

Regresní přímka

Motivace: Cíl regresní analýzy - popsat závislost hodnot veličiny Y na hodnotách veličiny X.

Nutnost vyřešení dvou problémů:

- a) jaký typ funkce se použije k popisu dané závislosti;
- b) jak se stanoví konkrétní parametry daného typu funkce?

ad a) Při určení typu funkce je třeba provést teoretický rozbor zkoumané závislosti.

Často však nemáme dostatek informací k provedení teoretického rozboru. Pak se snažíme odhadnout typ funkce pomocí dvourozměrného tečkového diagramu.

Zde se omezíme na polynomiální funkci 1. stupně $y = \beta_0 + \beta_1 x$.

ad b) Odhady b_0, b_1 neznámých parametrů β_0, β_1 získáme na základě dvourozměrného datového souboru $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$ metodou

nejmenších čtverců, tj. z podmínky, aby součet čtverců odchylek zjištěných a odhadnutých hodnot byl minimální.

Specifikace klasického modelu regresní přímky

Máme regresní model $Y = \beta_0 + \beta_1 X + \varepsilon$, kde

$y = \beta_0 + \beta_1 x$ - **teoretická regresní přímka** (deterministická složka modelu).

(Parametr β_0 interpretujeme jako teoretickou hodnotu Y při $x = 0$ a β_1 udává změnu Y , když X se změní o jednotku.)

Složka ε - **náhodná složka** modelu. Je to náhodná odchylka od deterministické závislosti Y na X . Popisuje závislost vysvětlované proměnné na neznámých nebo nepozorovaných proměnných a popisuje i vliv náhody. Nelze ji funkčně vyjádřit.

Veličina Y - **závisle proměnná (též vysvětlovaná) veličina**.

Veličina X - **nezávisle proměnná (též vysvětlující) veličina**.

Pořídíme n dvojic pozorování $(x_1, y_1), \dots, (x_n, y_n)$, tj. dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$.

Pro $i = 1, \dots, n$ platí: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

O náhodných odchylkách $\varepsilon_1, \dots, \varepsilon_n$ předpokládáme, že

- $E(\varepsilon_i) = 0$ (odchylky nejsou systematické)
- $D(\varepsilon_i) = \sigma^2 > 0$ (všechna pozorování jsou prováděna s touž přesností)
- $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$ (mezi náhodnými odchylkami neexistuje žádný lineární vztah)
- $\varepsilon_i \sim N(0, \sigma^2)$

V tomto případě hovoříme o **klasickém modelu regresní přímky**.

Předpoklady použití regresní přímky:

- Závislost Y na X má lineární charakter.
- Pro celý rozsah uvažovaných hodnot nezávisle proměnné X je reziduální rozptyl s^2 konstantní (hovoříme o homoskedasticitě a znamená to, že variabilita hodnot závisle proměnné veličiny Y kolem regresní přímky je stejná pro všechny uvažované hodnoty nezávisle proměnné veličiny X).
- Hodnoty závisle proměnné veličiny Y mají normální rozložení pro dané hodnoty x_i a jsou stochasticky nezávislé (to souvisí s uspořádáním experimentu).

Poznámka: Menší odchylky od normality a homoskedasticity je možno tolerovat.

Označení

$Q(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ - rozptyl hodnot znaku Y kolem přímky $y = \beta_0 + \beta_1 x$

b_0, b_1 - odhady regresních parametrů β_0, β_1 (získáme je metodou nejmenších čtverců, tj. z podmínky, že výraz $Q(\beta_0, \beta_1)$ nabývá svého minima pro $\beta_j = b_j, j = 0, 1$)

$y = b_0 + b_1 x$ - empirická regresní přímka

$\hat{y}_i = b_0 + b_1 x_i$ - regresní odhad i-té hodnoty veličiny Y (i-tá predikovaná hodnota veličiny Y)

$e_i = y_i - \hat{y}_i$ - i-té reziduum

$S_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ - reziduální součet čtverců

$s^2 = \frac{S_E}{n-2}$ - odhad rozptylu σ^2

$S_R = \sum_{i=1}^n (\hat{y}_i - m_2)^2$ - regresní součet čtverců ($m_2 = \frac{1}{n} \sum_{i=1}^n y_i$)

$S_T = \sum_{i=1}^n (y_i - m_2)^2$ - celkový součet čtverců ($S_T = S_R + S_E$)

