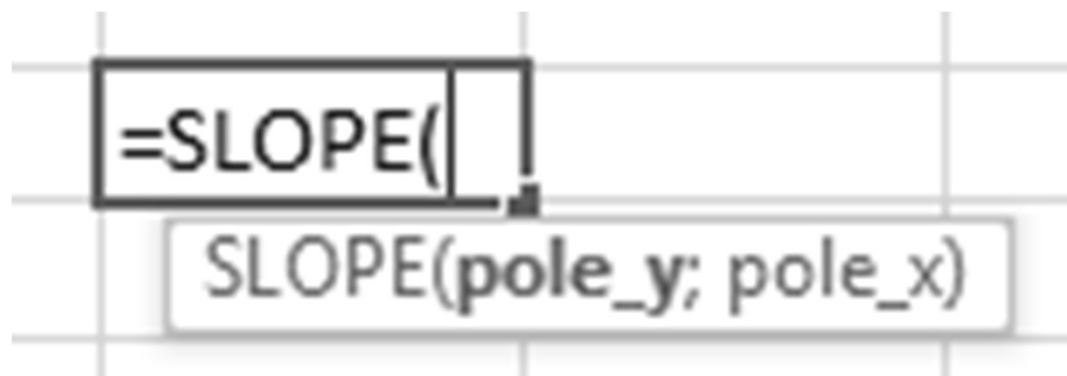
$$b_0=110-2 \cdot 50=\underline{\underline{10,0}} \quad (\text{ne vždy celočíselně})$$

Nalezený model: $Y=10+2X$

Jednoduchá lineární regrese-příklad

V Excelu:

$b_1 =$

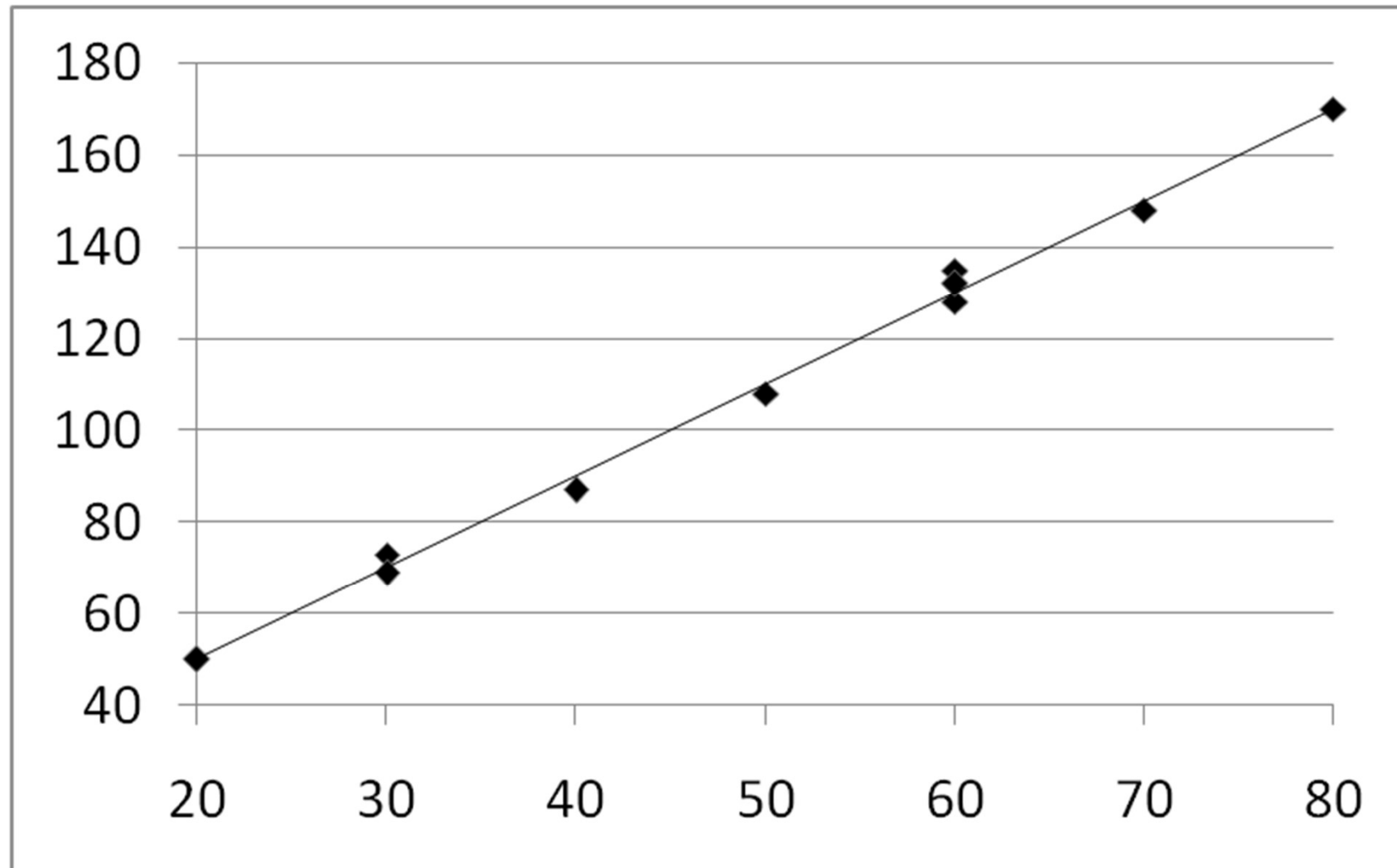


$b_0 =$



Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company



Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

Interpretace výsledku (konkr. b_1):

Při jednotkovém nárůstu velikosti staveniště vzrůstá potřebný počet odpracovaných hodin (v průměru)

o 2. (*Pozn.: Co když $b_1 < 0$?*)

Ad b_0 : Při $X=0$ (co to je?)... $Y=10$.

Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

Určete pro nalezený model Q_e :

$$Y_1 = 10 + 2 \cdot 30 = 70, \quad e_1 = 73 - 70 = 3$$

...

$$Y_{10} = 10 + 2 \cdot 60 = 130,$$

$$e_{10} = 132 - 130 = 2$$

$$Q_e = 3^2 + 0^2 + (-2)^2 + \dots + 2^2 = \underline{\underline{60}}$$

A k čemu dál využít tuto hodnotu?

Korelovanost je obecně

míra lineární závislosti

V každém typu regresního modelu
lze určit tzv. index determinace:

$$I^2 = 1 - Q_e / Q_Y$$

kde $Q_Y = \sum (y_i - \bar{y})^2$

Korelovanost

Př: Data - Westwood Company

Určete pro nalezený model (pro nějž vyšlo $Q_e=60$) hodnotu I^2 :

$$Q_Y = (73-110)^2 + \dots + (132-110)^2 = \\ = 1369 + \dots + 484 = 13660;$$

$$I^2 = 1 - 60/13660 = 1 - 0,004 = \underline{\underline{0,996}}$$

Korelovanost

Př: Data - Westwood Company

*Interpretace: Nalezený model
($Y=10+2X$) vysvětluje z 99,6 %
variabilitu proměnné Y*

ANEK

*jde o model velmi silné závislosti
proměnné Y na proměnné X.*

Korelace

- Korelace obecně je míra kvality (vhodnosti, těsnosti) nalezeného regresního modelu pro daná data; vychází z hodnot reziduí
- V každém typu regresního modelu lze použít index determinace I^2 (0 až 1, resp. 0 % až 100 %); vyjadřuje, z kolika % je variabilita závisle proměnné (Y) vysvětlena daným modelem

Korelace spec. pro model jednoduché lineární regrese

- Korelační koeficient (verze se sumami,
resp. s průměry):

$$r = \frac{(n \sum x_i y_i - \sum x_i \sum y_i)}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2] \cdot [n \sum y_i^2 - (\sum y_i)^2]}}$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2) \cdot (\overline{y^2} - \bar{y}^2)}}$$

← „*kovariance*“

Korelace spec. pro model jednoduché lineární regrese

- Korelační koeficient
- vždy v rozmezí -1 až +1 (NE % !)
- záporný při “klesající regresní přímce”
- kladný při “rostoucí regresní přímce”
- čím DÁL od 0, tím silnější je lineární závislost („korelovanost“) mezi X a Y
- platí: $r^2 = |^2$

Korelace spec. pro model jednoduché lineární regrese

Př: Data - Westwood Company

$$r = (6180 - 50 \cdot 110) / \sqrt{(2840 - 50^2) \cdot (13466 - 110^2)} =$$
$$= \underline{\underline{0,998}} \quad (\text{platí: } 0,998^2 = r^2 = 0,996)$$

Silná přímá* lineární závislost počtu
prac. hodin na velikosti staveniště.

