Regresní analýza

Eva Jarošová

Obsah

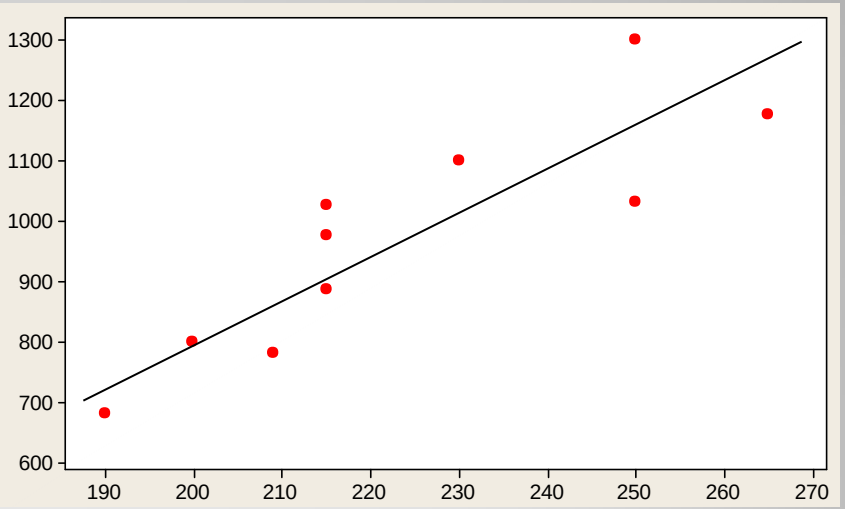
1. Regresní přímka
2. Možnosti zlepšení modelu
3. Testy v regresním modelu
4. Regresní diagnostika
5. Speciální využití

Lineární model

1. Regresní přímka

Jednoduchá regrese

Studium závislosti jedné numerické veličiny (měřitelné na spojité stupnici) na jiné numerické veličině

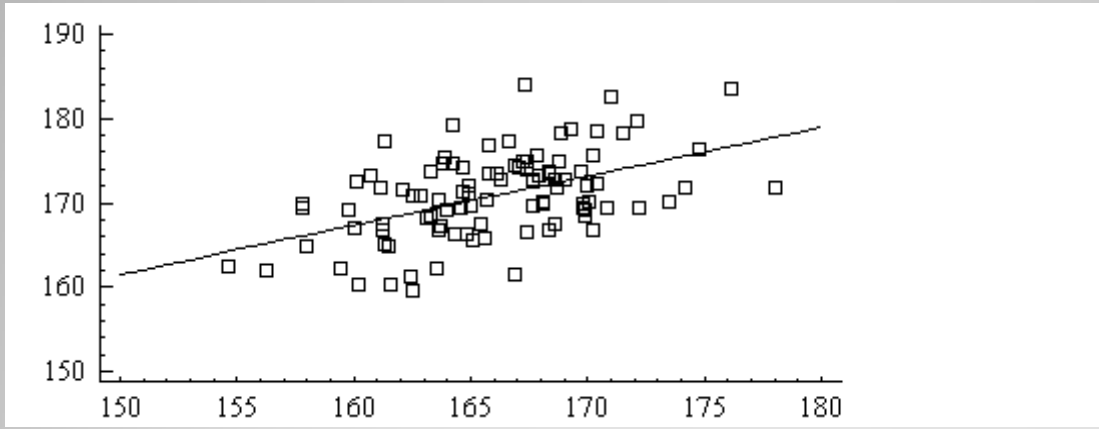


- Zkoumání existence vztahu
- Popis vztahu pomocí modelu
- Odhad (předpověď)
- Kalibrace
- Hledání optimálních podmínek

Původ pojmu regrese

Francis Galton

zkoumání závislosti výšky synů na výšce otců



y … výška syna

x … výška otce

$y_i = 74 + 0{,}58\, x_i$

$y_i - 170{,}8 = 0{,}58\,(x_i - 166)$

x = 176 y = 176,6

x = 156 y = 165

Regresní model

$$y_i = \eta_i + \varepsilon_i$$

střední hodnota proměnné Y v bodě x_i

funkce proměnné X

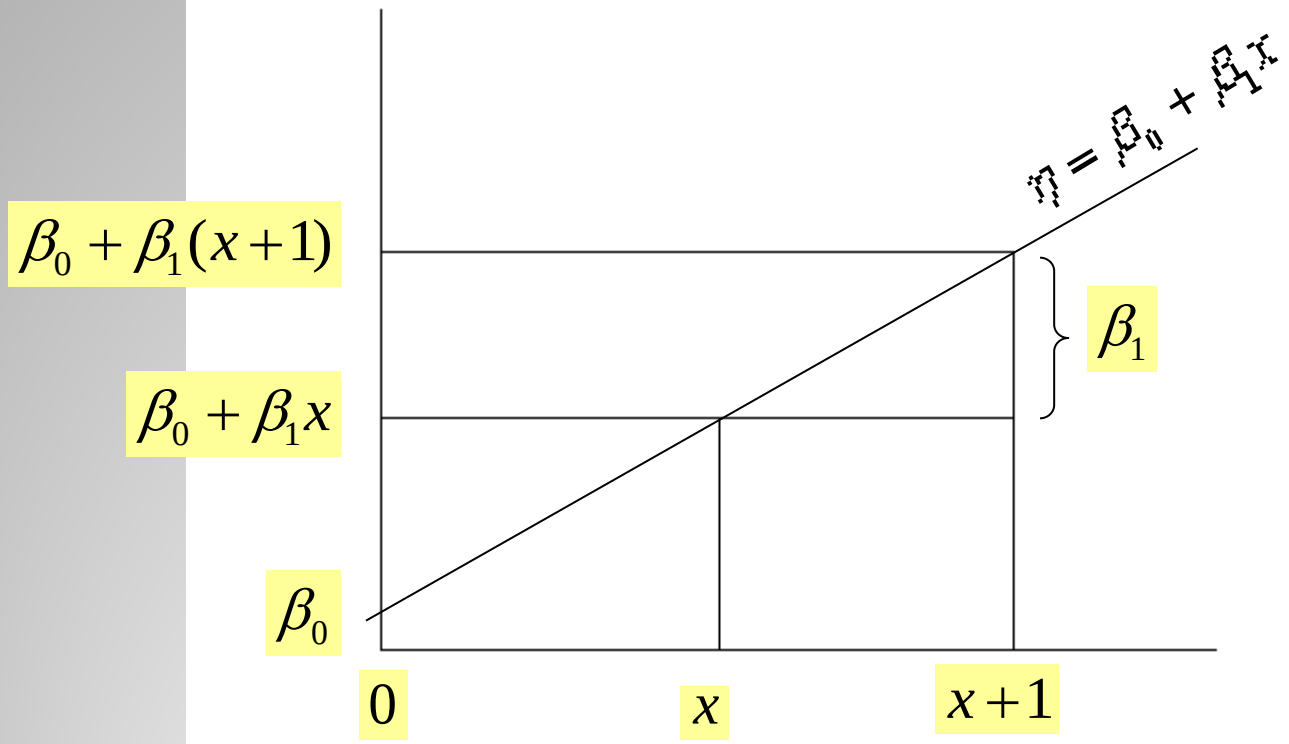
chyba (náhodná složka)

projev vlivů v modelu neuvažovaných

Model regresní přímky

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

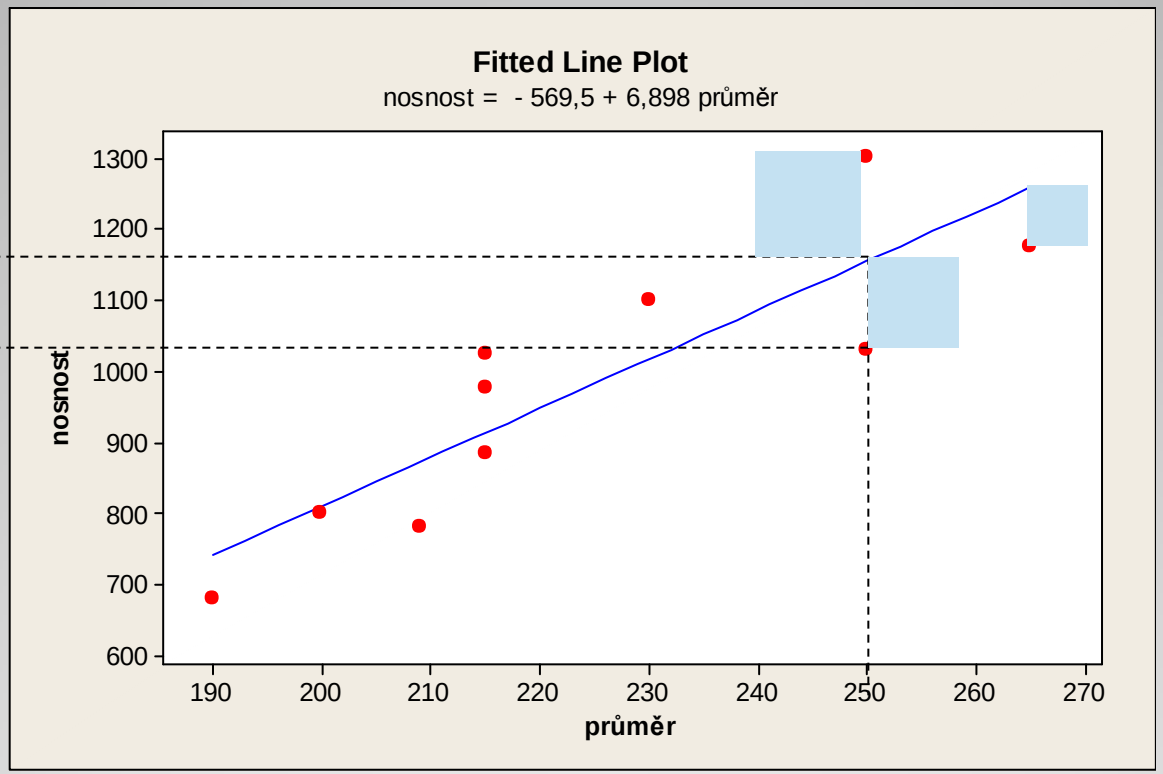
Interpretace parametrů



Odhad parametrů

vyrovnaná hodnota

pozorovaná hodnota



pozorovaná hodnota - vyrovnaná hodnota = reziduum

minimum reziduálního součtu čtverců

Metoda nejmenších čtverců

Reziduální součet čtverců (část variability závisle proměnné, která není vysvětlena modelem)