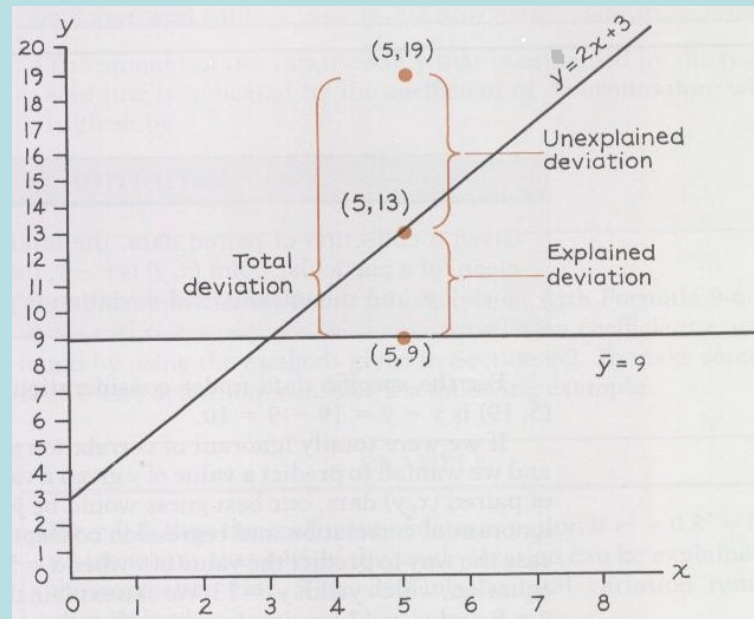
Význam jednotlivých typů součtů čtverců

Předpokládejme, že máme dvourozměrný datový soubor, v němž průměr hodnot závisle proměnné veličiny Y je 9 a závislost veličiny Y na veličině X je popsána regresní přímkou $y = 2x + 3$. Dvourozměrný tečkový diagram obsahuje bod o souřadnicích (5, 19), který pochází z datového souboru. Na regresní přímce leží bod o souřadnicích (5, 13).

Odchylka zjištěné hodnoty 19 od průměru 9 je v obrázku označena „Total deviation“ a po umocnění je to jedna ze složek celkového součtu čtverců S_T , tj. složka $y_i - \bar{y}$.

Odchylka zjištěné hodnoty 19 od hodnoty 13 na regresní přímce je v obrázku označena „Unexplained deviation“ a po umocnění je to jedna ze složek reziduálního součtu čtverců S_E , tj. složka $y_i - \hat{y}_i$.

Odchylka hodnoty 13 na regresní přímce od průměru 9 je v obrázku označena „Explained deviation“ a po umocnění je to jedna ze složek regresního součtu čtverců S_R , tj. složka $\hat{y}_i - \bar{y}$.



System normálních rovnic pro regresní přímku

Uvažujeme regresní model $Y = \beta_0 + \beta_1 X + \varepsilon$.

System normálních rovnic pro odhad regresních parametrů β_0 a β_1 získáme derivováním výrazu

$$Q(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

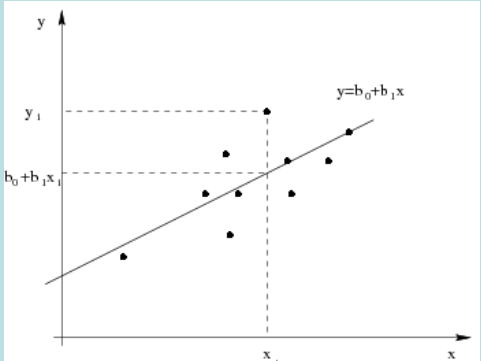
parciálně podle β_0 a β_1 :

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 2 \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-1) = 0, \quad \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 2 \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-x_i) = 0$$

Řešením tohoto systému získáme odhady $b_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$

Po jednoduchých úpravách dospějeme ke tvaru $b_1 = \frac{s_{12}}{s_1^2}$, kde s_{12} je kovariance hodnot $(x_i, y_i), i = 1, \dots, n$ a s_1^2 je rozptyl

hodnot x_1, \dots, x_n . Dále dostáváme $b_0 = m_2 - b_1 m_1$, tedy regresní přímku můžeme vyjádřit ve tvaru $y = m_2 + \frac{s_{12}}{s_1^2} (x - m_1)$.



Index determinace regresní přímky

Kvalitu regresních modelů posuzujeme mj. pomocí indexu determinace: $ID^2 = \frac{S_R}{S_T}$, kde

$S_R = \sum_{i=1}^n (\hat{y}_i - m_2)^2$ je regresní součet čtverců a $S_T = \sum_{i=1}^n (y_i - m_2)^2$ je celkový součet čtverců.

Pro regresní přímku má regresní součet čtverců tvar:

$$S_R = \sum_{i=1}^n (y_i - m_2)^2 = \sum_{i=1}^n \left[m_2 + \frac{s_{12}}{s_1} (x_i - m_1) - m_2 \right]^2 = \frac{s_{12}^2}{s_1^4} \sum_{i=1}^n (x_i - m_1)^2 = \frac{s_{12}^2}{s_1^4} \cdot n \cdot \frac{1}{n} \sum_{i=1}^n (x_i - m_1)^2 = n \frac{s_{12}^2}{s_1^4} s_1^2 = n \frac{s_{12}^2}{s_1^2}.$$

Celkový součet čtverců $S_T = \sum_{i=1}^n (y_i - m_2)^2 = ns_2^2$, tedy index determinace

$$ID^2 = \frac{S_R}{S_T} = \frac{n \frac{s_{12}^2}{s_1^2}}{ns_2^2} = \frac{s_{12}^2}{s_1^2 s_2^2} = r_{12}^2$$

Vidíme tedy, že v případě regresní přímky **index determinace je roven kvadrátu koeficientu korelace**. Index determinace nabývá hodnot z intervalu $\langle 0,1 \rangle$. Často se vyjadřuje v procentech a informuje nás o tom, jakou část variability hodnot závisle proměnné veličiny Y vyčerpává regresní model.