* tj. dle „rostoucí přímky“ (nepřímá=?)

Jednoduchá lineární regrese-příklad

V Excelu:

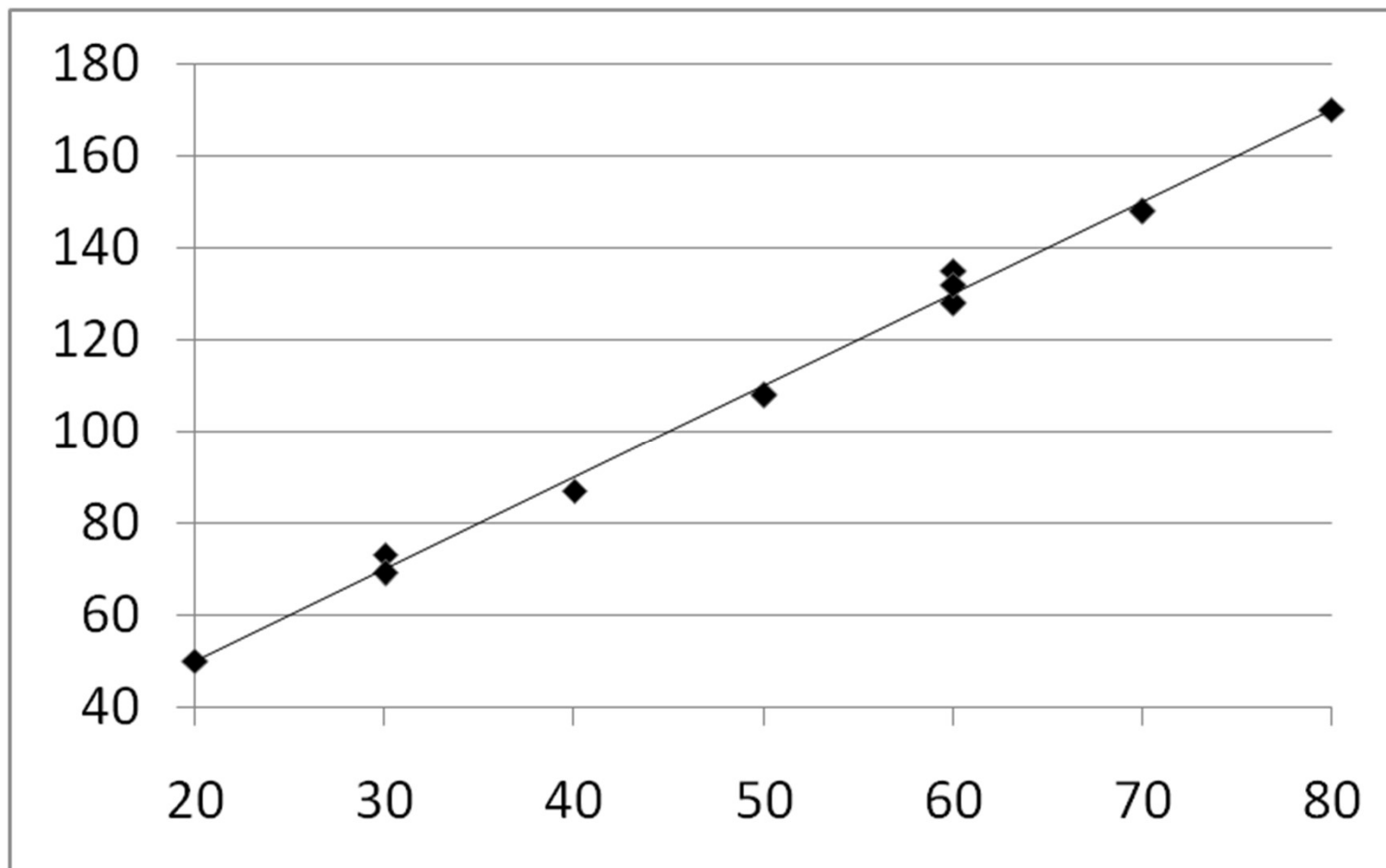
$r =$

=CORREL(

CORREL(matice1; matice2)