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

Hledání b_0 a b_1 tak, aby S_R byl minimální

$$\frac{\partial S_R}{\partial b_0} = 0 \qquad \frac{\partial S_R}{\partial b_1} = 0$$

Normální rovnice

$$\sum y_i = n b_0 + b_1 \sum x_i$$

$$\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2$$

Odhad parametrů

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad b_0 = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n}$$

	x_i	y_i	$x_i y_i$	x_i^2
	190	680	129200	36100
	200	800	160000	40000
	209	780	163020	43681
	215	885	190275	46225
	215	975	209625	46225
	215	1025	220375	46225
	230	1100	253000	52900
	250	1030	257500	62500
	265	1175	311375	70225
	250	1300	325000	62500
Σ	2239	9750	2219370	506581

Výstup v Minitabu

Regression Analysis: nosnost versus průměr

The regression equation is
nosnost = - 569 + 6,90 průměr

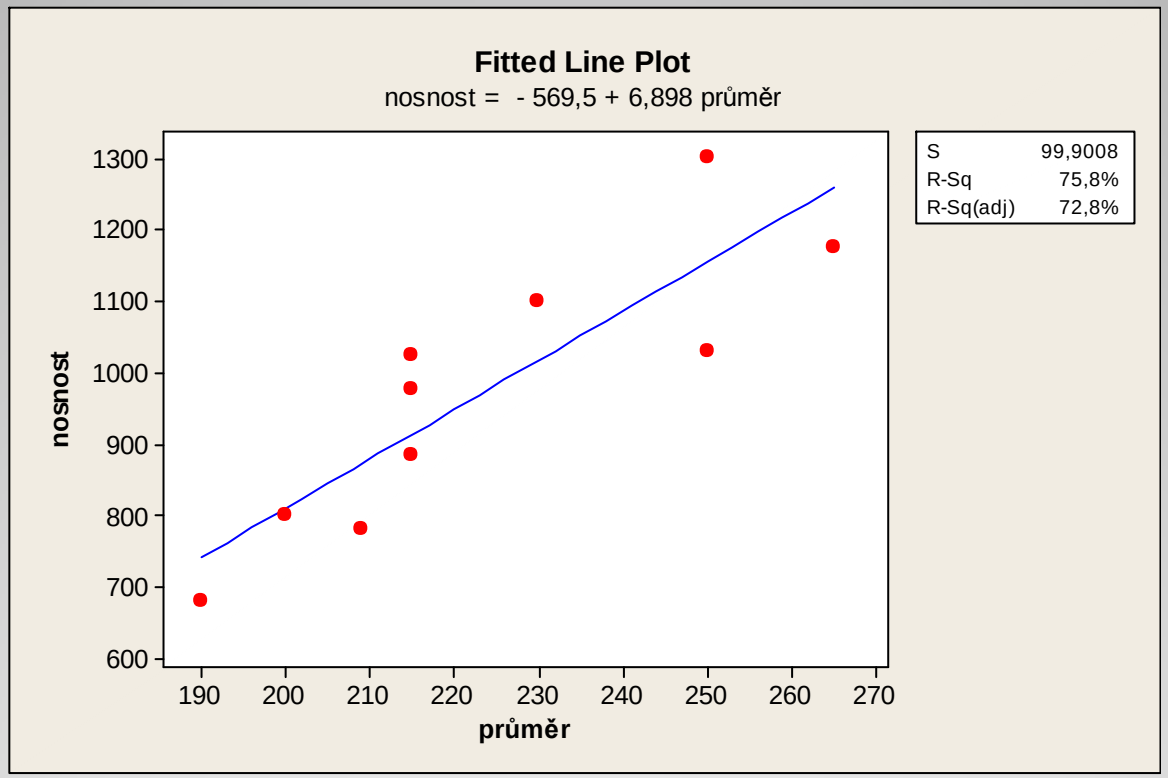
Predictor	Coef	SE Coef	T	P
Constant	-569,5	309,8	-1,84	0,103
průměr	6,898	1,376	5,01	0,001

S = 99,9008 R-Sq = 75,8% R-Sq(adj) = 72,8%

Analysis of Variance

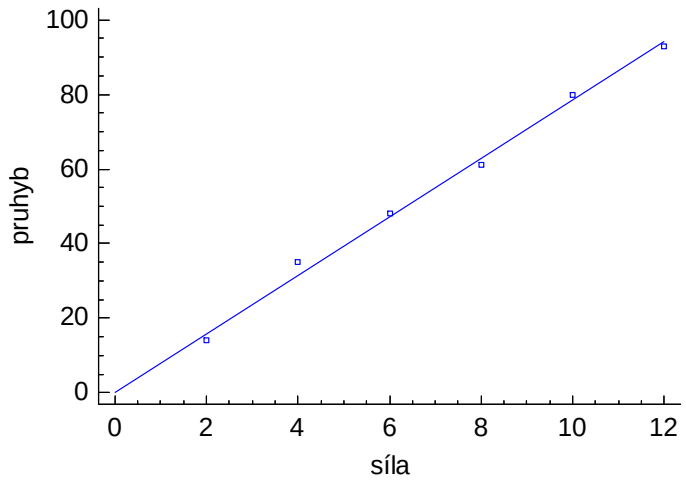
Source	DF	SS	MS	F	P
Regression	1	250709	250709	25,12	0,001
Residual Error	8	79841	9980		
Total	9	330550			

Graf regresní přímky



Nosnost svaru v závislosti na průměru, Duncan (1965)

Přímka procházející počátkem



Regression Analysis: průhyb versus síla

The regression equation is
průhyb = 7,86 síla

Predictor	Coef	SE Coef	T	P
Noconstant				
síla	7,8571	0,1138	69,04	0,000

S = 2,17124

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	22471	22471	4766,67	0,000
Residual Error	5	24	5		
Total	6	22495			

Určení modulu pružnosti plastické hmoty, Hátle, Likeš (1972)

Kvalita modelu, koeficient determinace

celková variabilita proměnné Y (celkový součet čtverců) $S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$

variabilita proměnné Y vysvětlená modelem (teoretický součet čtverců)

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

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

Platí $S_y = S_T + S_R$

koeficient determinace (model s konstantou) $R^2 = \frac{S_T}{S_y}$

hodnoty z intervalu $< 0 ; 1 >$

Výstup v Minitabu

Regression Analysis: nosnost versus průměr

The regression equation is
nosnost = - 569 + 6,90 průměr

Predictor	Coef	SE Coef	T	P
Constant	-569,5	309,8	-1,84	0,103
průměr	6,898	1,376	5,01	0,001

S = 99,9008 R-Sq = 75,8% R-Sq(adj) = 72,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	250709	250709	25,12	0,001
Residual Error	8	79841	9980		
Total	9	330550			

Kvalita modelu, směrodatná chyba odhadu

$$s = \sqrt{\frac{S_R}{n-p}}$$

Regression Analysis: nosnost versus průměr

The regression equation is
nosnost = - 569 + 6,90 průměr

Predictor	Coef	SE Coef	T	P
Constant	-569,5	309,8	-1,84	0,103
průměr	6,898	1,376	5,01	0,001

S = 99,9008 R-Sq = 75,8% R-Sq(adj) = 72,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	250709	250709	25,12	0,001
Residual Error	8	79841	9980		
Total	9	330550			

Odhad, předpověď

Bodový odhad pro průměr 220

$$Y = -569{,}5 + 6{,}898 \cdot 220 = 948{,}1$$

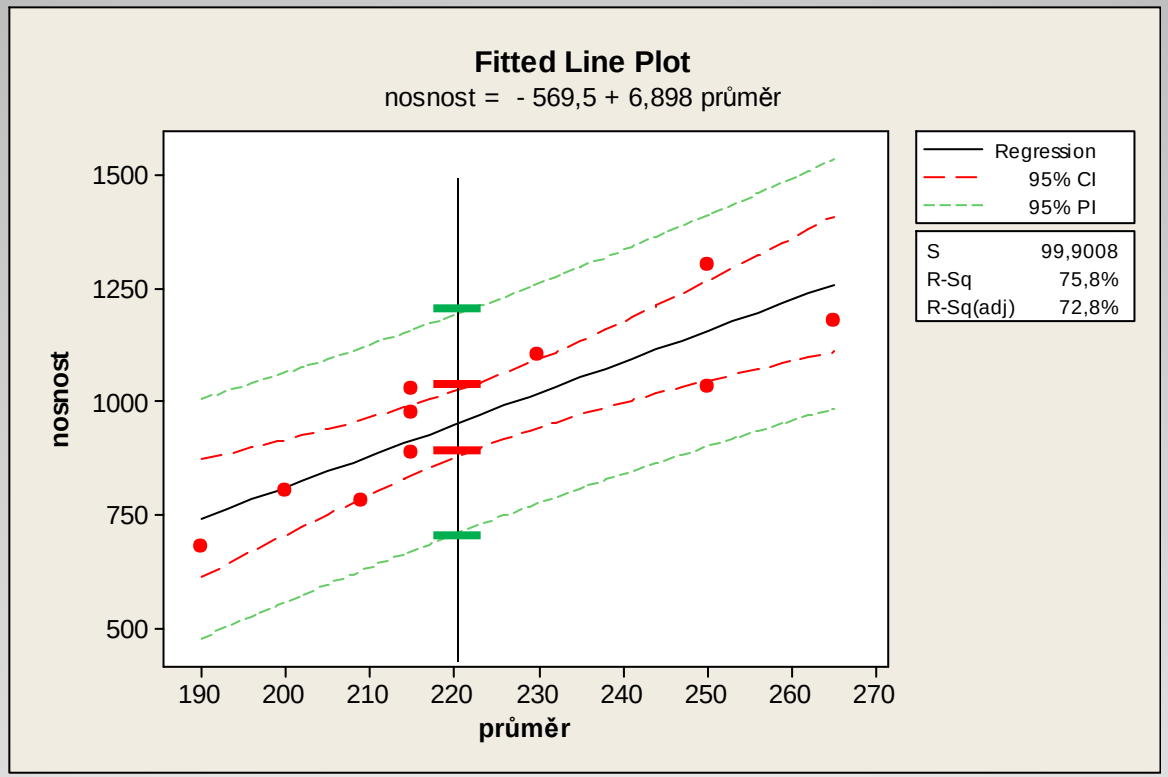
odhad střední hodnoty nosnosti při průměru 220

odhad nosnosti při průměru 220

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	948,1	32,0	(874,2; 1022,0)	(706,2; 1190,0)

