XII. Metody stanovení závislosti:

Regresní analýza

Jaké a k čemu jsou metody stanovení závislosti

- závislostí 1. **kvantitativního** znaku na 2. **kvantitativním** znaku
  (nebo více kvantitativních znacích) - **regresní a korelační analýza**
- závislost dvou znaků **- jednoduchá regresní analýza** (**jednoduchá korelační analýza**)
- závislost znaku na více znacích - **vícenásobná regresní analýza**
- znalost závislostí umožňuje:

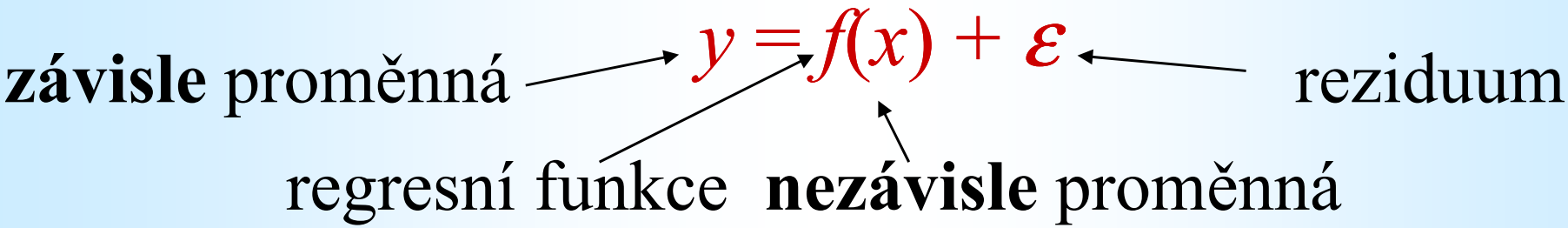
  předvídat chování (prognózovat, predikovat) závislé veličiny

Příklad 1. Zisk z reklamy 1

nezávislá - *závislá* veličina (proměnná)

Firma č.	Výdaje na reklamu (tis. Kč)	Zisk z prodeje (10 tis. Kč)
1	6	5
2	8	8
3	9	9
4	9	12
5	12	21
6	15	25
7	16	32
8	20	36
9	22	51
10	23	59

Jednoduché regresní modely



Lineární regresní funkce:

$$f(x) = \beta_0 + \beta_1 x$$

Parabolická regresní funkce :

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Exponenciální a logaritmická regresní funkce :

$$f(x) = \beta_0 \beta_1^x \qquad f(x) = \beta_0 + \beta_1 \log x$$

Jednoduchá lineární regrese (JLR)

- výběr párových hodnot:
  $(y_1, x_1), (y_2, x_2), (y_3, x_3),...,(y_n, x_n)$
- 2 způsoby získání dat:

(A) hodnoty nezávisle proměnné x_i se předem pevně zvolí a k nim se „změří“ příslušné hodnoty y_i

(B) hodnoty (y_i, x_i) se „změří“ na n náhodně zvolených jednotkách základního souboru

