Jednoduchá lineární regrese

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. Teoretická analýza může upozornit například na to, že
s růstem hodnot veličiny X budou mít hodnoty veličiny Y tendenci monotónně růst či klesat,
tato tendence má charakter zrychlujícího se či zpomalujícího se růstu či poklesu,
jde o závislost, kdy s růstem hodnot veličiny X dochází zpočátku k růstu hodnot veličiny Y, který je po dosažení určitého maxima vystřídán poklesem,
apod.
Můžeme např. zkoumat závislost ceny ojetého auta (veličina Y) na jeho stáří (veličina X). Je zřejmé, že s rostoucím stářím bude klesat cena, ale není jasné, zda lineárně, kvadraticky či dokonce exponenciálně.
Vždy se snažíme o to aby regresní model byl jednoduchý, tj. aby neobsahoval příliš mnoho parametrů. Připadá-li v úvahu více funkcí, posuzujeme jejich vhodnost pomocí různých kritérií – viz dále.
Často však nemáme dostatek informací k provedení teoretického rozboru. Pak se snažíme odhadnout typ funkce pomocí dvourozměrného tečkového diagramu.
Zde se omezíme na funkce, které závisejí lineárně na parametrech $\beta_0, \beta_1, \ldots, \beta_p$.

ad b) Odhady $b_0, b_1, \ldots, b_p$ neznámých parametrů $\beta_0, \beta_1, \ldots, \beta_p$ získáme na základě dvourozměrného datového souboru $\begin{pmatrix} x_1 & y_1 \\ \ldots & \ldots \\ x_n & y_n \end{pmatrix}$ metodou nejmenších čtverců, tj. z podmínky, aby součet čtverců odchylek zjištěných a odhadnutých hodnot byl minimální.

Specifikace klasického modelu lineární regrese

$Y = m(x;\beta_0,\beta_1,\ldots,\beta_p) + \varepsilon$, kde

$m(x;\beta_0,\beta_1,\ldots,\beta_p)$ - teoretická regresní funkce, která lineárně závisí na neznámých regresních parametrech $\beta_0,\beta_1,\ldots,\beta_p$ a známých funkcích $f_1(x),\ldots,f_p(x)$, které již neobsahují neznámé parametry, tj. $m(x;\beta_0,\beta_1,\ldots,\beta_p) = \sum_{j=0}^{p} \beta_j f_j(x)$, přičemž $f_0(x) \equiv 1$.

Jde o deterministickou složku modelu.
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 $(x_1,y_1),\ldots,(x_n,y_n)$, tj. dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \ldots & \ldots \\ x_n & y_n \end{pmatrix}$.

Pro i = 1, ..., n platí: $y_i = m(x_i;\beta_0,\beta_1,\ldots,\beta_p) + \varepsilon_i$.

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

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

V tomto případě hovoříme o klasickém modelu lineární regrese.

Označení

$b_0, b_1, \ldots, b_p$ - odhady regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$ (nejčastěji je získáme metodou nejmenších čtverců, tj. z podmínky, že výraz

$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j f_j(x_i) \right)^2$ nabývá svého minima pro $\beta_j = b_j$, $j = 0, 1, \ldots, p$)

$\hat{m}(x; b_0, \ldots, b_p)$ - empirická regresní funkce

$\hat{y}_i = \hat{m}(x_i; b_0, \ldots, b_p) = \sum_{j=0}^{p} b_j f_j(x_i)$ - 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

$s^2 = \frac{S_E}{n - p - 1}$ - 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 ($S_T = S_R + S_E$)