Intervalový odhad



2. Možnosti zlepšení modelu

Příčiny nízké hodnoty koeficientu determinace

- Nevhodně zvolená vysvětlující proměnná

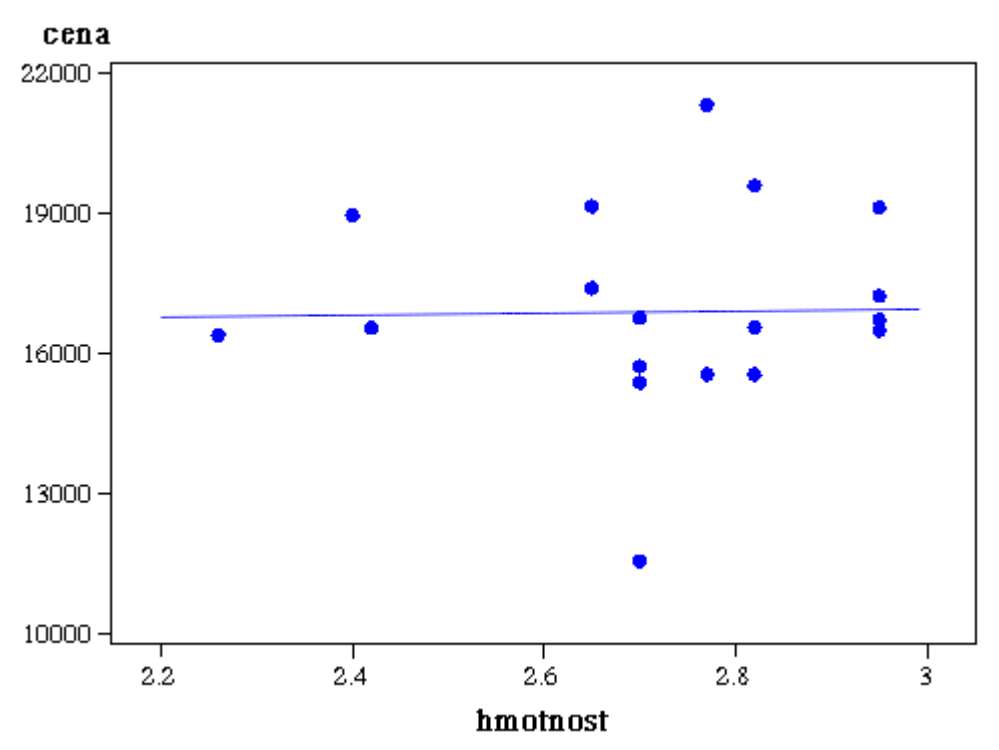
  (závislost neexistuje nebo je slabá)
- Nevhodně zvolený tvar modelu

  (závislost existuje, ale není lineární)
- Nejsou zařazeny všechny důležité vysvětlující proměnné

  (je třeba hledat další proměnné, které mají vliv na Y)

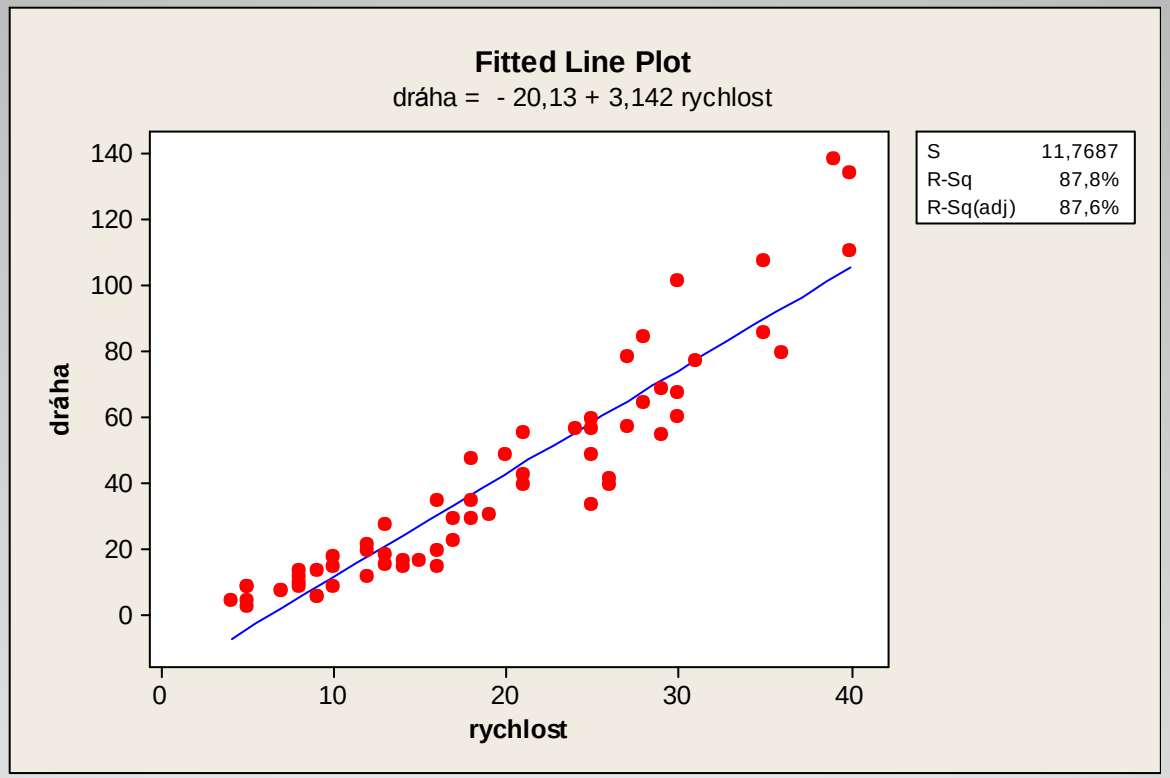
Neexistující závislost

Cena notebooku vs hmotnost



$Y = 16\,315 + 210\,x$ $R^2 = 0{,}0004$

Nevhodný tvar modelu



Další modely lineární z hlediska parametrů

$$\eta = \beta_0 + \beta_1 \ln x$$

$$\eta = \beta_0 + \beta_1 \frac{1}{x}$$

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Modely po transformaci, např. $\eta = \beta_0 \beta_1^x \longrightarrow$

$$\ln \eta = \ln \beta_0 + x \ln \beta_1$$

Regresní parabola

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2$$

jedna vysvětlující proměnná

v lineárním a kvadratickém tvaru

speciální případ regresního polynomu

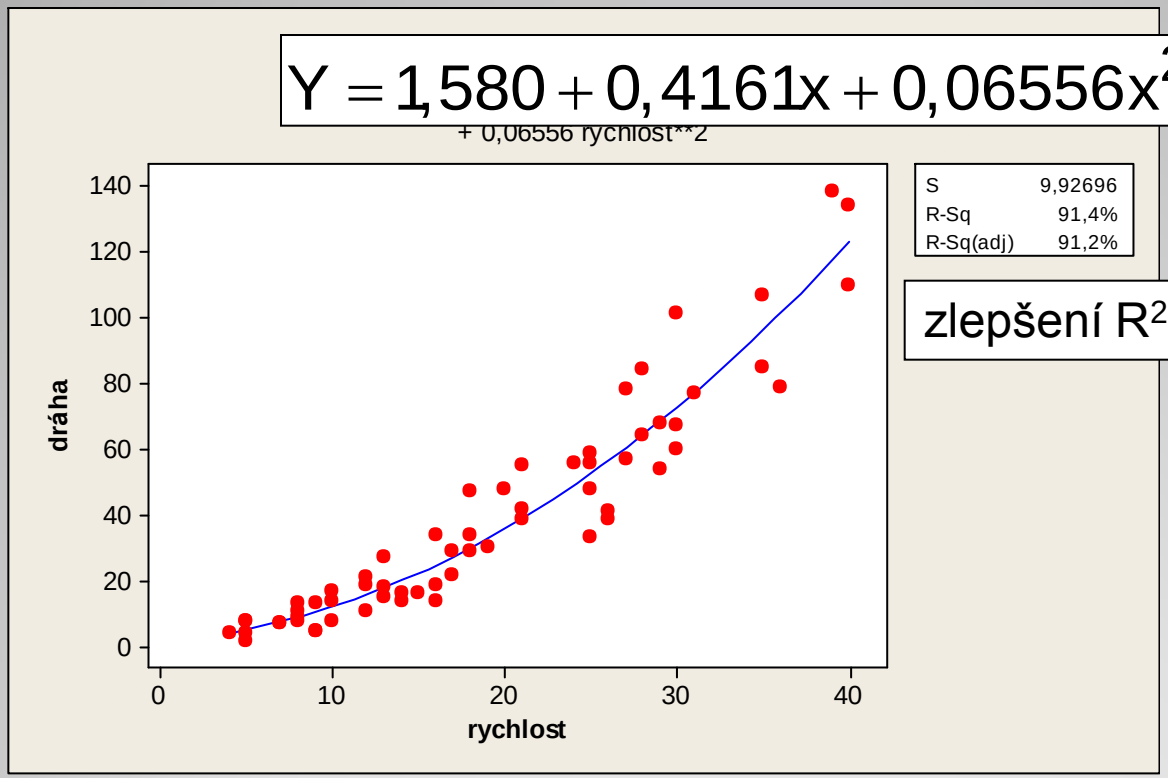
ačkoli regresní funkce obsahuje kvadratický člen,

je lineární v parametrech

odhad parametrů metodou nejmenších čtverců

Regresní parabola

$$Y = 1{,}580 + 0{,}4161x + 0{,}06556x^2$$



zlepšení R^2 z 87,8 % na 91,4 %

$$Y = b_0 + b_1x + b_2x^2$$

Nejsou zařazeny důležité vysvětlující proměnné

Regression Analysis: spotřeba versus stáří

The regression equation is
spotřeba = 6,96 + 0,243 stáří

Predictor	Coef	SE Coef	T	P
Constant	6,9595	0,5565	12,51	0,000
stáří	0,2429	0,1015	2,39	0,024

S = 1,51898 R-Sq = 17,5% R-Sq(adj) = 14,4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	13,202	13,202	5,72	0,024
Residual Error	27	62,297	2,307		
Total	28	75,499			

Regression Analysis: spotřeba versus stáří; obsah

The regression equation is
spotřeba = 0,756 + 0,200 stáří + 0,00397 obsah

Predictor	Coef	SE Coef	T	P
Constant	0,7559	0,5736	1,32	0,199
stáří	0,20013	0,04132	4,84	0,000
obsah	0,0039693	0,0003375	11,76	0,000

S = 0,615676 R-Sq = 86,9% R-Sq(adj) = 85,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	65,643	32,822	86,59	0,000
Residual Error	26	9,855	0,379		
Total	28	75,499			

$$Y = 6,9595 + 0,2429x_1$$

$$R^2 = 17,5\,\%$$

$$Y = 0,7559 + 0,20013\ x_1 + 0,0039693\ x_2$$

$$R^2 = 86,9\,\%$$

Porovnání modelů s různým počtem parametrů

Upravený koeficient determinace

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p}$$

x_1: R^2 adj = 14,4 %

x_1, x_2 R^2 adj = 85,9 %

p… počet parametrů modelu

3. Testy v regresním modelu

Ověření existence závislosti

model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

t-test

Testovaná hypotéza H_0: $\beta_1 = 0$

(Y nezávisí na X)

Alternativní hypotéza H_1: $\beta_1 \neq 0$

(Y závisí na X)

Postup při t-testu

Testová statistika $$t = \frac{b_j}{s(b_j)}$$

směrodatná chyba odhadu,
vyjadřuje přesnost odhadu

Kritický obor $$W_\alpha = \{ t : |t| > t_{1-\alpha/2} \}$$

Platí-li $|t| > t_{1-\alpha/2}$, zamítneme H_0.