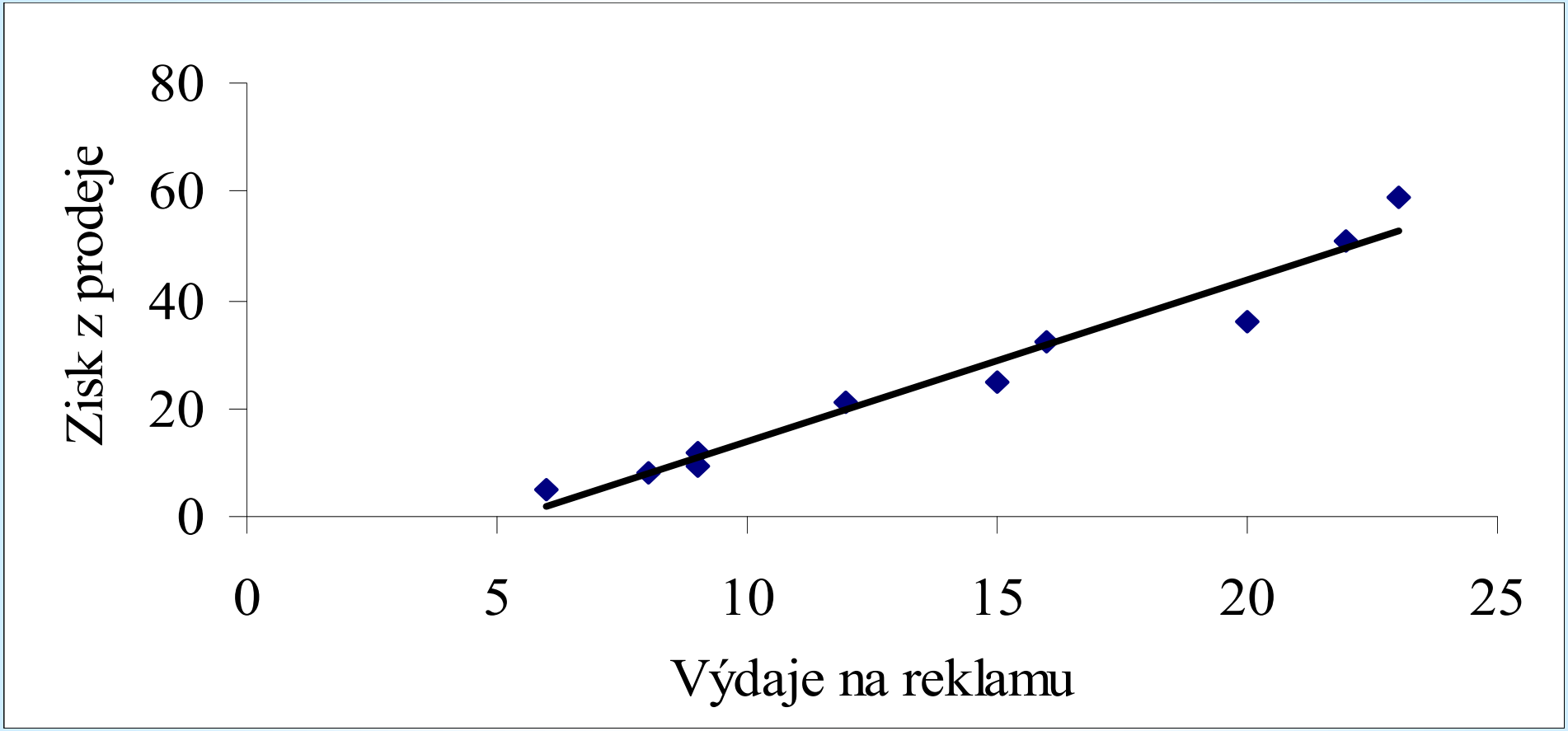
soubor párových hodnot se geometricky znázorní v rovině **bodovým grafem:**

