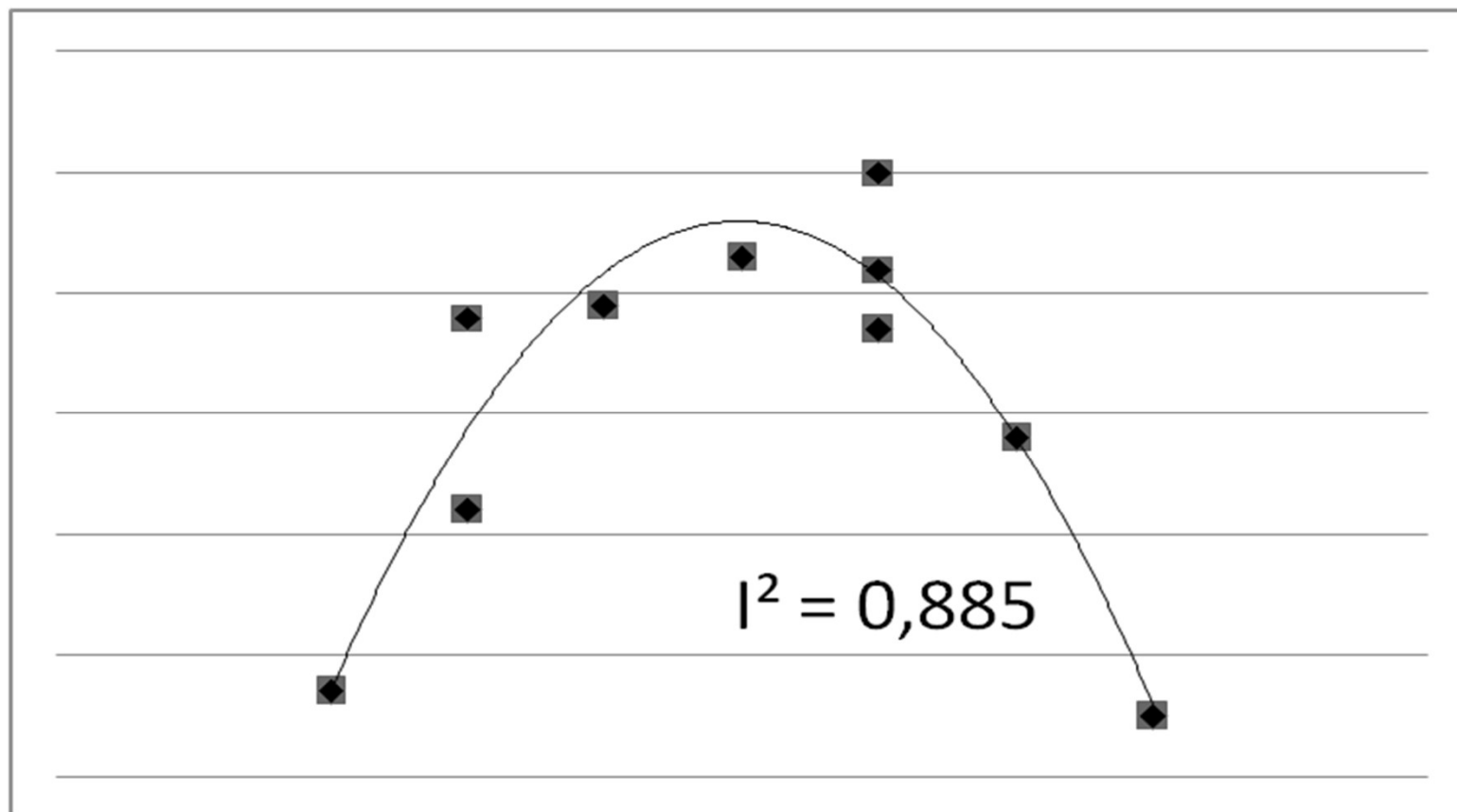


Korelace spec. pro model jednoduché nelineární regrese

Př: Stejná data, ale jiný, kvadratický model
(*kde už tedy nepočítáme r , jen l^2 !*)



Jednoduchá regrese – různé modely

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

Vektor $\mathbf{b}$ odhadů (b_0, b_1, b_2) pro parametry $\beta_0, \beta_1, \beta_2$ je opět dán vzorcem $(\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{y}$,
přičemž matice $\mathbf{F}$ má zde tvar:

$$\mathbf{F} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Jednoduchá regrese – různé modely

Pro jedna data lze tedy najít jak model lineární (L), tak kvadratický (K).