Sdružené regresní přímky

Předpokládáme, že obě veličiny Y a X jsou náhodné a veličina X nezávisí na náhodné složce ε . Pak jde o případ oboustranné závislosti.

Závislost Y na X vystihuje regresní model $y = \beta_0 + \beta_1 x + \varepsilon$,

závislost X na Y vystihuje regresní model $x = \alpha_0 + \alpha_1 y + \delta$.

Odhady a_0, a_1 regresních parametrů α_0, α_1 v modelu $x_i = \alpha_0 + \alpha_1 y_i + \delta_i$ získáme opět MNČ ve tvaru

$$a_1 = \frac{s_{12}}{s_2}, a_0 = m_1 - a_1 m_2 = m_1 - \frac{s_{12}}{s_2} m_2.$$

Empirická regresní přímka závislosti X na Y má tedy rovnici:

$$x = m_1 + \frac{s_{12}}{s_2} (y - m_2).$$

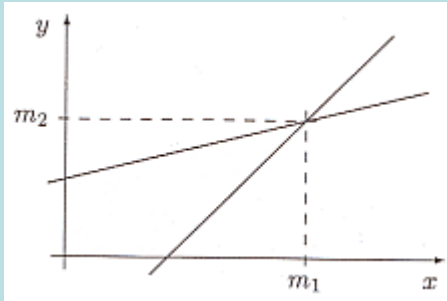
Obě empirické regresní přímky $y = b_0 + b_1 x$, $x = a_0 + a_1 y$ se nazývají **sdružené regresní přímky** a odhady regresních parametrů b_1, a_1 se nazývají **odhady párově sdružených regresních parametrů**.

Je zřejmé, že $b_1 a_1 = r_{12}^2$. Rovnice sdružených regresních přímek můžeme tedy psát ve tvaru:

$$y = m_2 + \frac{s_{12}}{s_1} (x - m_1), \quad y = m_1 + \frac{1}{r_{12}} \frac{s_2}{s_1} (x - m_2).$$

Vlastnosti sdružených regresních přímek

a) Sdružené regresní přímky se protínají v bodě o souřadnicích \bar{x}, \bar{y} (tj. v těžišti dvourozměrného tečkového diagramu).



b) Je-li $r_{12} = 0$ (tj. náhodné veličiny X, Y jsou nekorelované), pak sdružené regresní přímky mají rovnice $y = \bar{y}$, $x = \bar{x}$ (tj. jsou to kolmice rovnoběžné se souřadnými osami).

c) Je-li $r_{12}^2 = 1$ (tj. mezi náhodnými veličinami X, Y existuje úplná lineární závislost), pak sdružené regresní přímky splývou a $a_1 = \frac{1}{b_1}$.

d) Je-li $0 < r_{12}^2 < 1$, pak sdružené regresní přímky se liší a svírají úhel, který je tím menší, čím je těsnější lineární závislost veličin X, Y.

e) Označíme-li φ úhel, který svírají sdružené regresní přímky, pak z předešlých úvah plyne:

$\cos \varphi = 0 \Leftrightarrow$ mezi X a Y neexistuje žádná lineární závislost;

$\cos \varphi = 1 \Leftrightarrow$ mezi X a Y existuje úplná přímá lineární závislost;

$\cos \varphi = -1 \Leftrightarrow \Leftrightarrow$ mezi X a Y existuje úplná nepřímá lineární závislost.

Příklad:

Z fiktivního základního souboru všech vzorků oceli odpovídajících „všem myslitelným tavbám“ bylo do laboratoře dodáno 60 vzorků a zjištěny a hodnoty proměnné X – mez plasticity a Y – mez pevnosti. Datový soubor má tvar:

154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

- Určete regresní přímku meze pevnosti na mez plasticity.
- Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
- Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
- Vypočtěte index determinace a interpretujte ho.
- Najděte reziduální součet čtverců a odhad rozptylu náhodných odchylek.
- Určete regresní přímku meze plasticity na mez pevnosti.
- Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
- Obě regresní přímky zakreslete do téhož dvourozměrného tečkového diagramu.