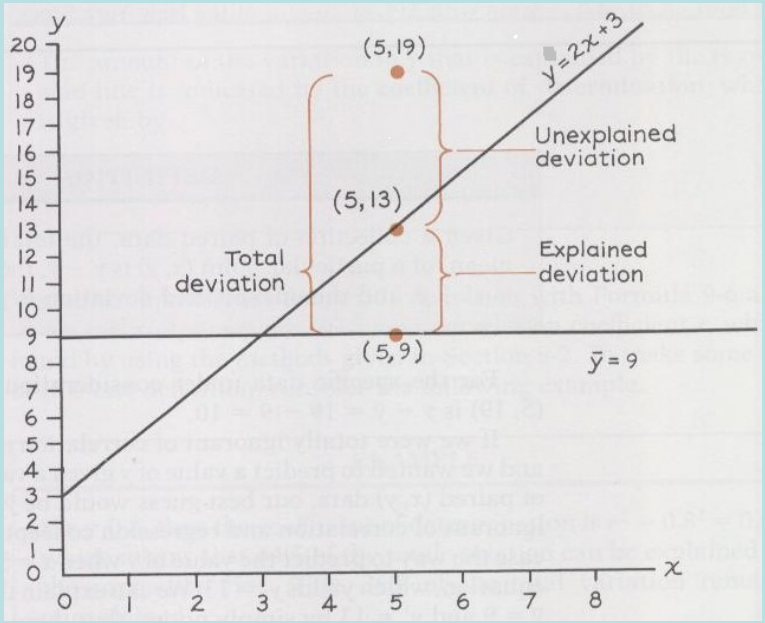
Význam jednotlivých typů součtů čtverců

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 $y_i - m_2$.

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 $\hat{y}_i - m_2$.



Maticový zápis klasického modelu lineární regrese

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kde

$\mathbf{y} = (y_1, \ldots, y_n)'$ - vektor pozorování závisle proměnné veličiny Y,

$$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix}$$ - regresní matice

(předpokládáme, že $h(\mathbf{X}) = p+1 < n$)

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ - vektor náhodných odchylek.

Podmínky (a) až (d) lze zkráceně zapsat ve tvaru $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X'X\beta} = \mathbf{X'y}$ - systém normálních rovnic

$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ – odhad vektoru $\boldsymbol{\beta}$ získaný metodou nejmenších čtverců

$\hat{\mathbf{y}} = \mathbf{Xb}$ – vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ - vektor reziduí

Vlastnosti odhadu **b**:

- odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(\mathbf{X'X})^{-1}\mathbf{X}'$;
- odhad **b** je nestranný, neboť $E(\mathbf{b}) = \boldsymbol{\beta}$;
- odhad **b** má varianční matici $\text{var}\,\mathbf{b} = \sigma^2(\mathbf{X'X})^{-1}$;
- odhad $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X'X})^{-1})$ vzhledem k platnosti podmínky (d);
- pro odhad **b** platí Gaussova - Markovova věta: Odhad $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ je nejlepší nestranný lineární odhad vektoru $\boldsymbol{\beta}$.

Příklad

Sestrojte regresní matici **X** pro lineární regresní model

a) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, provedeme-li 4 měření,

b) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 \ln x_{i2} + \varepsilon_i$, provedeme-li 5 měření.

Řešení:

ad a) $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{pmatrix}$, ad b) $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{11}^2 & \ln x_{12} \\ 1 & x_{21} & x_{21}^2 & \ln x_{22} \\ 1 & x_{31} & x_{31}^2 & \ln x_{32} \\ 1 & x_{41} & x_{41}^2 & \ln x_{42} \\ 1 & x_{51} & x_{51}^2 & \ln x_{52} \end{pmatrix}$

Intervaly spolehlivosti pro regresní parametry

$s_{b_j} = s\sqrt{v_{jj}}$ - směrodatná chyba odhadu b_j, kde v_{jj} je j-tý diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$.

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

$b_j \pm t_{1-\alpha/2}(n-p-1)\, s_{b_j}$.

(S intervaly spolehlivosti souvisí relativní chyby odhadů regresních parametrů. Získají se tak, že se vypočítá absolutní hodnota podílu poloviční šířky intervalu spolehlivosti a hodnoty odhadu. 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 $y = \beta_0 + \beta_1 x + \varepsilon$.

a) MNČ najděte odhady neznámých regresních parametrů β_0, β_1.
b) Sestrojte 95% intervaly spolehlivosti pro regresní parametry β_0, β_1.
c) Najděte relativní chyby odhadů regresních parametrů β_0, β_1.

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_technicka.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 mezí 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_technicka.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%.

Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme

H_0: $(\beta_1,\ldots,\beta_p) = (0,\ldots,0)$ proti H_1: $(\beta_1,\ldots,\beta_p) \neq (0,\ldots,0)$.