$t_{1-\alpha/2}$ kvantil t-rozdělení s ($n - p$) stupni volnosti,
n je rozsah výběrového souboru
p je počet parametrů modelu (u přímky p = 2)

Využití p-hodnoty

P-hodnota … pravděpodobnost, že při platnosti testované hypotézy H_0 nabude testová statistika hodnoty svědčící ještě více v neprospěch H_0 než vypočtená hodnota testové statistiky

Je-li p-hodnota menší než α, H_0 zamítneme na hladině významnosti α.

Výhoda používání p-hodnoty: Ihned vidíme, jak silný „důkaz“ proti platnosti H_0 máme.

$0,01 < p < 0,05$ slabší důkaz

$0,001 < p < 0,01$ silnější důkaz

$p < 0,001$ silný důkaz

Výstup v Minitabu

Regression Analysis: nosnost versus průměr

The regression equation is
nosnost = - 569 + 6,90 průměr

Predictor	Coef	SE Coef	T	P
Constant	-569,5	309,8	-1,84	0,103
průměr	6,898	1,376	5,01	0,001

S = 99,9008 R-Sq = 75,8% R-Sq(adj) = 72,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	250709	250709	25,12	0,001
Residual Error	8	79841	9980		
Total	9	330550			

Ověření existence závislosti

model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

F-test

Testovaná hypotéza H_0: $\beta_1 = \beta_2 = 0$

(Y nezávisí na žádné z vysvětlujících proměnných)

Alternativní hypotéza H_1: non H_0

(Y závisí alespoň na jedné z vysvětlujících proměnných)

Postup při F-testu

Testová statistika $$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}}$$

p – počet parametrů (zde *p* = 3)
n – rozsah výběrového souboru)

Kritický obor $$W_\alpha = \{F: F > F_{1-\alpha}\}$$

Platí-li $F > F_{1-\alpha}$, zamítneme H_0.

$F_{1-\alpha}$ - kvantil F rozdělení s (p – 1) a (*n* – *p*) stupni volnosti

Výstup v Minitabu

Regression Analysis: spotřeba versus stáří; obsah

The regression equation is
spotřeba = 0,756 + 0,200 stáří + 0,00397 obsah

Predictor	Coef	SE Coef	T	P
Constant	0,7559	0,5736	1,32	0,199
stáří	0,20013	0,04132	4,84	0,000
obsah	0,0039693	0,0003375	11,76	0,000

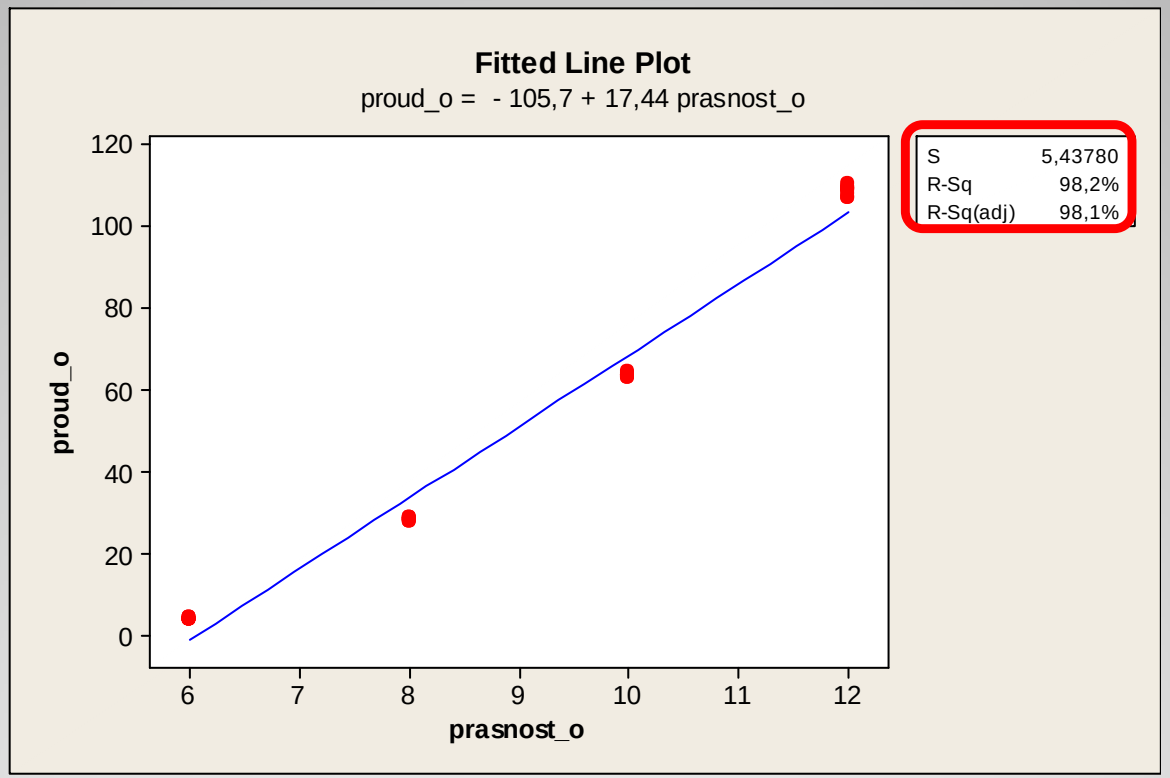
S = 0,615676 R-Sq = 86,9% R-Sq(adj) = 85,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	65,643	32,822	86,59	0,000
Residual Error	26	9,855	0,379		
Total	28	75,499			