Řešení v systému STATISTICA:

Ad a) Odhad parametrů 1. regresní přímky:

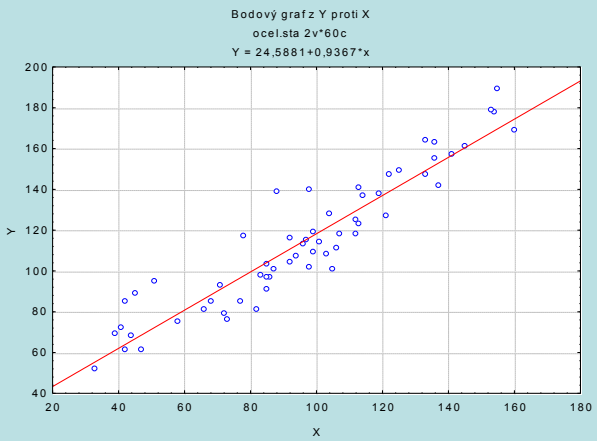
Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

Výsledky regrese se závislou proměnnou : Y (ocel.sta)						
R= ,93454811 R2= ,87338017 Upravené R2= ,87119707						
F(1,58)=400,06 p<0,0000 Směrod. chyba odhadu : 11,768						
N=60	Beta	Sm.chyba beta	B	Sm.chyba B	t(58)	Úroveň p
Abs.člen			24,58814	4,740272	5,18707	0,000003
X	0,934548	0,046724	0,93668	0,046830	20,00160	0,000000

Vidíme, že $y = 24,5881 + 0,9367x$.

Ad b) Zakreslení regresní přímky do dvourozměrného tečkového diagramu:

Grafy – Bodové grafy – Proměnné X, Y – OK – OK.



Ad c) Výpočet predikované hodnoty: Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi - Předpovědi závisle proměnné X: 60 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď: 80,79

Proměnná	Předpovězené hodnoty (ocel.sta) proměnné: Y		
	b-v áha	Hodnota	b-v áha * Hodnot
X	0,936679	60,00000	56,20071
Abs. člen			24,58814
Předpověď			80,78885
-95,0%LS			76,25426
+95,0%LS			85,32344

Regresní odhad meze pevnosti pro mez plasticity 60 je tedy 80,8.

Ad d) Index determinace najdeme ve výstupní tabulce regrese pod označením R2:

Výsledky regrese se závislou proměnnou : Y (ocel.sta) R= ,93454811 R2= ,87338017 Upravené R2= ,87119707 F(1,58)=400,06 p<0,0000 Směrod. chyba odhadu : 11,768						
N=60	Beta	Sm.chyba beta	B	Sm.chyba B	t(58)	Úroveň p
Abs.člen			24,58814	4,740272	5,18707	0,000003
X	0,934548	0,046724	0,93668	0,046830	20,00160	0,000000

Vidíme, že variabilita meze pevnosti je regresní přímkou vyčerpána z 87,3 %.

Ad e) Reziduální součet čtverců a odhad rozptylu najdeme v tabulce ANOVA: Vrátime se do Výsledky – Vícenásobná regrese – na záložce Detailní výsledky zvolíme ANOVA (Celk. vhodnost modelu)

Efekt	Analýza rozptylu (ocel.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	55400,60	1	55400,60	400,0641	0,000000
Rezid.	8031,80	58	138,48		
Celk.	63432,40				

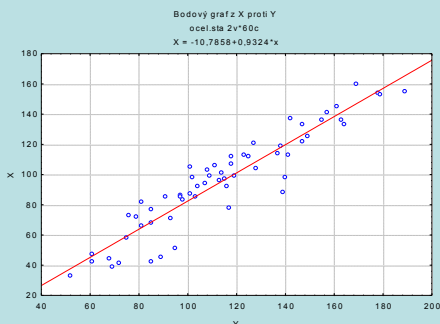
Vidíme, že reziduální součet čtverců je 8031,8 a reziduální rozptyl nabývá hodnoty 138,48.

Ad f) Výsledky pro 2. regresní přímku:

Výsledky regrese se závislou proměnnou : X (ocel.sta)						
R= ,93454811 R2= ,87338017 Upravené R2= ,87119707						
F(1,58)=400,06 p<0,0000 Směrod. chyba odhadu : 11,741						
N=60	Beta	Sm.chyba beta	B	Sm.chyba B	t(58)	Úroveň p
Abs. člen			-10,7858	5,544250	-1,94540	0,056579
Y	0,934548	0,046724	0,9324	0,046617	20,00160	0,000000

Vidíme, že $x = -10,7858 + 0,9324y$.