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

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

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

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

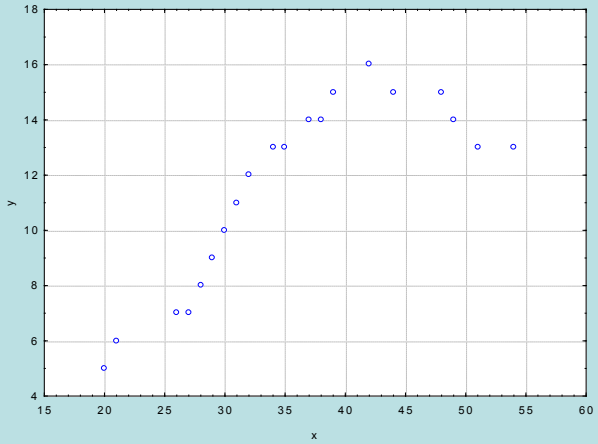
Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

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

Příklad:
Majitelé prodejny počítačových her nechali své prodavače absolvovat kurz prodejních dovedností. Poté zjišťovali po dobu 20 dnů, kolik osob navštíví během otevírací doby prodejnu (proměnná X) a jaká je v tento den tržba (proměnná Y, udává se v tisících Kč a je zaokrouhlená).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	20	21	2	27	28	29	30	31	32	34	35	37	38	39	42	44	48	49	51	54
y_i	5	6	7	7	8	9	10	11	12	13	13	14	14	15	16	15	15	14	13	13

Dvourozměrný tečkový diagram



Z grafu závislosti Y na X vyplývá, že s rostoucím počtem zákazníků se tržby zvyšují, avšak při denním počtu zákazníků asi 42 dosahují svého maxima a pak už zase klesají (vyšší počet zákazníků obsluha prodejny nezvládá a zákazníci odcházejí, aniž by nakoupili). Zdá se tedy, že vhodným modelem závislosti tržeb na počtu zákazníků bude regresní parabola $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.

Odhadněte parametry regresního modelu a proveďte celkový F-test.

Řešení:

Vytvoříme nový datový soubor se třemi proměnnými X, Xkv, Y a o 20 případech. Do proměnných X a Y napíšeme zjištěné hodnoty a do Dlouhého jména proměnné Xkv napíšeme = X^2.

Získání odhadů b_0, b_1, b_2:

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

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

Regresní parabola má tedy tvar: $y = -20{,}7723 + 1{,}5651x - 0{,}0173x^2$.

Výsledky celkového F-testu jsou uvedeny v záhlaví výstupní tabulky. Testová statistika F nabývá hodnoty 88,524, odpovídající p-hodnota je blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že dostačující je model konstanty.

Podrobnější výsledky získáme v tabulce analýzy rozptylu:

Aktivujeme Výsledky–vícenásobná regrese – Detailní výsledky – ANOVA

Analýza rozptylu (prodejna_software.sta)

Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	199,8141	2	99,90706	88,52445	0,000000
Rezid.	19,1859	17	1,12858		
Celk.	219,0000				

Testování významnosti regresních parametrů (dílčí t-testy)

Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu
$H_0: \beta_j = 0$ proti $H_1: \beta_j \neq 0$.

Testová statistika: $T_j = \frac{b_j}{s_{b_j}}$ má rozložení t(n-p-1), pokud H_0 platí.

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

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

Příklad:

V předešlém příkladě, kde byla modelována závislost tržby na počtu zákazníků regresní parabolou, proveďte dílčí t-testy o nevýznamnosti jednotlivých regresních parametrů

Řešení:

Stačí interpretovat výstupní tabulku vícenásobné regrese:

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

Sloupec označený t(17) obsahuje realizace testových statistik a sloupec p-hodn. pak odpovídající p-hodnoty. Ve všech třech případech jsou p-hodnoty menší než 0,05, tedy na hladině významnosti 0,05 zamítáme hypotézy o nevýznamnosti regresních parametrů β_0, β_1, β_2.