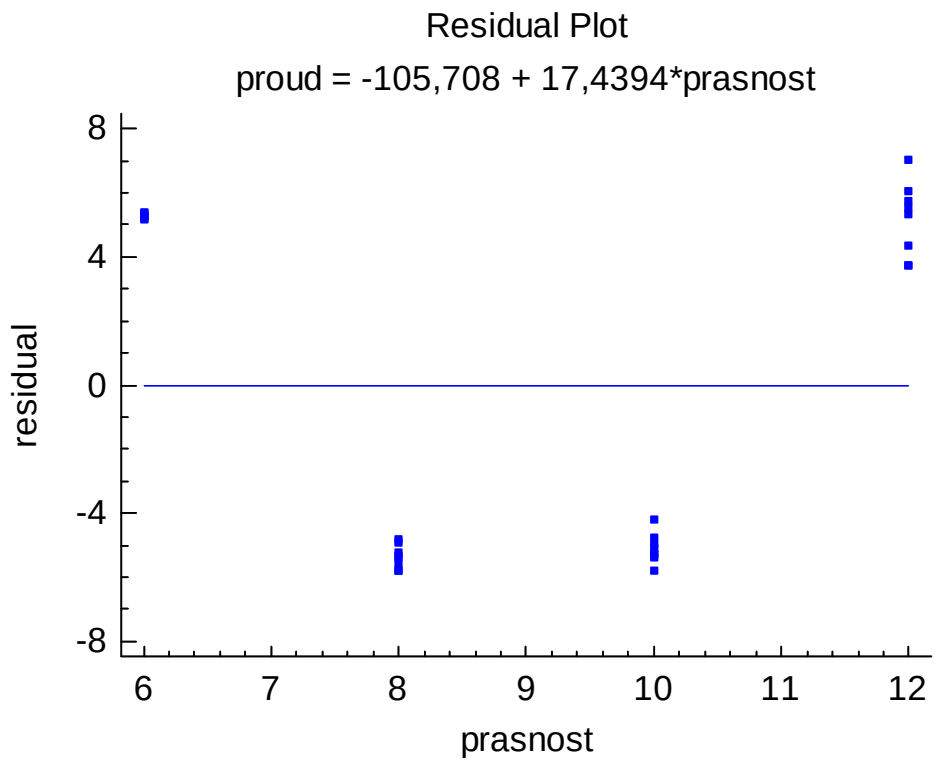
4. Regresní diagnostika

Zkoumání vhodnosti tvaru regresní funkce

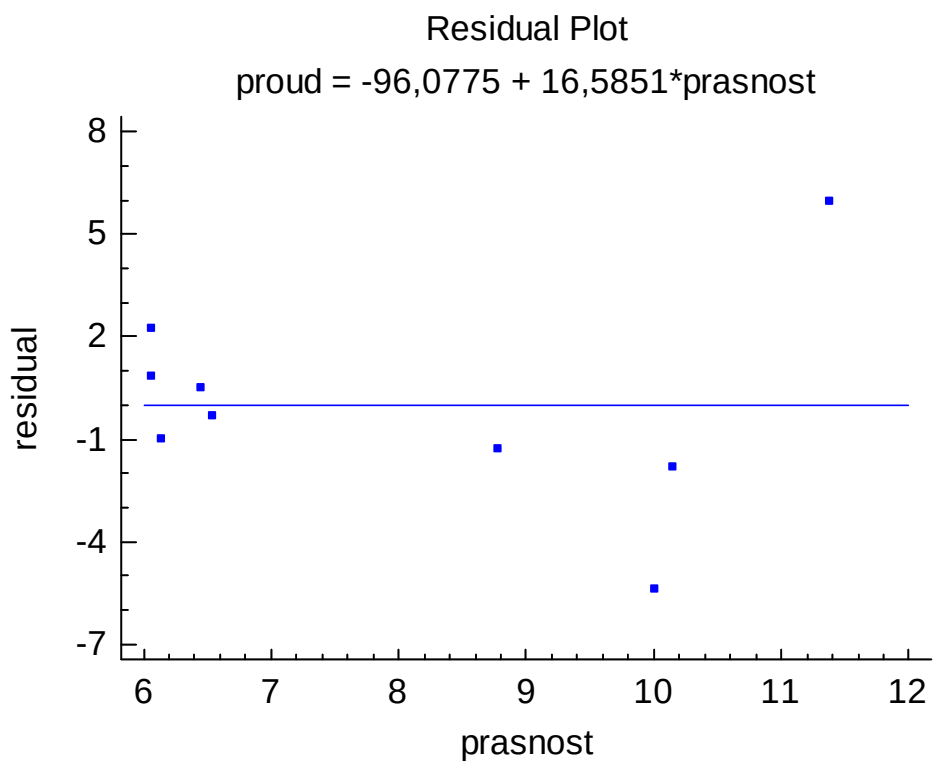


Odlučovač popílku, opakovaná data

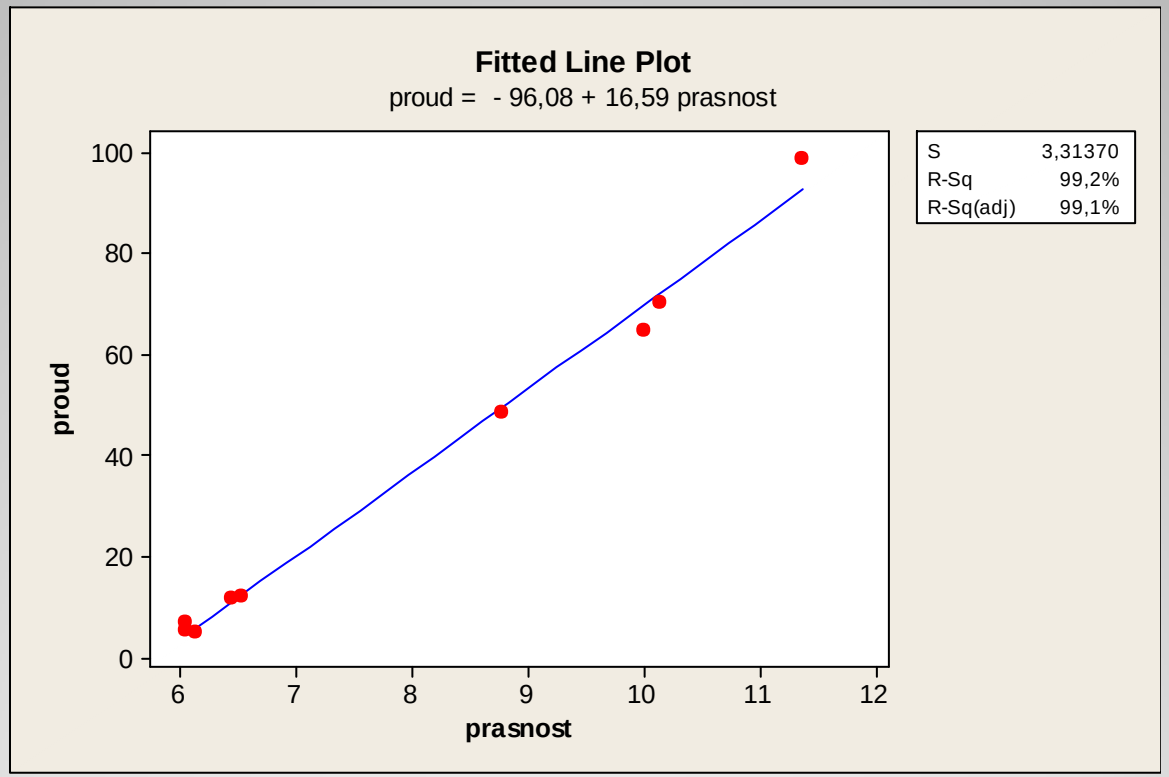
Graf rezidua vs vysvětlující proměnná



Rezidua vs vysvětlující proměnná

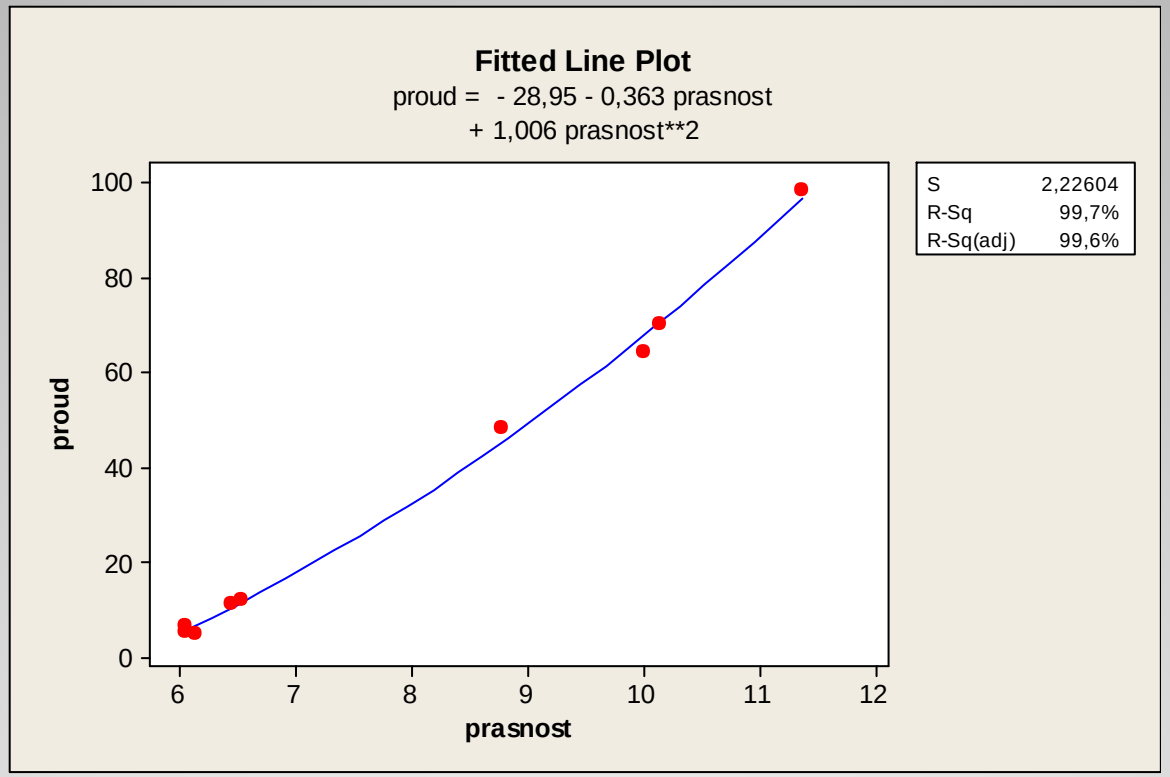


Zkoumání vhodnosti tvaru regresní funkce

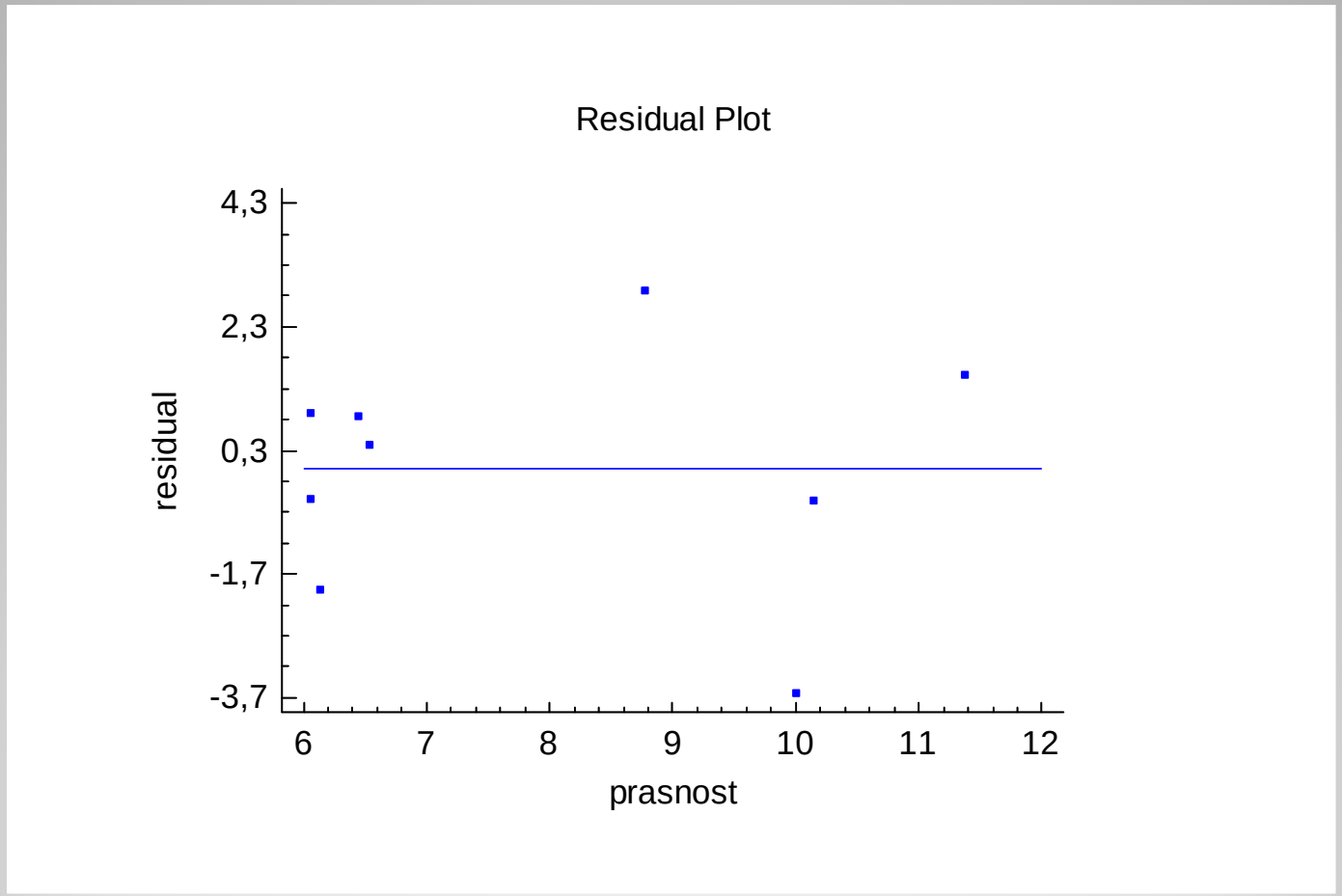


Kalibrace filtru v odlučovači popílku, Vaněk

Volba jiné regresní funkce



Rezidua vs vysvětlující proměnná



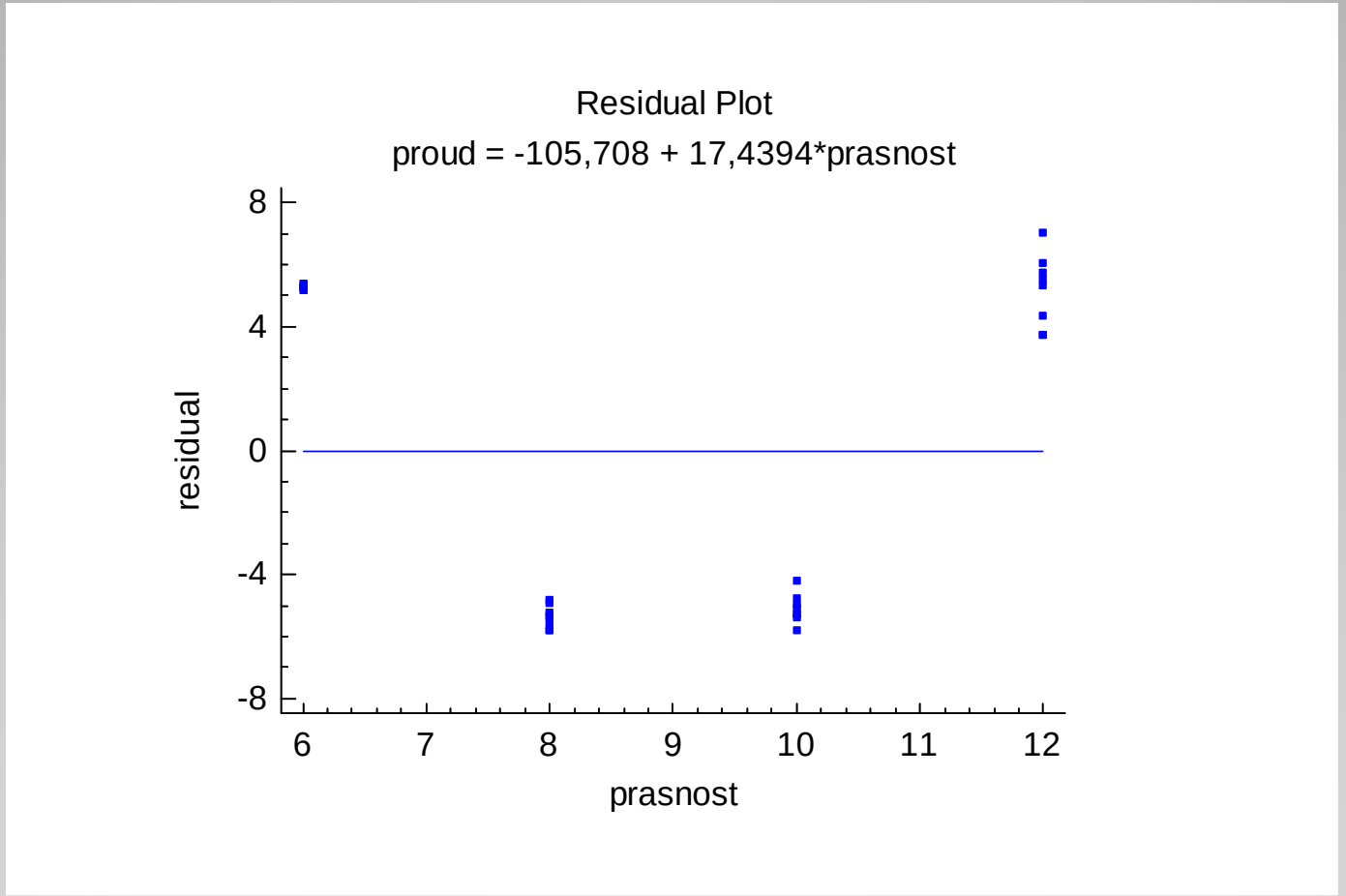
Ověření předpokladů o náhodné složce