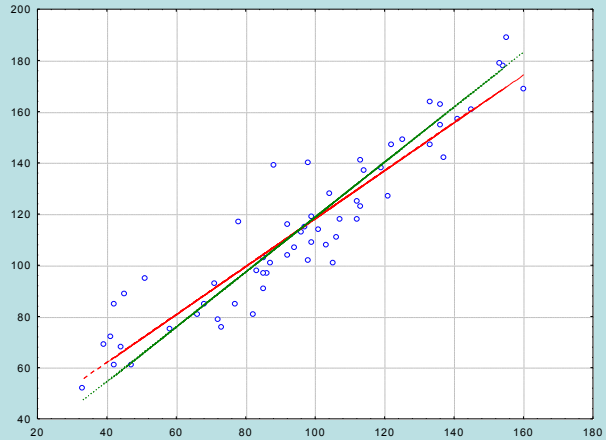
Ad g) Dvourozměrný tečkový diagram se zakreslenou 2. regresní přímkou



Ad h) Nakreslení sdružených regresních přímek do jednoho diagramu:

K datovému souboru ocel.sta přidáme dvě nové proměnné y1 a y2. Do proměnné y1 uložíme predikované hodnoty meze pevnosti na mezi plasticity (do Dlouhého jména proměnné y1 napíšeme $=24,58814 + 0,93668*x$ a do Dlouhého jména proměnné y2 napíšeme $=(x+10,7858)/0,9324$

Grafy – Bodové grafy – zaškrtneme Vícenásobný – Proměnné X: X, Y: Y, y1, y2 – OK. Ve vytvořeném grafu pak vypneme zobrazování značek pro y1, y2 a naopak zapneme Spojnici.



Intervaly spolehlivosti pro regresní parametry

Bodový odhad směrnice β_1 směrnice regresní přímky $y = \beta_0 + \beta_1 x$ je dán vzorcem $b_1 = \frac{s_{12}}{s_1}$ a bodový odhad úseku β_0 je dán vzorcem $b_0 = m_2 - b_1 m_1$. Vedle bodových odhadů nás však zajímají též intervalové odhady regresních parametrů β_0, β_1 .

Označme $s_{b_j} = s\sqrt{v_{jj}}$ - **směrodatná chyba odhadu b_j** , kde

$$v_{00} = \frac{s_1^2 + m_1^2}{ns_1^2}, \quad v_{11} = \frac{1}{ns_1^2}$$

Pro $j = 0, 1$ statistika $T_j = \frac{b_j - \beta_j}{s_{b_j}} \sim t_{n-2}$, tedy $100(1 - \alpha)\%$ interval spolehlivosti pro β_j má meze:

$$d_j = b_j - t_{1-\frac{\alpha}{2}; n-2} s_{b_j}, \quad h_j = b_j + t_{1-\frac{\alpha}{2}; n-2} s_{b_j}, \quad j = 0, 1.$$

(S intervaly spolehlivosti souvisí relativní chyby odhadů regresních parametrů: $\frac{1}{2} \left| \frac{h_j - d_j}{b_j} \right| 100\%$.

Relativní chyba odhadu by neměla přesáhnout 10 %.)

Příklad:

V tabulce jsou výnosy technické cukrovky v tunách na ha od roku 2000 do roku 2007.

i	rok	cukrovka technická
1	2000	45,83
2	2001	45,41
3	2002	49,45
4	2003	45,20
5	2004	50,34
6	2005	53,31
7	2006	51,48
8	2007	53,25

Předpokládejte, že závislost výnosu cukrovky na roku lze vyjádřit regresní přímkou $y = \beta_0 + \beta_1 x$.

- MNČ najděte odhady neznámých regresních parametrů β_0 , β_1 .
- Sestrojte 95% intervaly spolehlivosti pro regresní parametry β_0 , β_1 .
- Najděte relativní chyby odhadů regresních parametrů β_0 , β_1 .
- Regresní přímkou zakreslete do dvourozměrného tečkového diagramu.

Řešení:

Vytvoříme datový soubor se dvěma proměnnými rok, Y a osmi případy.

Získání odhadů b_0 , b_1 :

Statistiky – Vícerozměrná regrese – Závisle proměnná rok, nezávisle proměnné Y - OK – OK – Výpočet: Výsledky regrese.

Výsledky regrese se závislou proměnnou : Y (cukrovka_tecnicka.sta)						
R= ,84604287 R2= ,71578853 Upravené R2= ,66841995						
F(1,6)=15,111 p<,00810 Směrod. chyba odhadu : 1,9651						
N=8	b*	Sm.chyba z b*	b	Sm.chyba z b	t(6)	p-hodn.
Abs.člen			-2312,22	607,4943	-3,80616	0,008903
rok	0,846043	0,217643	1,18	0,3032	3,88729	0,008102

Výpočet mezi intervalu spolehlivosti a relativních chyb odhadů:

K výstupní tabulce přidáme tři nové proměnné DM, HM a chyba.