Kritéria pro posouzení vhodnosti zvolené regresní funkce

a) Index determinace

$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$ - index determinace $(0 \le ID^2 \le 1)$

- udává, jakou část variability závisle proměnné veličiny Y lze vysvětlit zvolenou regresní funkcí (často se udává v %);
- je zároveň mírou těsnosti závislosti proměnné Y na proměnné X;
- je to obecná míra, nezávislá na typu regresní funkce (lze použít i pro měření nelineární závislosti);
- je to míra, která nebere v úvahu počet parametrů regresní funkce. U regresních funkcí s více parametry vychází tedy obvykle vyšší než u regresních funkcí s méně parametry;
- tato míra není symetrická.

Za vhodnější se považuje ta regresní funkce, pro niž je index determinace vyšší. V případě, že porovnáváme několik modelů s rozdílným počtem parametrů, používáme adjustovaný index determinace:

$ID_{adj}^2 = ID^2 - \frac{(1 - ID^2)p}{n - p - 1}$ - adjustovaný index determinace

V příkladu s prodejem software najdeme index determinace ve výstupní tabulce regrese:

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

Index determinace je zde označen jako R2, nabývá hodnoty 0,9124 a říká nám, že 91,24% variability tržeb je vysvětleno regresní parabolou. Adjustovaný index determinace je označen Upravené R2.

b) Testové kritérium F

Za vhodnější je považována ta regresní funkce, u níž je hodnota testové statistiky $F = \frac{S_R / p}{S_E / (n - p - 1)}$ pro test významnosti modelu jako celku vyšší.

Ve výstupní tabulce regrese je testová statistika F uvedena v záhlaví:

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

V našem příkladě je označena F(2,17) a nabývá hodnoty 88,524.

c) Reziduální součet čtverců a reziduální rozptyl

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

Za vhodnější považujeme funkci, která má reziduální součet čtverců nižší. Reziduální součet čtverců lze použít pouze tehdy, když srovnáváme funkce se stejným počtem parametrů.

Reziduální rozptyl: $s^2 = \frac{S_E}{n - p}$

Za vhodnější považujeme tu funkci, která má reziduální rozptyl nižší. Reziduální rozptyl můžeme použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.
Obě charakteristiky najdeme v tabulce ANOVA:

Efekt	Analýza rozptylu (prodejna_software.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	199,8141	2	99,90706	88,52445	0,000000
Rezid.	19,1859	17	1,12858		
Celk.	219,0000				

Reziduální součet čtverců je 19,1859 a reziduální rozptyl je 1,12858.

d) Střední absolutní procentuální chyba predikce (MAPE)

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

Za vhodnější považujeme tu funkci, která má MAPE nižší.

Systém STATISTICA MAPE neposkytuje, tuto chybu musíme vypočítat.

Statistiky – Vícerozměrná regrese – Závisle proměnná y, nezávisle proměnné x, xkv - OK – OK – zvolíme Rezidua/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*abs((v1-v2)/v1)
Pomocí Statistiky – Základní statistiky/tabulky – Popisné statistiky zjistíme průměr proměnné chyba. V našem případě je MAPE 9,31%.

e) 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 tržby na počtu zákazníků.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

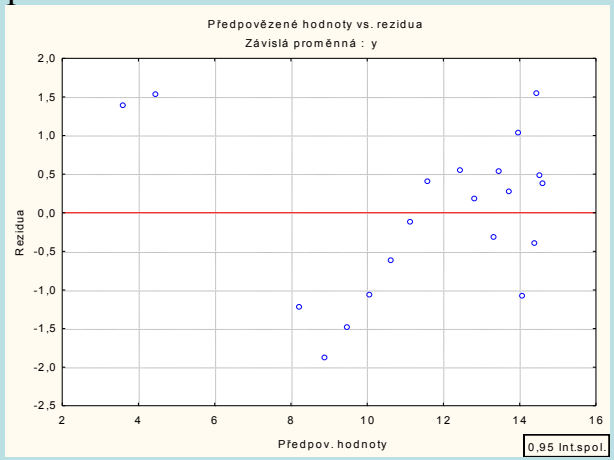
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x, xkv – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watson.d	Sériové korelace
Odhad	0,702506	0,599248

Hodnota této statistiky je nízká, svědčí o tom, že rezidua jsou kladně korelovaná.

Posouzení homoskedasticity reziduí

Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Je vidět, že rezidua nejsou kolem 0 rozmístěna náhodně. Model s regresní parabolou tedy není úplně vhodný.