Označme:

- * odhady parametrů v L: $b_0^{(L)}$, $b_1^{(L)}$

- * odhady parametrů v K: $b_0^{(K)}$, $b_1^{(K)}$, $b_2^{(K)}$

(pozor, obecně např. $b_0^{(L)} \neq b_0^{(K)}$)

Jednoduchá regrese – různé modely

Dále označme:

- * součet rez.čtverců pro L: $Q_e^{(L)}$
- * součet rez.čtverců pro K: $Q_e^{(K)}$
- * index determinace pro L: $I^2(L)$
- * index determinace pro K: $I^2(K)$

Vždy (u modelů pro tatáž data): $Q_e^{(L)} \geq Q_e^{(K)}$

Jednoduchá regrese – různé modely

Hodnota Q_Y je dána pouze y-ovými hodnotami, nezávisí na modelu (*je tedy stejná u každého modelu pro tatáž data*)

=>

Vždy (u modelů pro tatáž data): $I^2(L) \leq I^2(K)$

(Logické – parabola se dle potřeby může „prohnout“ a o trochu lépe vysvětlit data.)

Jednoduchá regrese – různé modely

? Lze tedy říct, že parabola je vždy LEPŠÍ model než přímka ?

NE: Parabola je vždy VÝSTIŽNĚJŠÍ,
ale výhodou přímky je její

JEDNODUCHOST

Každý model = kompromis mezi
výstižností a jednoduchostí

Jednoduchá regrese – různé modely

Reziduální rozptyl s_e^2

je míra kvality modelu, zohledňující jak jeho výstižnost (Q_e), tak složitost (p značí počet parametrů):

$$s_e^2 = Q_e / (n - p)$$

Jednoduchá regrese – různé modely

Reziduální rozptyl s_e^2

hodnotu nelze interpretovat,

slouží pouze k porovnání různých modelů
pro tatáž data,

a to dle hesla „čím menší (je s_e^2), tím lepší
(je pro daná data příslušný model)“

Jednoduchá regrese – různé modely

Adjustovaný index determinace R_{adj}^2

$$R_{adj}^2 = 1 - (1 - I^2) \cdot (n - 1) / (n - p)$$

= % kvalita modelu při zohlednění počtu parametrů (p), i ten slouží k porovnání různých modelů pro tatáž data,

a to dle hesla „čím větší (je R_{adj}^2), tím lepší (je pro daná data příslušný model)“

Jednoduchá regrese – různé modely

Př: Data - Westwood Company (pokr.)

Pro data: $Q_Y=13660$

Pro lin.model:	$Q_e^{(L)}=60,$
<i>(už víme)</i>	$ ^2 (L)=0,995608$

Pro kvadr.model:	$Q_e^{(K)}=59,958907$
<i>(neodvozovat)</i>	$ ^2 (K)=0,995611$

Jednoduchá regrese – různé modely

Př: Data - Westwood Company (pokr.)

Kdybychom daty proložili místo přímky parabolou, znamenalo by to sice vylepšení výstižnosti modelu, ale zanedbatelné (z 99,5608 % na 99,5611 %);

na grafu by nebyl průběh paraboly mezi daty k rozeznání od přímky

Jednoduchá regrese – různé modely

Př: Data - Westwood Company (pokr.)

Pro lin.model:

$$s_e^2 (L) = 60/(10-2) = \underline{\underline{7,5}}$$

Pro kvadr.model:

$$s_e^2 (K) = 59,958907/(10-3) = \underline{\underline{8,6}}$$

Z obou modelů je přímka „vítězem“.

Testování regresních parametrů

■ Uvažujme model s p parametry,
např. jednoduchá lineární regrese:

$$Y = \beta_0 + \beta_1 X \rightarrow p=2$$

např. jednoduchá kvadratická regrese:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \rightarrow p=3$$

např. 2-násobná lineární regrese:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z \rightarrow p=3$$

Testování regresních parametrů

$$H_0: \beta_1 = \dots = \beta_{p-1} = 0$$

(model jako celek je nevýznamný)

versus $H_1: \text{non } H_0$

(aspoň jeden parametr modelu je významný);

v testu není zahrnut β_0

Testování - možnosti

jednoduchá lineární regrese ($p=2$):

$$H_0: \beta_1=0 \quad \text{versus} \quad H_1: \beta_1 \neq 0$$

H_0místo lineární funkce by jako model „bývala stačila“ funkce konstantní ($Y=\beta_0$) aneb „přímka s nulovou směrnici“;

H_1do vhodného modelu je potřeba zahrnout nenulovou „směrnici“

Testování - možnosti

jednoduchá kvadratická regrese ($p=3$):

$H_0: \beta_1 = \beta_2 = 0$ versus $H_1: \text{non } H_0$

H_0místo kvadratické funkce by jako model „bývala stačila“ funkce konstantní;