reziduum

JLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1,2,...,n$

regresní koeficienty a jejich odhady $\boldsymbol{b}_0, \boldsymbol{b}_1$

Příklad 1. Zisk z reklamy

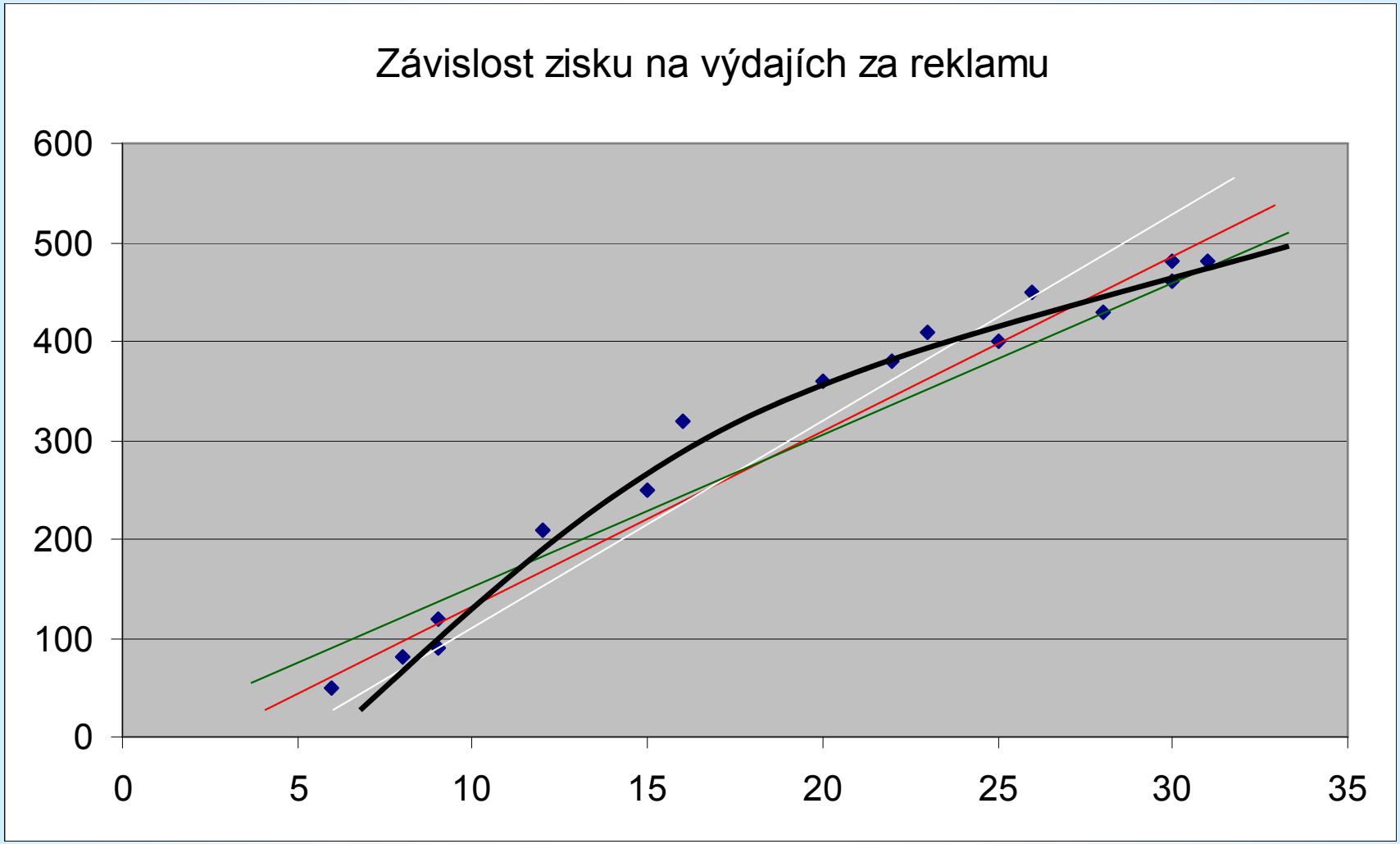


Příklad 2: Výdaje na reklamu

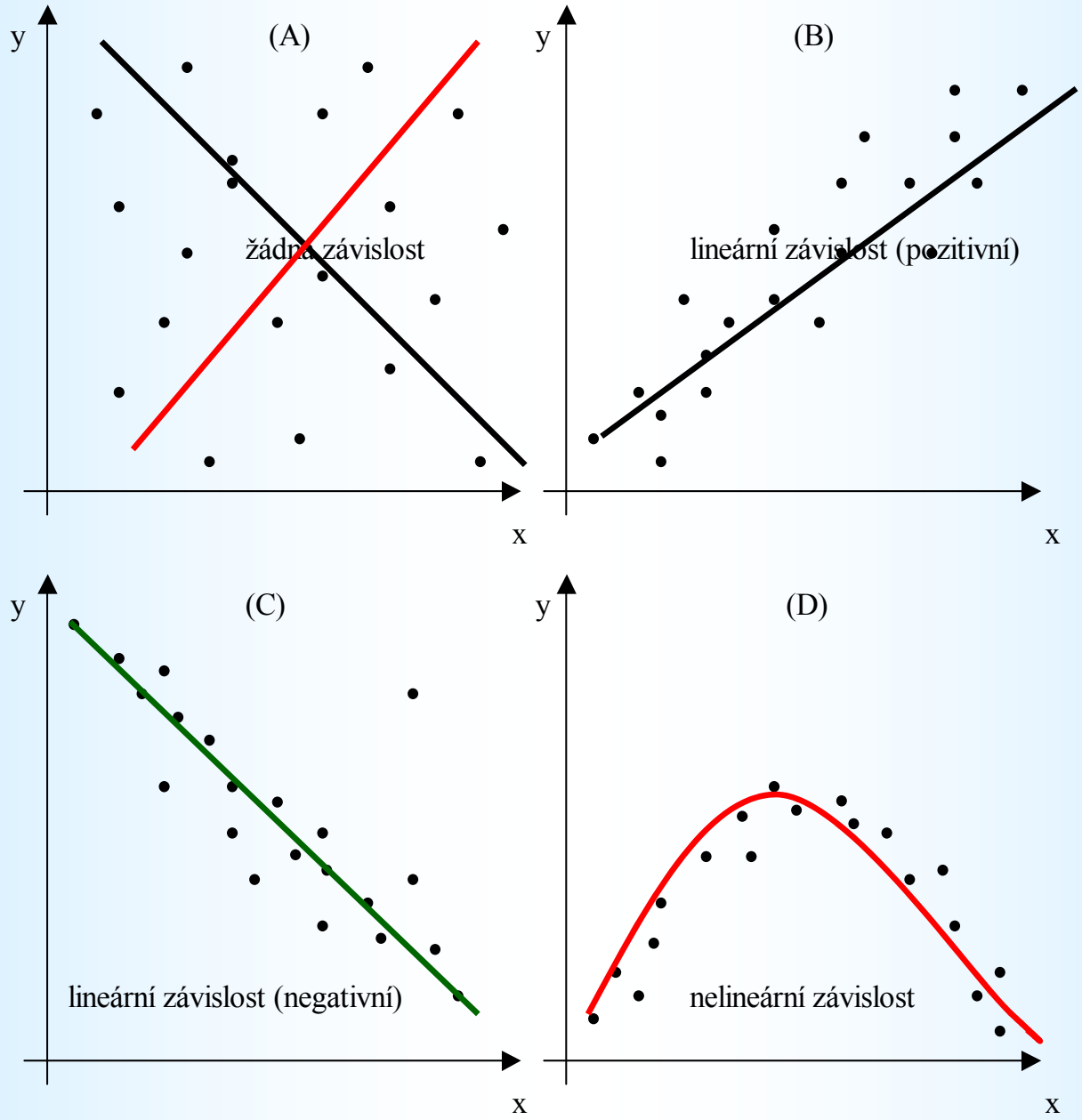
JRA

č. firmy	Výdaje na reklamu	Výdaje na reklamu	Zisk
1	malé	6	50
2	malé	8	80
3	malé	9	90
4	malé	9	120
5	středně velké	12	210
6	středně velké	15	250
7	středně velké	16	320
8	středně velké	20	360
9	středně velké	22	380
10	středně velké	23	410
11	velké	25	400
12	velké	26	450
13	velké	28	430
14	velké	30	460
15	velké	30	480
16	velké	31	480

Příklad 2: Grafické znázornění



Bodový diagram (Scatter diagram)



Metoda nejmenších čtverců MNČ

Idea MNČ: minimalizovat reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n}(y_i - Y_i)^2 = \sum_{i=1}^{n}(y_i - \underbrace{(b_0 + b_1 x_j)}_{Y_i})^2$$