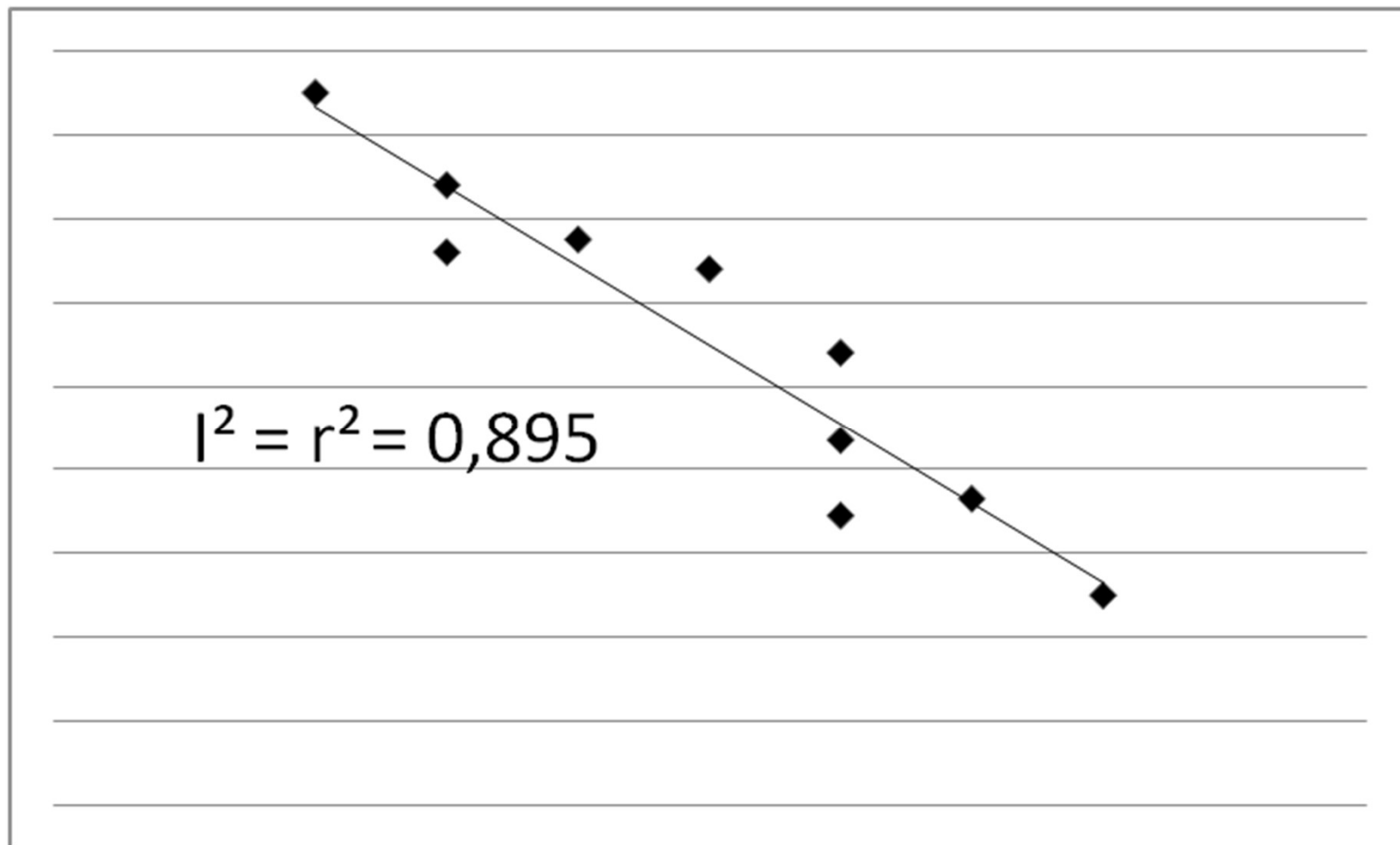
Korelace spec. pro model jednoduché lineární regrese

Př: Data Westwood Company ($r=0,998$)