Testování nulovosti střední hodnoty reziduí:

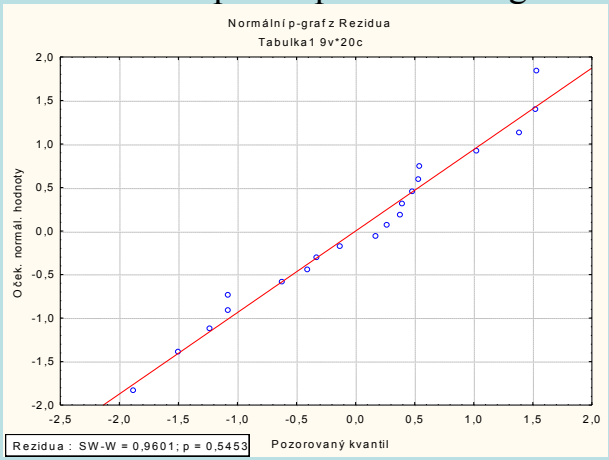
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000000	1,004880	20	0,224698	0,00	-0,000000	19	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.

Závěr: V neprospěch regresní paraboly hovoří hodnota Durbinovy – Watsonovy statistiky a graf závislosti reziduí na predikovaných hodnotách.

Model regresní přímky

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

$y = \beta_0 + \beta_1 x$ - teoretická regresní přímka (deterministicka 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.

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.

Systém normálních rovnic pro regresní přímku

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

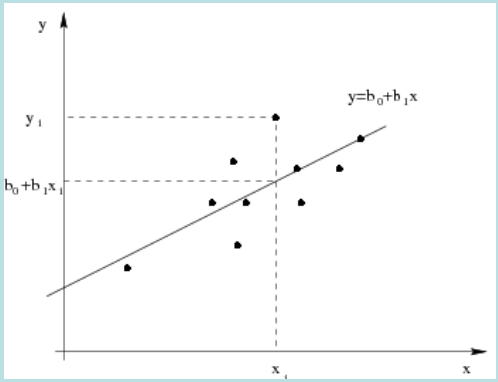
Systém normálních rovnic pro odhad regresních parametrů β_0 a β_1 získáme derivováním výrazu

$q(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ 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 $b_0 = \frac{\sum_{i=} {}_i^2 \sum_{i=} {}_i - \sum_{i=} {}_i \sum_{i=} {}_i y_i}{n\sum_{i=} {}_i^2 - \left(\sum_{i=} {}_i\right)^2}$, $b_1 = \frac{n\sum_{i=} {}_i y_i - \sum_{i=} {}_i \sum_{i=} {}_i}{n\sum_{i=} {}_i^2 - \left(\sum_{i=} {}_i\right)^2}$

Po jednoduchých úpravách dospějeme ke tvaru $b_1 = \frac{s_{12}}{s_1^2}$, kde s_{12} je kovariance hodnot (x_i, y_i), $i = 1, ..., n$ a s_1^2 je rozptyl hodnot $x_1, \ldots, x_n$. 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:

$$S_R = \sum_{i=} (\quad - n_2)^2 = \sum_{i=} \left[n_2 + \frac{s_{12}}{s_1^2} (x_i - n_1 - n_2) \right]^2 = \frac{s_{12}^2}{s_1^4} \sum_{i=} (\quad_i - n_2)^2 = n \frac{s_{12}^2}{s_1^2}.$$

Celkový součet čtverců $S_T = \sum_{i=} (\quad_i - n_2)^2 = n s_2^2$, tedy index determinace

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

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

$$a_1 = \frac{s_{12}}{s_2^2}, \quad a_0 = m_1 - a_1 m_2 = m_1 - \frac{s_{12}}{s_2^2} m_2 .$$

Empirická regresní přímka závislosti X na Y má tedy rovnici:

$$x = m_1 + \frac{s_{12}}{s_2^2}(y - m_2).$$

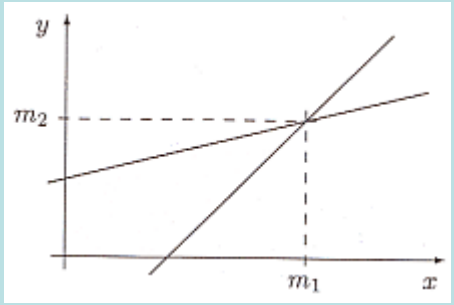
Obě empirické regresní přímky $y = b_0 + b_1 x$, $x = a_0 + a_1 y$ se nazývají sdružené regresní přímky a odhady regresních parametrů b_1, a_1 se nazývají odhady párově sdružených regresních parametrů.