Příklad 1:(pokračování)

$$b_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{462{,}1 - 14\cdot 25{,}8}{230 - 14^2} = \frac{100{,}9}{34} = 2{,}97$$

$$b_0 = \bar{y} - b_1\bar{x} = 25{,}8 - 2{,}97\cdot 14 = -15{,}78$$

Regresní funkce: $Y = -15{,}78 + 2{,}97x$

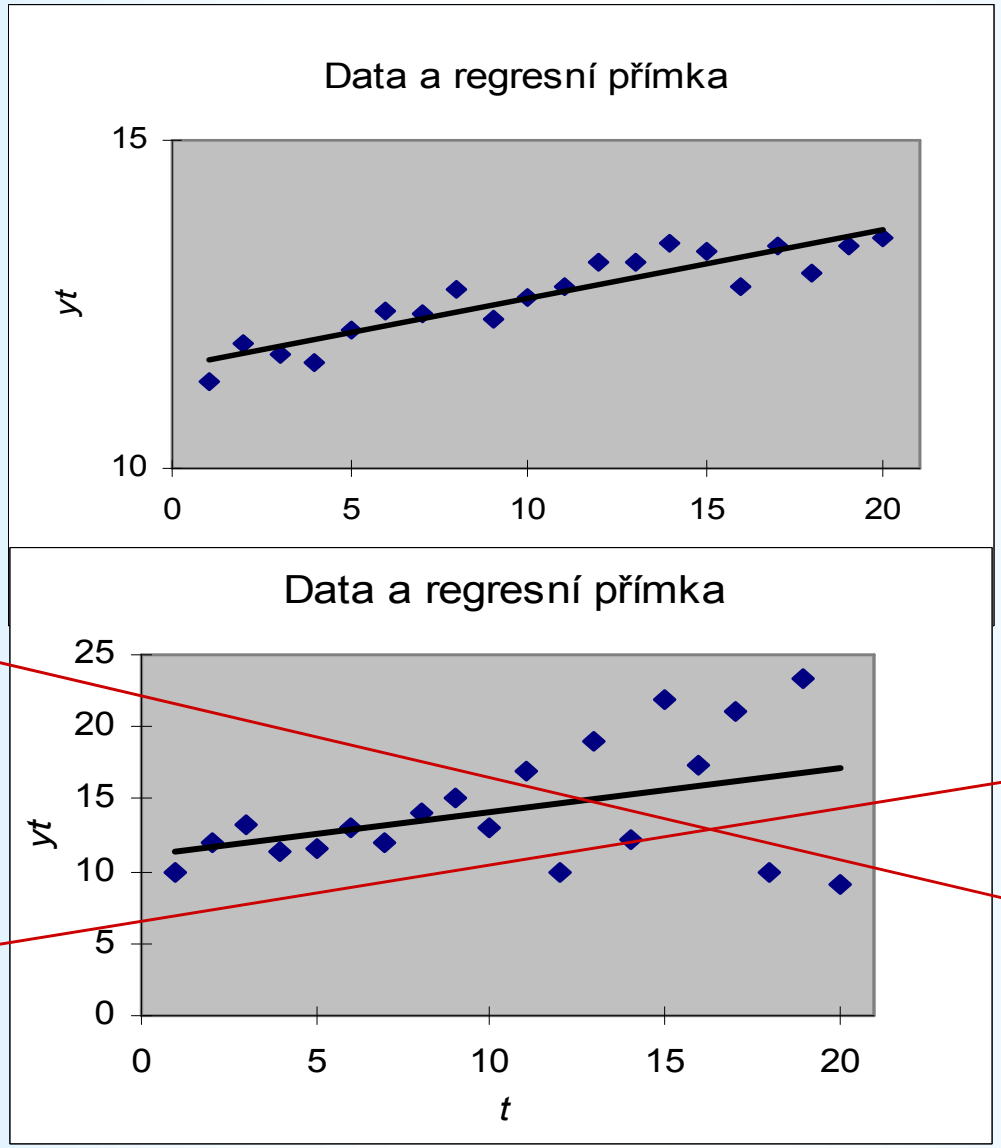
Příklad 1: „Ruční výpočty“

i	x_i	y_i	x_i^2	$x_i y_i$	Y_i	$(Y_i - \bar{y})^2$	$(y_i - \bar{y})^2$
1	6	5	36	30	2,04	565,21	432,64
2	8	8	64	64	7,98	318,22	316,84
3	9	9	81	81	10,95	221,15	282,24
4	9	12	81	108	10,95	221,15	190,44
5	12	21	144	252	19,86	35,62	23,04
6	15	25	225	375	28,77	8,61	0,64
7	16	32	256	512	31,74	34,84	38,44
8	20	36	400	720	43,62	315,88	104,04
9	22	51	484	1122	49,56	562,08	635,04
10	23	59	529	1357	52,53	711,60	1102,24
Součet	140	258	2300	4621	258	2994,3	3125,6
Průměr	14	25,8	230	462,1			

Předpoklady (klasického) lineárního modelu

1. Hodnoty vysvětlující proměnné x_i se volí předem, ***nejsou*** to tedy náhodné veličiny.
2. Náhodné složky (rezidua) ε_i mají **normální rozdělení** pravděpodobnosti se střední hodnotou 0 a (neznámým) konstantním rozptylem σ^2 - tzv. **homoskedasticita**
3. Náhodné složky jsou **nekorelované**, tj. $\rho(\varepsilon_i , \varepsilon_j) = 0$ pro každé $i \neq j$, $i,j = 1,2,...,n$. (ρ - **korelační koeficient**)

Předpoklady (klasického) lineárního modelu



Koeficient determinace R^2

Koeficient determinace (KD) charakterizuje **přiléhavost dat k regresnímu modelu** (číslo mezi 0 a 1):

$$R^2 = \frac{S_T}{S_y} = \frac{S_y - S_R}{S_y} = 1 - \frac{S_R}{S_y}$$