H_1do vhodného modelu je potřeba zahrnout alespoň jeden z obou testovaných parametrů (lineární či kvadratický)

Testování - možnosti

2-násobná lineární regrese ($p=3$):

$$H_0: \beta_1 = \beta_2 = 0 \quad \text{versus} \quad H_1: \text{non } H_0$$

H_0místo lineární funkce 2 proměnných (X a Z) by jako model „bývala stačila“ funkce konstantní;

H_1do vhodného modelu je potřeba zahrnout alespoň jeden z obou testovacích parametrů (aneb proměnnou X či Z)

Testování regresních parametrů
 atd. (modely složitější, s více parametry).
 Provedení testu regresních modelů:

$$F = \frac{(Q_Y - Q_e) / (p-1)}{Q_e / (n-p)} \quad \leftarrow = s_e^2$$

$W = \langle F_{1-\alpha}(p-1, n-p); \infty \rangle \dots$ celkový F test

Testování regresních parametrů

■ Př: Data - Westwood Company (pokr.)

Pomocí reziduálního rozptylu byl ze dvou modelů vybrán lineární. Jde ale o model VÝZNAMNÝ (tj. „dobrý“)?

$$F = 1813,3$$

$$p = 1,02 \cdot 10^{-10} < 0,05$$

zamítáme $H_0 \rightarrow$ model JE VÝZNAMNÝ

Testování regresních parametrů

The screenshot shows the Microsoft Excel interface with the 'Data' tab selected in the ribbon. The 'Analýza dat' (Data Analysis) button is highlighted. Below the ribbon, the 'Data' tab group contains options like 'Seskupit', 'Oddělit', and 'Souhrn'. The 'Analýza' group contains 'Citlivostní analýza' and 'List prognózy'. The 'Data Analysis' dialog box is open, showing a list of analytical tools. The 'Regrese' (Regression) option is selected at the bottom of the list. The background spreadsheet shows columns A and B with data points.

	A	B
1	x	y
2	30	73
3	20	50
4	60	128
5	80	170
6	40	87
7	50	108
8	60	135
9	30	69
10	70	148
11	60	132

Analýza dat

Analytické nástroje:

- Kovariance
- Popisná statistika
- Exponenciální vyrovnání
- Dvouvýběrový F-test pro rozptyl
- Fourierova analýza
- Histogram
- Klouzavý průměr
- Generátor pseudonáhodných čísel
- Pořadová statistika a percentily
- Regrese**

OK Storno Nápoředa

Testování regresních parametrů

	A	B
1	x	y
2	30	73
3	20	50
4	60	128
5	80	170
6	40	87
7	50	108
8	60	135
9	30	69
10	70	148
11	60	132

Regrese

Vstup

Vstupní oblast y: SBS1:SBS11

Vstupní oblast x: SAS1:SAS11

☒ Popisky ☐ Konstanta je nula

☐ Hladina spolehlivosti 95 %

Možnosti výstupu

☐ Výstupní oblast:

☒ Nový list:

☐ Nový sešit

OK

Storno

Nápověda

Testování regresních parametrů

VÝSLEDEK						
<i>Regresní statistika</i>						
Násobné R	0,997801	korel.koef.				
Hodnota spolehlivosti R	0,995608	index determinace				
Nastavená hodnota spole	0,995059	adjustovaný index determinace				
Chyba stř. hodnoty	2,738613					
Pozorování	10					
ANOVA		df			p-hodnota (celek)	
		<i>Rozdíl</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Významnost F</i>
Regrese	1	13600	13600	1813,333333	1,01959E-10	
Rezidua	8	60	7,5			
Celkem	9	13660				
	<i>Koeficienty</i>	<i>stř. hod</i>	<i>t Stat</i>	<i>Hodnota P</i>	<i>Dolní 95%</i>	<i>Horní 95%</i>
Hranice	10,00000	2,5029	3,995302	0,00397576	4,228211282	15,77179
x	2,00000	0,047	42,58325	1,01959E-10	1,891694315	2,108306

Regrese pomocí R:

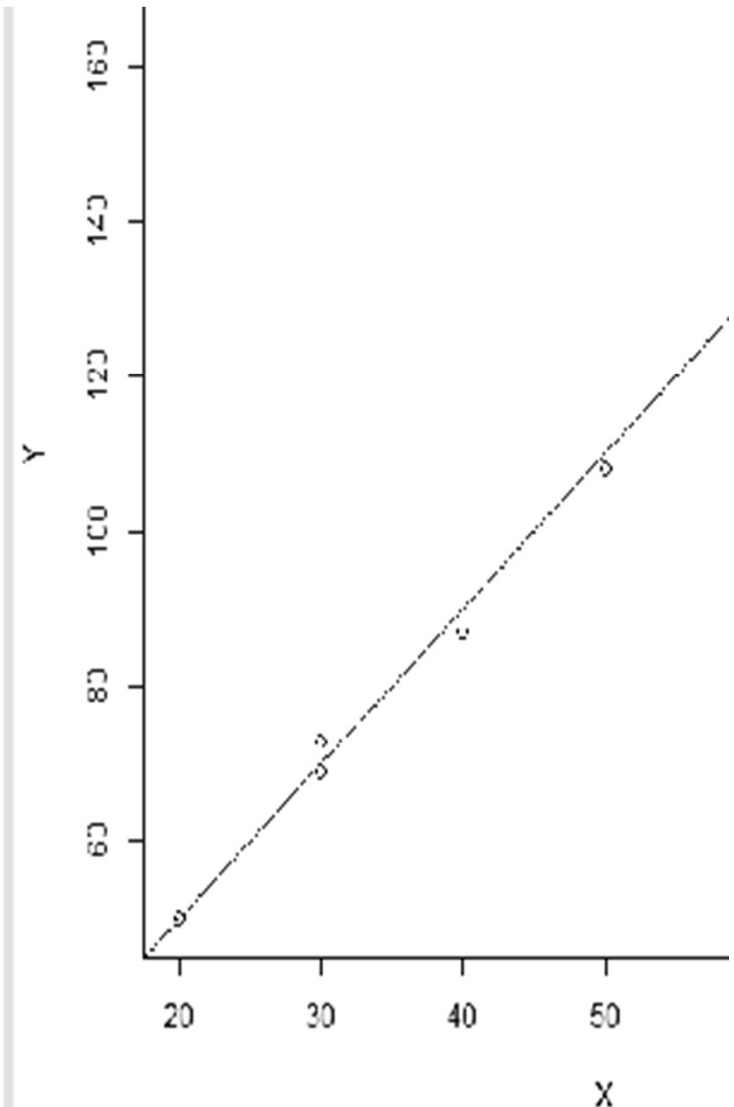
```
> X=c(30,20,60,80,40,50,60,30,70,60)
> Y=c(73,50,128,170,87,108,135,69,148,132)
> MODEL=lm(Y~X)
> summary(MODEL)
Call:
lm(formula = Y ~ X)

Residuals:
    Min       1Q   Median       3Q      Max
   -3.0    -2.0    -0.5     1.5     5.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.00000    2.50294   3.995  0.00398 **
X           2.00000    0.04697  42.583 1.02e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-squared:  0.9956,    Adjusted R-squared:  0.9951
F-statistic: 1813 on 1 and 8 DF,  p-value: 1.02e-10

> R=cor(X,Y)
> R*R
[1] 0.9956076
> plot(Y~X)
> abline(MODEL)
```



Regrese pomocí R:

■ Pokrač.: Nebude lepší kvadrat.model?

```
> X2=X*X
> MODEL2=lm(Y~X+X2)
> summary(MODEL2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.043e+01	6.703e+00	1.555	0.163784
X	1.980e+00	2.957e-01	6.695	0.000279 ***
X2	2.055e-04	2.966e-03	0.069	<u>0.946717</u>

Residual standard error: 2.927 on 7 degrees of freedom
Multiple R-squared: 0.9956, Adjusted R-squared: 0.9944
F-statistic: 793.9 on 2 and 7 DF, p-value: 5.603e-09

■ Je také významný, ale u lin.modelu vyšlo $R^2_{adj}=0,9951$... byl tedy vhodnější

Korelace – poznámky

Korelační koeficienty lze určovat i u lineárních modelů s více regresory →

- koeficient mnohonásobné korelace

(míra závislosti Y na všech regresorech)

- parciální (dílčí) korelační koeficienty

(míra závislosti Y vždy na jednom z regresorů při „zohlednění“ regresorů zbylých)

Regrese – poznámky

U modelů s více regresory lze vybrat postupně model jen s významnými regresory - kroková regrese (stepwise)

- regrese typu forward

(přidávání regresorů, dokud je model jako celek stále ještě významný)

- regrese typu backward

(naopak ubírání regresorů)

Regrese – poznámky

Regresor může být i kategoriální (např. nominální se 3 kategoriemi), pak jej vyjádříme jako dummy proměnné - zde např. V a W :

- $V=1$ pro kategorii 1, jinak $V=0$
- $W=1$ pro kategorii 2, jinak $W=0$
- (čili kategorii 3 odpovídá $V=W=0$)

Model může být: $Y = \beta_0 + \beta_1 V + \beta_2 W$

- pak např. β_1 odpovídá rozdílu efektu mezi 1. a 3. kategorií