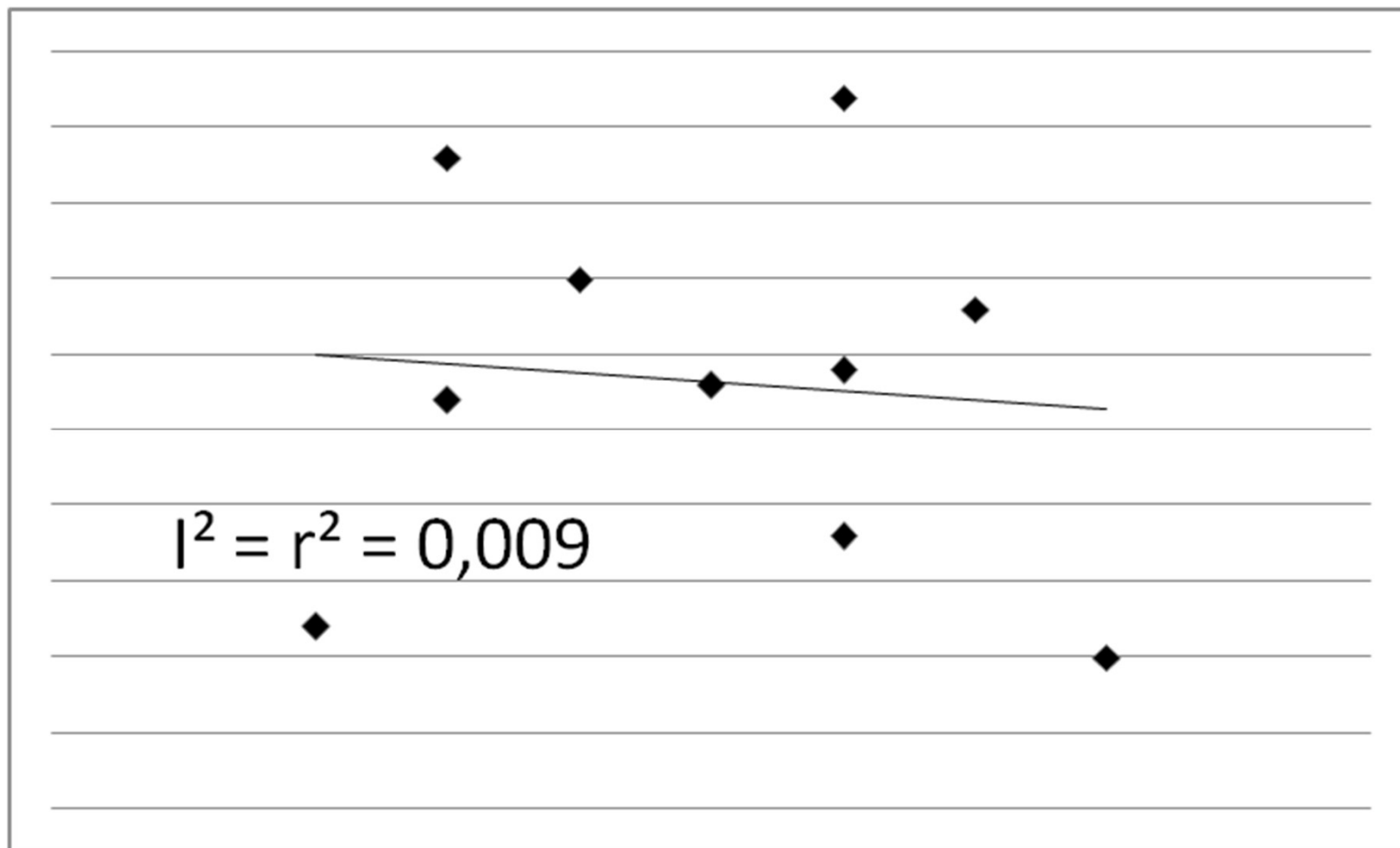
Korelace spec. pro model jednoduché lineární regrese

Př: Jiná data ($r = -0,946$)



Korelace spec. pro model jednoduché lineární regrese

Př: Jiná data ($r = -0,098$)



Korelace spec. pro model jednoduché lineární regrese

Př: Jiná data ($r = 0,075$)