Do Dlouhého jméne proměnné DM napíšeme

$$=v3-v4*VStudent(0,975;6)$$

Do Dlouhého jméne proměnné HM napíšeme

$$=v3+v4*VStudent(0,975;6)$$

Do Dlouhého jména proměnné chyba napíšeme

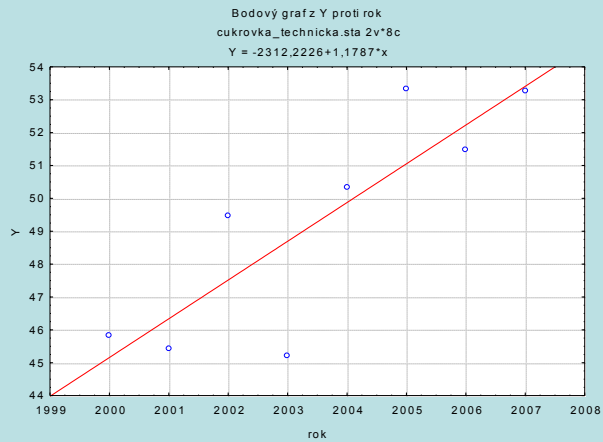
$$=100*abs(0,5*(v8-v7)/v3)$$

Výsledky regrese se závislou proměnnou : Y (cukrovka_tecnicka.sta)									
R= ,84604287 R2= ,71578853 Upravené R2= ,66841995									
F(1,6)=15,111 p<,00810 Směrod. chyba odhadu : 1,9651									
N=8	b*	Sm.chyba z b*	b	Sm.chyba z b	t(6)	p-hodn.	DM =v3-v4*V	HM =v3+v4*V	chyba =100*abs
Abs.člen			-2312,22	607,4943	-3,80616	0,008903	-3798,71	-825,738	64,28814
rok	0,846043	0,217643	1,18	0,3032	3,88729	0,008102	0,436747	1,920634	62,94643

S pravděpodobností 95% se bude úsek β_0 regresní přímky nacházet v intervalu (-3798,71; -825,738). Odhad b_0 úseku β_0 je zatížen relativní chybou 64,3%.

S pravděpodobností 95% se bude směrnice β_1 regresní přímky nacházet v intervalu (-3798,71; -825,738). Odhad b_1 úseku β_1 je zatížen relativní chybou 62,9%.

Vytvoření dvourozměrného tečkového diagramu s proloženou regresní přímkou:
Grafy – Bodové grafy – Proměnné X: rok, Y: Y – OK – OK.



Testování hypotézy o regresních parametrech

Na hladině významnosti α pro $j = 0, 1$ testujeme hypotézu

$H_0: \beta_j = c$ proti $H_1: \beta_j \neq c$, kde c je reálná konstanta.

Testová statistika: $T_j = \frac{b_j - c}{s_{b_j}}$ má rozložení $t(n-2)$, pokud H_0 platí.

Kritický obor: $w = (-\infty, -t_{1-\alpha/2, n-2}) \cup (t_{1-\alpha/2, n-2}, \infty)$

$T_j \in w \Rightarrow H_0$ zamítáme na hladině významnosti α .

Je-li $c = 0$, jde o **testy významnosti regresních parametrů** (dílní t-testy)

Příklad:

V předešlém příkladě, kde byla modelována závislost výnosu technické cukrovky na roku regresní přímkou, proveďte dílní t-testy významnosti jednotlivých regresních parametrů.

Řešení:

Stačí interpretovat výstupní tabulku vícenásobné regrese:

Výsledky regrese se závislou proměnnou : Y (cukrovka_tecnicka.sta)						
R= ,84604287 R2= ,71578853 Upravené R2= ,66841995						
F(1,6)=15,111 p<,00810 Směrod. chyba odhadu : 1,9651						
N=8	b*	Sm.chyba z b*	b	Sm.chyba z b	t(6)	p-hodn.
Abs. člen			-2312,22	607,4943	-3,80616	0,008903
rok	0,846043	0,217643	1,18	0,3032	3,88729	0,008102

Sloupec označený t(6) obsahuje realizace testových statistik a sloupec p-hodn. pak odpovídající p-hodnoty. V obou případech jsou p-hodnoty menší než 0,05, tedy na hladině významnosti 0,05 zamítáme hypotézy o nevýznamnosti úseku a směrnice regresní přímky.

Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme

$H_0: \beta_1 = 0$ proti $H_1: \beta_1 \neq 0$.

(Nulová hypotéza říká, že dostačující je model konstanty.)

Testová statistika: $F = \frac{S_R}{S_E / (n-2)}$ má rozložení $F(1, n-2)$, pokud H_0 platí.

Kritický obor: $w = (F_{1-\alpha/2}, n-2, \infty)$.

$F \in w \Rightarrow H_0$ zamítáme na hladině významnosti α .

Výsledky F-testu zapisujeme do tabulky, která se nazývá tabulka analýzy rozptylu:

zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R	1	$S_R/1$	$\frac{S_R}{S_E / (n-2)}$
reziduální	S_E	n-2	$S_E/(n-2)$	-
celkový	S_T	n-1	-	-

V případě regresní přímky je celkový F-test ekvivalentní dílčímu t-testu významnosti směrnice regresní přímky.

V příkladu s technickou cukrovkou má tabulka ANOVA tvar:

Efekt	Analýza rozptylu (cukrovka_tecnicka.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	58,35107	1	58,35107	15,11104	0,008102
Rezid.	23,16892	6	3,86149		
Celk.	81,51999				

Vidíme, že testová statistika F se realizuje hodnotou 15,111 a odpovídající p-hodnota je 0,0081, tedy na hladině významnosti 0,05 zamítáme hypotézu, že dostačující je model konstanty.