- Konstantní rozptyl (homoskedasticita)
- Nezávislost
- Normalita

Důsledky zanedbání předpokladů vliv na

- přesnost odhadů
- na spolehlivost intervalového odhadu
- závěry t-testů a F-testu

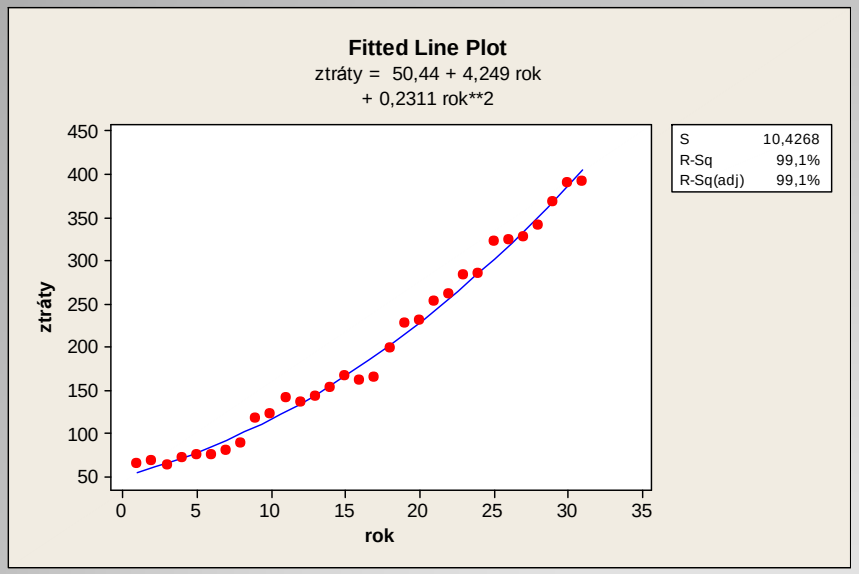
Heteroskedasticita



Heteroskedasticita

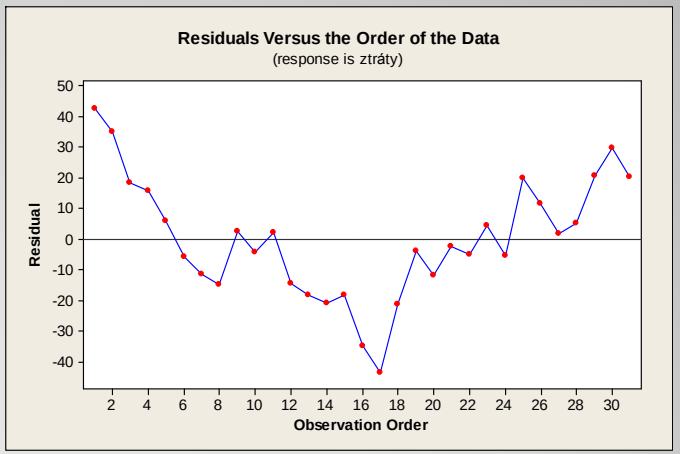
- Zjišťování heteroskedasticity
  - Grafy reziduí
  - Testy (Bartlett, Levene, Glejser, Goldfeld-Quandt, Breusch-Pagan, …)
- Opatření
  - Transformace, např. logaritmická
  - Vážená metoda nejmenších čtverců

Autokorelace



Ztráty při výrobě vody v letech 1953 – 1983, Zvára (1989)