$S_T = \sum_{i=1}^{n} (Y_i - \bar{y})^2$ - **teoretický součet čtverců**: $S_y = S_R + S_T$

$S_R = \sum_{i=1}^{n} (y_i - Y_i)^2$ - **reziduální součet čtverců**

Pro malé soubory: $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-2}$

Upravený KD

Příklad 1. Koeficient determinace

Závislost zisku z prodeje na velikosti nákladů na reklamu (viz Příklad 1):

$$R^2 = \frac{S_T}{S_y} = \frac{2994{,}3}{3125{,}6} = 0{,}958 \qquad R^2_{adj} = 0{,}953$$

Koeficient korelace („odmocnina KD“)

$$R = 0{,}979 \qquad R_{adj} = 0{,}979$$

Trendová funkce v časové řadě

- Hodnotami nezávisle proměnné jsou **ekvidistantní** (tj. stejně vzdálené) **časové okamžiky** t_i, $i=1,2,\ldots,n$
- Situace je častá v ekonomických aplikacích, kdy máme k dispozici tzv. **časové řady** ekonomických veličin, např. tržby v jednotlivých měsících, HDP v jednotlivých za sebou jdoucích rocích apod.
- Lineární **trendová** (regresní) **funkce**:

$$T_t = \beta_0 + \beta_1 t$$

Transformace časové osy v časové řadě

- Zavedení nové časové proměnné t' následujícím způsobem:

$t' = (t - \bar{t})$ je-li počet členů časové řady n lichý

$$\bar{t} = \frac{n+1}{2} \qquad \boxed{\sum_{t=1}^{n} t' = 0}$$

$t' = 2(t - \bar{t})$ je-li počet členů časové řady n sudý

Jednodušší odhad regresních koeficientů – MNČ:

$$b_0 = \frac{\sum y_t}{n} \qquad b_1 = \frac{\sum t' y_t}{\sum (t')^2}$$

Příklad 2. „ Časová řada “

Výrobu horských kol typu Superba / v tis. ks / udává následující tabulka:

Rok	2005	2006	2007	2008	2009	2010	2011
Výroba	22,3	22,0	22,3	???	21,3	21,4	21,1

a) Chybějící údaj za rok 2008 doplňte průměrem hodnot sousedních roků 2007 a 2009 a doplněnou časovou řadu schématicky načrtněte.

b) Z náčrtu odhadněte správný model trendu této časové řady, pak metodou regresní analýzy vypočtěte odhady neznámých regresních koeficientů.

c) Pomocí modelu z b) prognózujte velikost výroby v r. 2012 a 2013.

d) Vypočtěte koeficient determinace a na jeho základě slovně zhodnoťte„přiléhavost“ dat k regresnímu modelu.

$$b_0 = \frac{\sum y_t}{n} \quad b_1 = \frac{\sum t' y_t}{\sum (t')^2} \quad R^2 = \frac{S_T}{S_y}$$

Příklad 2. „Časová řada - výpočty“

V Excelu: Bodový graf → Přidat spojnici trendu (lineární, rovnice regrese…)

t	y_t	t'	$(t')^2$	$t' \cdot y_t$	T_t	$(y_t - T_t)^2$	$(T_t - y)^2$	$(y_t - y)^2$
1	20,5	-3	9	-61,5	20,75	0,0607	0,6117	1,0580
2	21,1	-2	4	-42,2	21,01	0,0086	0,2719	0,1837
3	21,4	-1	1	-21,4	21,27	0,0175	0,0680	0,0165
4	21,6	0	0	0	21,53	0,0051	0,0000	0,0051
5	21,8	1	1	21,8	21,79	0,0001	0,0680	0,0737
6	22,3	2	4	44,6	22,05	0,0625	0,2719	0,5951
7	22,0	3	9	66	22,31	0,0965	0,6117	0,2222
Sumy	150,7	0	28	7,3	150,70	0,2511	1,9032	2,1543

b_0 =	21,53		
b_1 =	0,26	T_{02} =	22,57
S_R =	0,251	T_{03} =	22,83
S_T =	1,903		
S_y =	2,154		

$R^2 = S_T/S_y =$ 0,88

Model vysvětluje 88 procent variability dat.
Data se těsně přimykají k regresní přímce!

Linearizované regresní funkce

- **regresní exponenciální funkce (např. Cobb-Douglasova produkční funkce):**

$$f(x) = \beta_0 \beta_1^{\,x}$$

Substituce: $y' = \ln y \qquad x' = x$

$$\beta_0' = \ln \beta_0 \quad \beta_1' = \ln \beta_1$$

MNČ vypočteme odhady: b_0', b_1'

Zpětná substituce: $b_0 = e^{b_0'} \quad b_1 = e^{b_1'}$

(odhady β_0, β_1)

Korelační analýza (KA)

- V KA **není předem známo**, které jsou vysvětlující a které vysvětlované proměnné!

Příklad: Závislost tržeb za zboží X na tržbách zboží Y

Oboustranný vztah - 2 regresní přímky:

$$y = \alpha_0 + \alpha_1 x + \varepsilon_1 \qquad x = \beta_0 + \beta_1 y + \varepsilon_2$$

Korelační koeficient: $\rho = \pm\sqrt{|\alpha_1 \beta_1|}$

Odhad ρ :

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

Příklad 3. Výsledky testů 10 studentů 1. ročníku OPF

Počet bodů z matematiky	56	79	50	84	63	91	46	56	74	76
Počet bodů z ekonomie	82	56	46	79	74	83	51	63	75	82

$$r = \frac{10 \cdot 47823 - 675 \cdot 691}{\sqrt{(10 \cdot 47687 - 675^2)(10 \cdot 49501 - 691^2)}} = 0{,}6112$$

$r > 0{,}6$ – „vysoká“ hodnota korelace!