Interval spolehlivosti pro teoretickou regresní přímku

V uvažovaném lineárním modelu $y = \beta_0 + \beta_1 x + \varepsilon$ můžeme na základě n dvojic pozorování (x_i, y_i) , $i = 1, \dots, n$ získat jak bodové, tak intervalové odhady neznámých regresních parametrů β_0, β_1 . Lze však spočítat též meze $100(1-\alpha)$ % intervalu spolehlivosti pro teoretickou regresní přímku při zadané hodnotě x_0 :

$$d = \hat{y}_0 - s \cdot t_{1-\alpha/2, n-2} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad h = \hat{y}_0 + s \cdot t_{1-\alpha/2, n-2} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Při spojitě změně argumentu x_0 mezní hodnoty tohoto intervalu spolehlivosti pro teoretickou regresní přímku vytvoří **100(1- α) % pás spolehlivosti kolem regresní přímky**.

(Tento pás spolehlivosti však nelze interpretovat tak, že pokrývá celou regresní přímku s pravděpodobností $1 - \alpha$, pouze ukazuje na šířku intervalu spolehlivosti pro vypočtenou hodnotu z modelu při pevně zvolené hodnotě argumentu.)

Predikční interval spolehlivosti pro budoucí pozorování

Uvažujme nyní situaci, že na základě pozorovaných hodnot najdeme bodový odhad ve tvaru $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \dots, n$. Nyní chceme provést ještě jedno další pozorování pro bod x_0 a zajímá nás otázka, jak daleko může být tato další nová pozorovaná hodnota od vypočtené hodnoty $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$? Na tuto otázku odpovídá tzv. predikční interval, který při zadané hodnotě x_0 má meze:

$$d = \hat{y}_0 - s \cdot t_{1-\alpha/2, n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad h = \hat{y}_0 + s \cdot t_{1-\alpha/2, n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Při spojitě změně argumentu x_0 mezní hodnoty predikčního intervalu na hladině $1-\alpha$ vytvoří **100(1- α)% predikční pás kolem regresního přímky**.

Tento predikční pás nelze interpretovat tak, že pokrývá celou regresní přímku s pravděpodobností $1-\alpha$, pouze ukazuje na šířku intervalu spolehlivosti, v němž se bude nacházet další jedno pozorování pro zvolenou hodnotu argumentu x_0 .

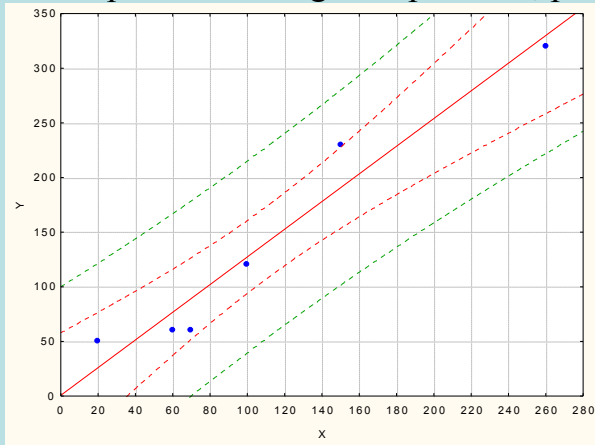
Srovnání intervalu spolehlivosti a predikčního intervalu při zadané hodnotě x_0 :

- oba intervaly jsou nejužší v místě $x_0 = m_1$,
- interval spolehlivosti pro dané x_0 je vždy užší než odpovídající predikční interval,
- predikční interval je určen pro individuální pozorování, zatímco interval spolehlivosti je určen pro hodnoty ležící na regresní přímce,
- s rostoucím rozsahem výběru se zmenšuje šířka obou intervalů.

Příklad: Pro data z příkladu o závislosti výnosů technické cukrovky na letech nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou, 95% pásem spolehlivosti kolem regresní přímky a 95% predikčním pásem.

Řešení: Grafy – Bodové grafy – Proměnné rok, Y – OK – na záložce Details zvolíme Regresní pásy, zaškrtneme Spolehl. – OK. Dále potřebujeme přidat predikční pás. 2x klikneme na vytvořený graf a v Možnostech grafu vybereme Regresní pásy – Přidat nový pár pásů – Typ Predikční – ve Vzoru změňme barvu na zelenou – OK.

Data s proloženou regresní přímkou, pásy spolehlivosti (červeně) a predikčními pásy (zeleně)



Výpočet střední absolutní procentuální chyby predikce (MAPE)

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Při použití MAPE nezáleží na použitých jednotkách, v nichž byly měřeny veličiny X, Y, protože se jedná o relativní míru. Systém STATISTICA MAPE neposkytuje, tuto chybu musíme vypočítat.