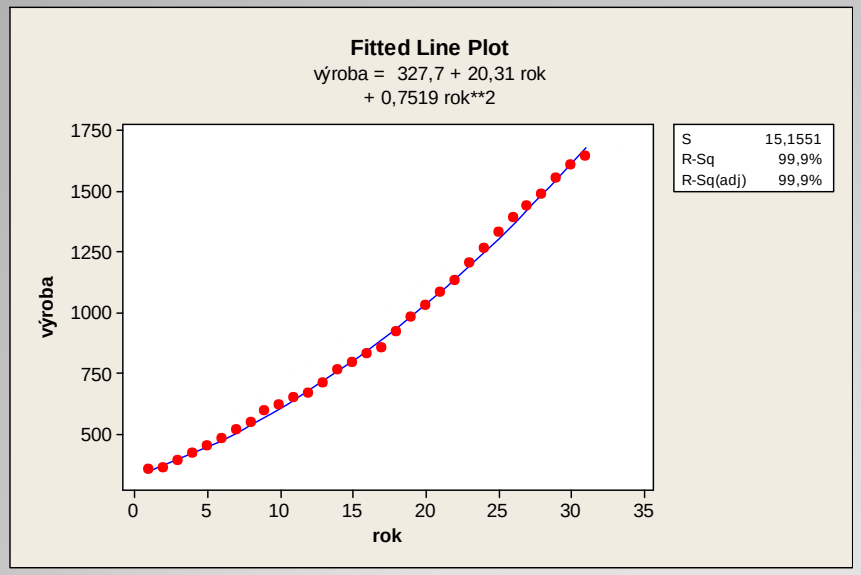
Graf reziduí

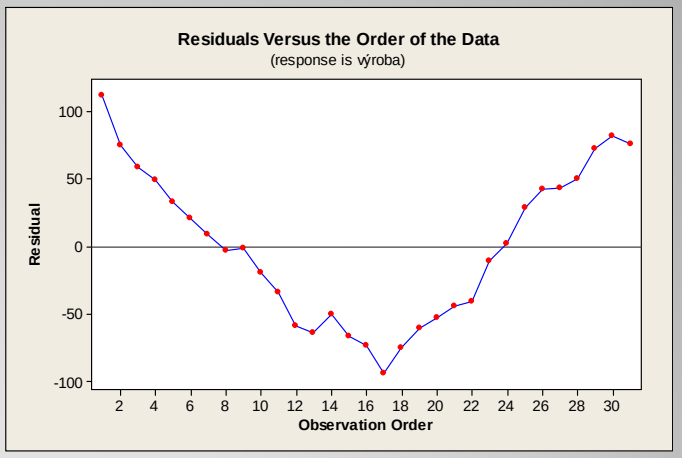


Durbin-Watsonův test

```
Durbin-Watson statistic = 0,350593
```

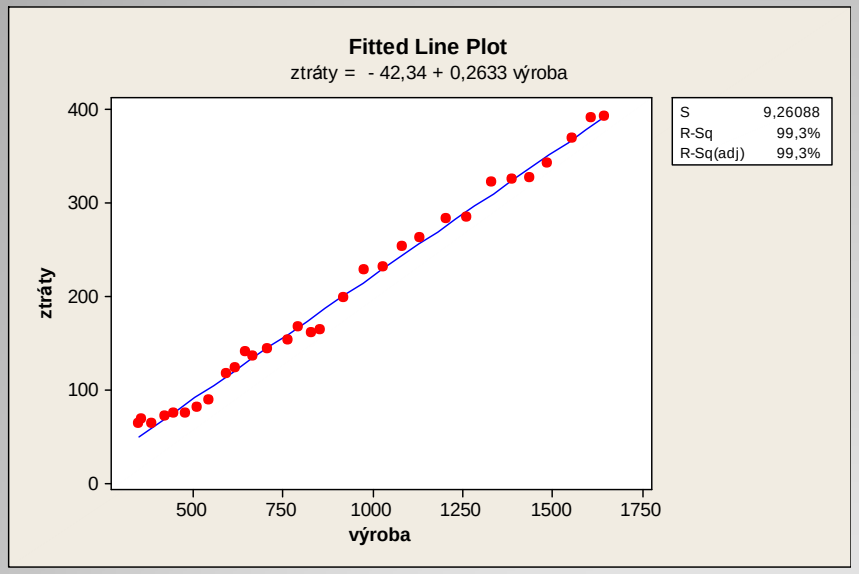
Autokorelace

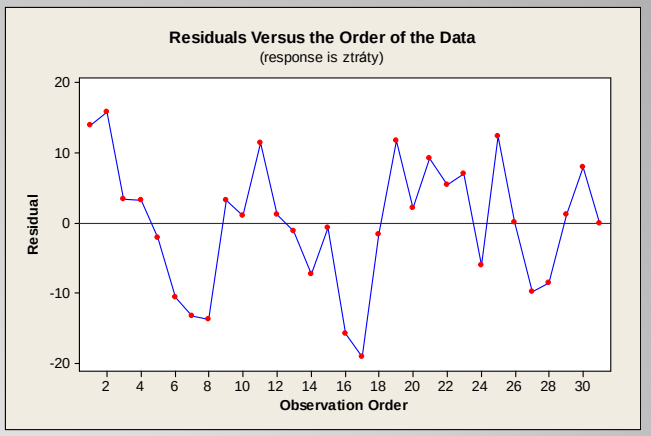




Durbin-Watson statistic = 0,0808722

Autokorelace





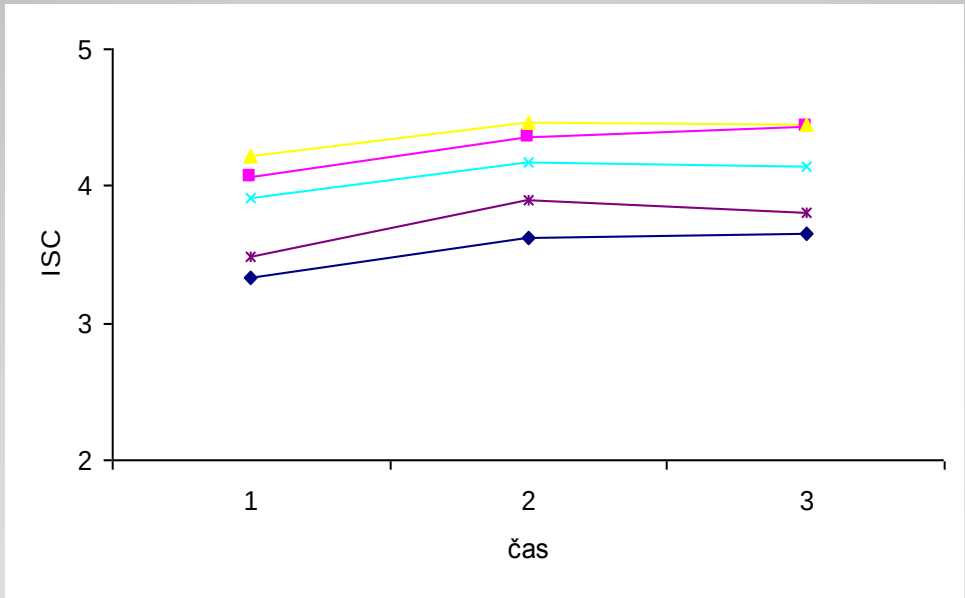
Durbin-Watson statistic = 1,08188

Závislost ztrát na výrobě vody, Zvára (1989)

Opatření: transformace (Cochran – Orcutt)

Longitudinální data

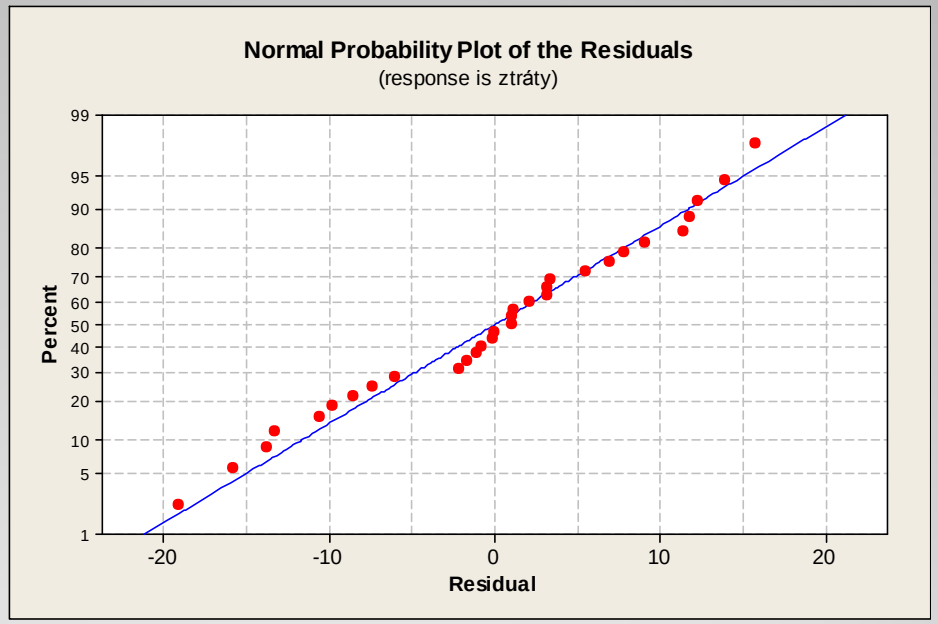
Speciální případ závislosti (kromě autokorelace ještě závislost pozorování náležejících stejné jednotce)



Degradace solárních článků, Kenett, Zacks (1998)

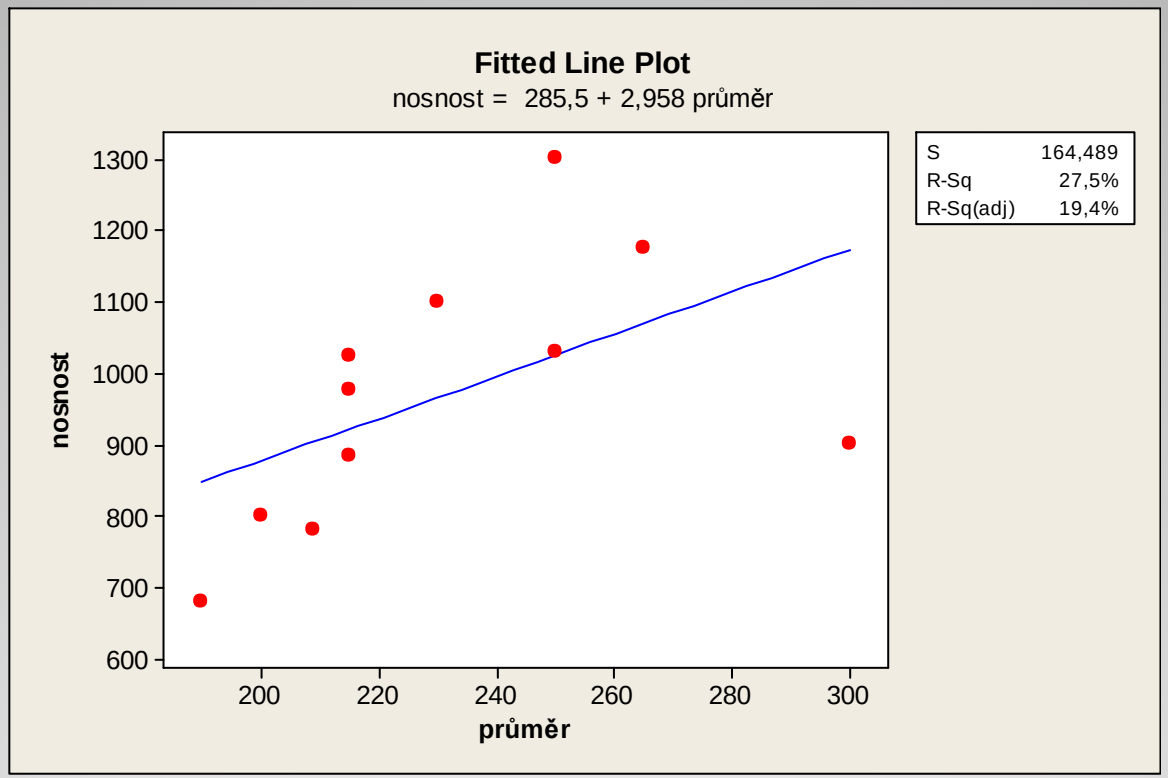
Lineární model s náhodnými efekty

Nesplněný předpoklad normálního rozdělení



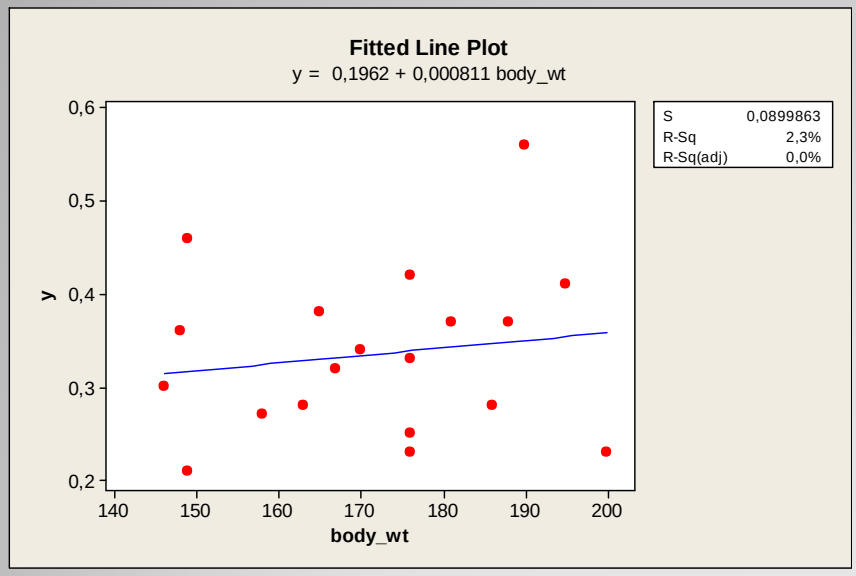
Testy normality: např. Shapiro-Wilk
Opatření:
Transformace
Zobecněný lineární model (gama model, logistická regrese)