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

$$y = m_2 + \frac{s_{12}}{s_1^2}(x - m_1), \quad y = m_1 + \frac{1}{r_{12}} \frac{s_2}{s_1}(x - m_2).$$

Vlastnosti sdružených regresních přímek

a) Sdružené regresní přímky se protínají v bodě o souřadnicích $[m_1, m_2]$ (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 $y = m_2$, $x = m_1$ (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 splynou 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;

$\cos \varphi = -1 \Leftrightarrow$ <=> 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

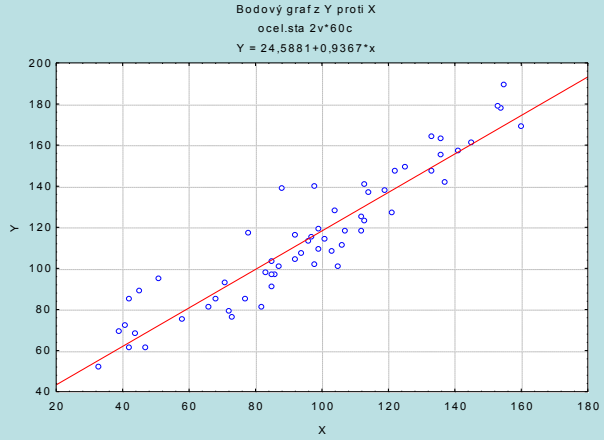
a) Určete regresní přímku meze pevnosti na mez plasticity.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
c) Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
d) Vypočtěte index determinace a interpretujte ho.
e) Najděte reziduální součet čtverců a odhad rozptylu náhodných odchylek.
f) Určete regresní přímku meze plasticity na mez pevnosti.
g) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
h) 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.

N=60	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					
	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

Ad b) Zakreslení regresních 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-váha	Hodnota	b-váha * 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:

N=60	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					
	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: Vrátíme 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				

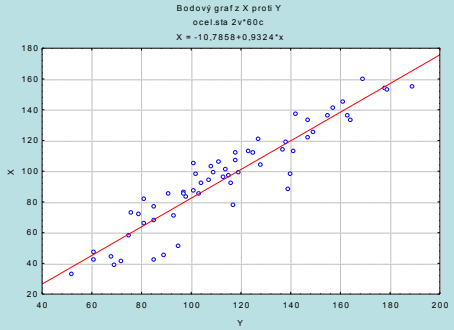
Vidíme, že reziduální součet čtverců je 8031,8 a reziduální rozptyl nabývá hodnoty 138,48.

Ad f) Výsledky pro 2. regresní přímku:

N=60	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					
	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.

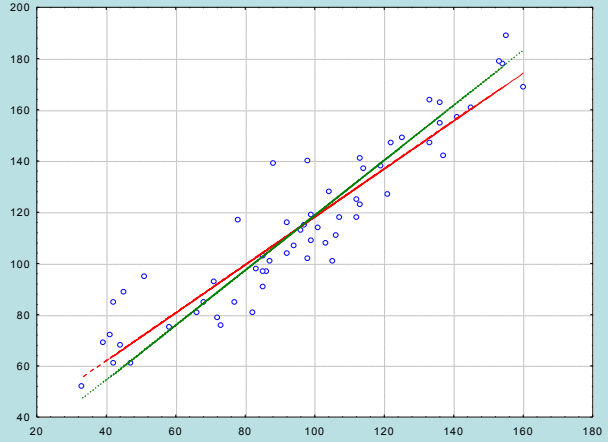
Ad g) Dvourozměrný tečkový diagram se zakreslenou 2. regresní přímkou



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.



Kritické hodnoty Durbinova-Watsonova testu pro autokorelaci 1. řádu pro $\alpha = 0{,}05$, rozsah výběru n a počet regresorů p (bez konstant)

	p=1		p=2		p=3		p=4		p=5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78