Příklad: Pro údaje z o mezích pevnosti a plasticity oceli vypočtete chybu MAPE při proložení regresní přímkou.

Řešení: Statistika – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné X - OK – OK – zvolíme Rezi-
dua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua & předpovědi – vybereme proměnnou y - OK.
K vzniklému datovému souboru přidáme jednu novou proměnnou, nazveme ji chyba a do jejího Dlouhého jména napíšeme
 $=100 * \text{abs}((v1-v2)/v1)$

Pomocí Statistika – Základní statistiky/tabulky – Popisné statistiky zjistíme průměr proměnné chyba. V našem případě je MAPE 8,72 %.

Analýza reziduí

Rezidua považujeme za odhady náhodných odchylek a klademe na ně stejné požadavky jako na náhodné odchylky, tj.

mají být nezávislá,

mají být normálně rozložená,

mají mít nulovou střední hodnotu,

mají mít konstantní rozptyl (tj. jsou homoskedastická).

Nezávislost reziduí (autokorelaci) posuzujeme např. pomocí Durbinovy – Watsonovy statistiky, která by se měla nacházet v intervalu $\langle 1,4;2,6 \rangle$ (to je ovšem pouze orientační vodítko, korektní postup spočívá v porovnání této statistiky s tabelovanou kritickou hodnotou).

Normalitu reziduí ověřujeme pomocí testů normality (např. Lilieforsovou variantou Kolmogorovova – Smirnovova testu nebo Shapirovým – Wilksovým testem) či graficky pomocí N-P plotu.

Testování nulovosti střední hodnoty reziduí provádíme pomocí jednovýběrového t-testu.

Homoskedasticitu reziduí posuzujeme pomocí grafu závislosti reziduí na predikovaných hodnotách. V tomto grafu by rezidua měla být rovnoměrně rozptýlena.

Příklad: Proveďte analýzu reziduí pro příklad s modelováním závislosti meze pevnosti oceli na mezi plasticity oceli.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

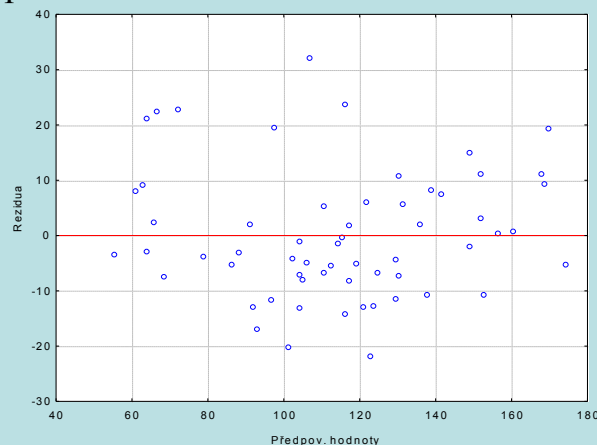
Statistiky – Vícenásobná regrese – proměnná Závislá: Y, nezávislá X – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Details – Durbin-Watsonova statistika:

Durbin-Watsonov o d (ocel.sta) a sériové korelace reziduí		
	Durbin- Watson.d	Sériové korelace
Odhad	2,179935	-0,108405

Hodnota této statistiky je blízká číslu 2, svědčí o tom, že rezidua nejsou korelovaná.

Posouzení homoskedasticity reziduí

Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Je vidět, že rezidua jsou kolem 0 rozmístěna náhodně. Model s regresní přímkou je tedy vhodný.

Testování nulovosti střední hodnoty reziduí:

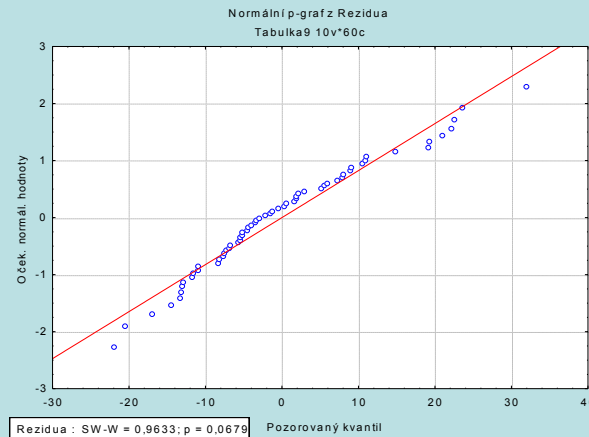
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka9)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000001	11,66757	60	1,506277	0,00	-0,000000	59	1,000000

Na hladině významnosti 0,05 nezamítáme hypotézu, že střední hodnota reziduí je 0.

Posouzení normality reziduí:

Na záložce Pravděpodobnostní grafy zvolíme Normální pravděpodobnostní graf reziduí:



Rezidua se řadí kolem ideální přímky, lze tedy soudit, že se řídí normálním rozložením. S-W test normality poskytuje p-hodnotu 0,0679, hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Závěr: Rezidua splňují všechny předpoklady, model regresní přímky je vhodný.