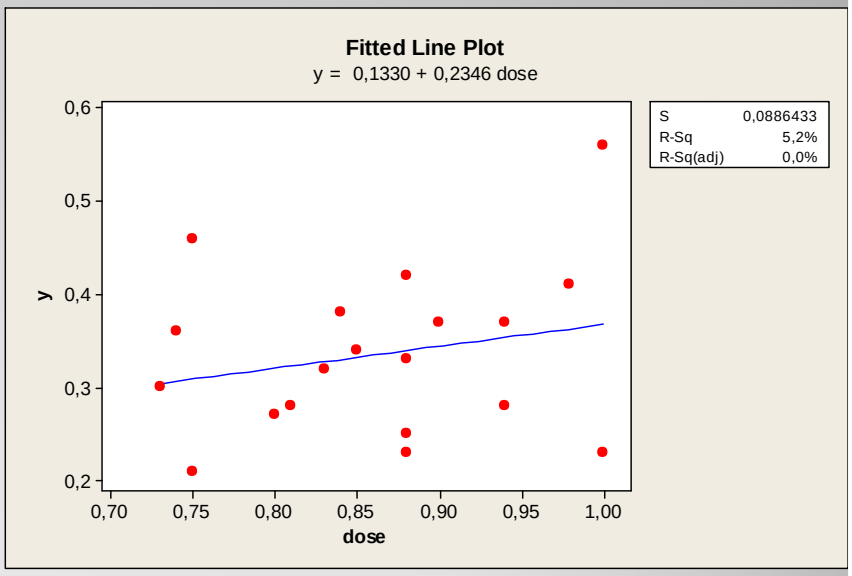
Odlehlá a vlivná pozorování



Nosnost svaru, přidáno (300;900)

Důsledky přítomnosti vlivného pozorování





Regression Analysis: y versus dose

The regression equation is
y = 0,133 + 0,235 dose

Predictor	Coef	SE Coef	T	P
Constant	0,1330	0,2109	0,63	0,537
dose	0,2346	0,2435	0,96	0,349

S = 0,0886433 R-Sq = 5,2% R-Sq(adj) = 0,0%

Unusual Observations

Obs	dose	y	Fit	SE Fit	Residual	St Resid
3	1,00	0,5600	0,3676	0,0393	0,1924	2,42R

R denotes an observation with a large standardized residual.

Regression Analysis: y versus body_wt

The regression equation is
y = 0,196 + 0,00081 body_wt

Predictor	Coef	SE Coef	T	P
Constant	0,1962	0,2216	0,89	0,388
body_wt	0,000811	0,001286	0,63	0,537

S = 0,0899863 R-Sq = 2,3% R-Sq(adj) = 0,0%

Unusual Observations

Obs	body_wt	y	Fit	SE Fit	Residual	St Resid
3	190	0,5600	0,3502	0,0315	0,2098	2,49R

R denotes an observation with a large standardized residual.

Regression Analysis: y versus body_wt; dose

The regression equation is
y = 0,286 - 0,0204 body_wt + 4,13 dose

Predictor	Coef	SE Coef	T	P
Constant	0,2855	0,1913	1,49	0,155
body_wt	-0,020444	0,007838	-2,61	0,019
dose	4,125	1,506	2,74	0,015

S = 0,0765380 R-Sq = 33,5% R-Sq(adj) = 25,1%

Unusual Observations

Obs	body_wt	y	Fit	SE Fit	Residual	St Resid
3	190	0,5600	0,5264	0,0697	0,0336	1,06 X

X denotes an observation whose X value gives it large influence.

Regression Analysis: y versus body_wt; dose
bez 3. pozorování

The regression equation is
y = 0,332 - 0,0044 body_wt + 0,88 dose

Predictor	Coef	SE Coef	T	P
Constant	0,3320	0,1954	1,70	0,110
body_wt	-0,00444	0,01693	-0,26	0,797
dose	0,875	3,400	0,26	0,800

S = 0,0762182 R-Sq = 0,5% R-Sq(adj) = 0,0%

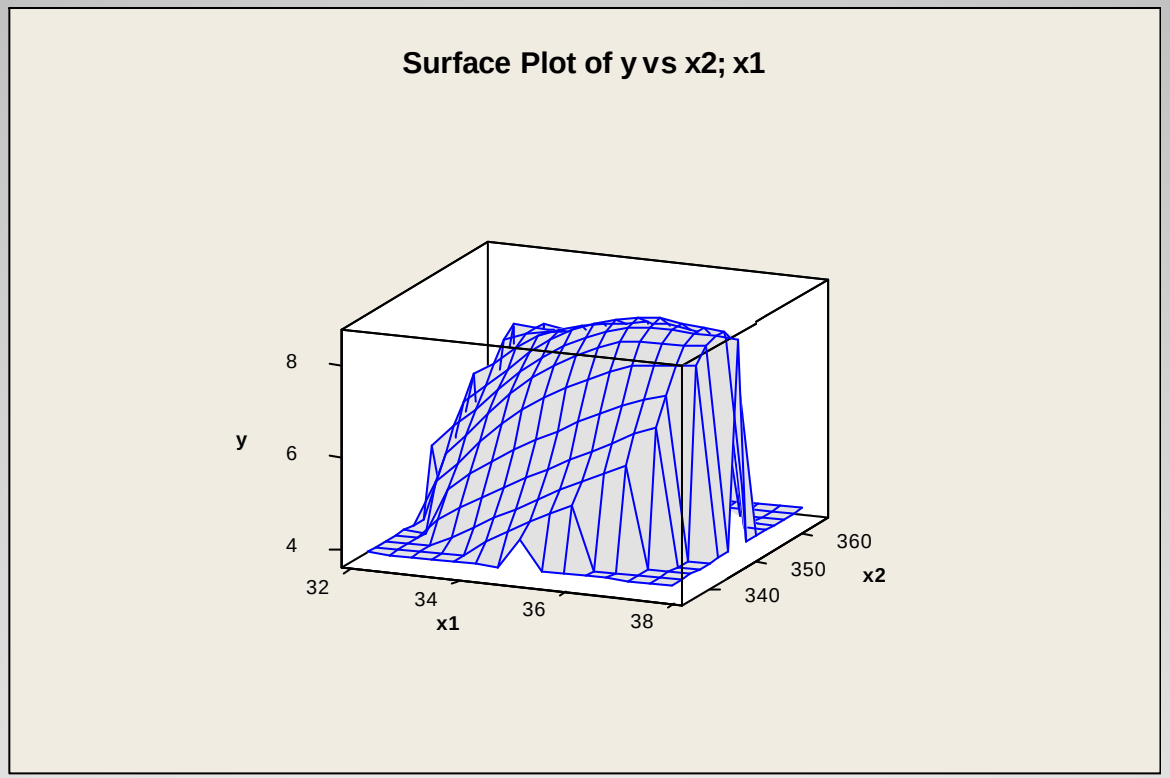
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0,000423	0,000211	0,04	0,964
Residual Error	15	0,087138	0,005809		
Total	17	0,087561			

5. Speciální využití

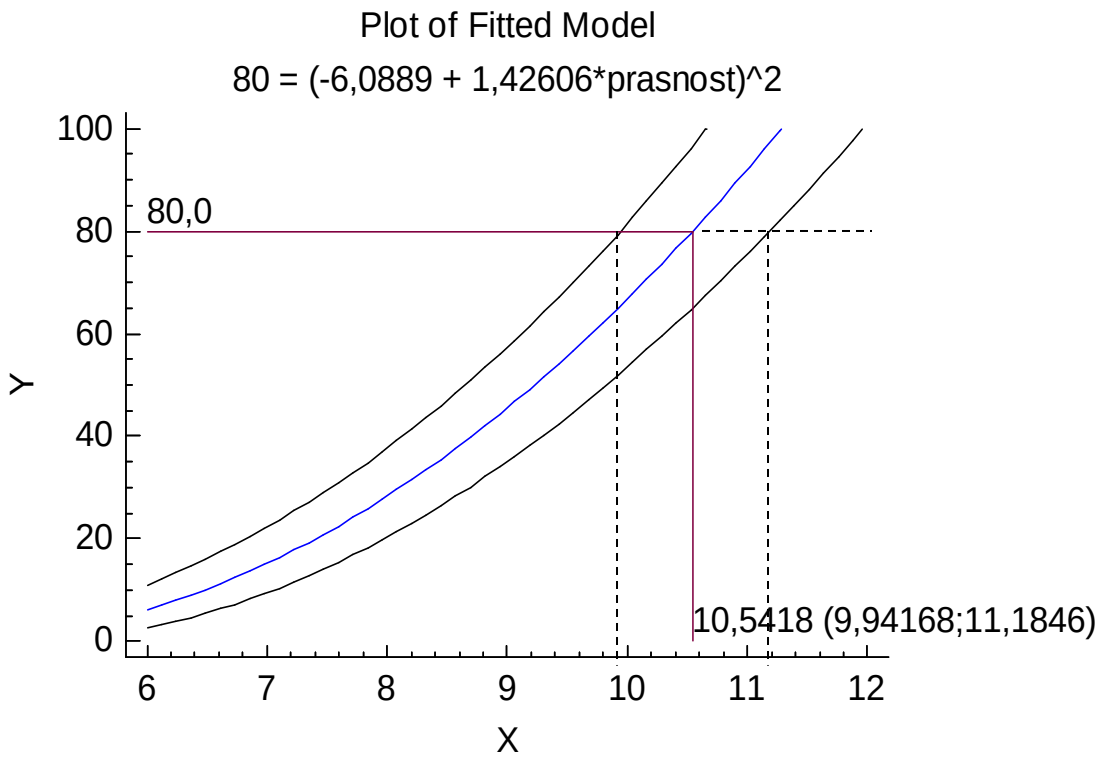
DOE – Odezvové plochy
hledání optimálních podmínek

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$



Chutnost koláče v závislosti na době a teplotě pečení, Weisberg (2005)

Kalibrace


Plot of Fitted Model
80 = (-6,0889 + 1,42606*prasnost)^2
100
80
60
40
20
0
Y
80,0
10,5418 (9,94168;11,1846)
6
7
8
9
10
11
12
X