Jednoduchá lineární regrese I

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. Ten může upozornit například na to, že s růstem hodnot veličiny X budou mít hodnoty veličiny Y tendenci monotónně růst či klesat,
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.
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.
Není-li dostatek informací k provedení teoretického rozboru, snažíme odhadnout typ funkce pomocí tečkových diagramů.
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

- lineárně závisí na neznámých regresních parametrech $\beta_0,\beta_1,\ldots,\beta_p$,

- lineárně závisí na 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, …, 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 $(S_T = S_R + S_E)$

Maticový zápis klasického modelu lineární regrese

$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, 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(**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 $\mathbf{b} = (\mathbf{X'X})^{-1}\,\mathbf{X'}\,\mathbf{y}$:

- odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$;

- odhad **b** je nestranný, neboť $E(\mathbf{b}) = \boldsymbol{\beta}$;

- odhad **b** má varianční matici $\text{var}\,\mathbf{b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$;

- odhad $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1})$ vzhledem k platnosti podmínky (d), tj $\varepsilon_i \sim N(0, \sigma^2)$;

- pro odhad **b** platí Gaussova - Markovova věta: Odhad $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}$ 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 $T_j = \dfrac{b_j - \beta_j}{s_{b_j}} \sim t(n-p-1)$, 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, tj $\frac{1}{2}\left|\dfrac{h_j - d_j}{b_j}\right| 100\%$. Relativní chyba odhadu by neměla přesáhnout 10 %.)

Příklad:

V tabulce jsou výnosy technické cukrovky v tunách na ha od roku 2000 do roku 2011.

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
9	2008	57,26
10	2009	57,91
11	2010	54,36
12	2011	66,84

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 12 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= ,88823463 R2= ,78896076 Upravené R2= ,76785684 F(1,10)=37,385 p<,00011 Směrod. chyba odhadu : 2,9958					
N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			-3019,37	502,4173	-6,00968	0,000130
rok	0,888235	0,145272	1,53	0,2505	6,11429	0,000114

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;10)

Do Dlouhého jméne proměnné HM napíšeme

=v3+v4*VStudent(0,975;10)

Do Dlouhého jména proměnné chyba napíšeme

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

	Výsledky regrese se závislou proměnnou : Y (cukrovka_technicka.sta) R= ,88823463 R2= ,78896076 Upravené R2= ,76785684 F(1,10)=37,385 p<,00011 Směrod. chyba odhadu : 2,9958								
N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.	DM =v3-v4*	HM =v3+v4*	chyba =100*a
Abs.člen			-3019,37	502,4173	-6,00968	0,000130	-4138,82	-1899,91	37,07583
rok	0,888235	0,145272	1,53	0,2505	6,11429	0,000114	0,973556	2,08994	36,44149

S pravděpodobností 95% se bude úsek β_0 regresní přímky nacházet v intervalu (-4138,82; -1899,91). Odhad b_0 úseku β_0 je zatížen relativní chybou 37,1 %.

S pravděpodobností 95% se bude směrnice β_1 regresní přímky nacházet v intervalu (0,9736; 2,0899). Odhad b_1 úseku β_1 je zatížen relativní chybou 36,4 %.

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.

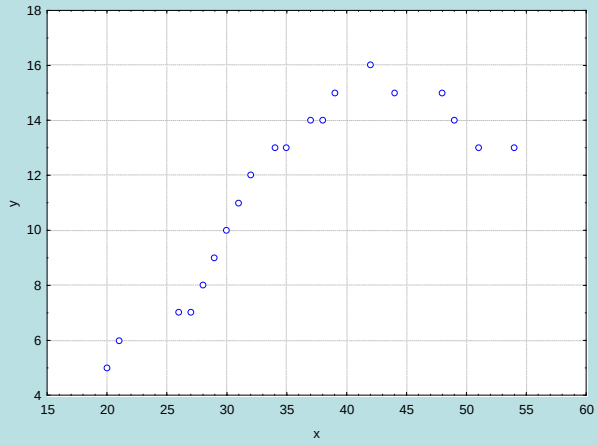
Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

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	26	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.

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
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						
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

Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Analýza rozptylu (prodejna_software.sta)					
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: $W = \left(-\infty, -t_{1-\alpha/2}(n-p-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-p-1), \infty\right)$.

$T_j \in W \Rightarrow H_0$ 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.

Interval spolehlivosti pro teoretickou regresní funkci

V uvažovaném lineárním modelu $Y = \sum_{j=0}^{p} \beta_j f_j(x_0) + \varepsilon$ můžeme na základě n dvojic pozorování (x_i, y_i), i = 1, …, n získat jak bodové, tak intervalové odhady neznámých regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$. Lze však spočítat též meze 100(1-α) % intervalu spolehlivosti pro teoretickou regresní funkci při zadané hodnotě x_0.

Vytvoříme vektor $\mathbf{x}_0 = (1, f_1(x_0), \ldots, f_p(x_0))'$ a zabýváme se lineární kombinací $\mathbf{x}_0'\beta$ složek vektoru regresních parametrů, tj. hodnotou $m(x_0; \beta_0, \beta_1, \ldots, \beta_p) = \sum_{j=0}^{p} \beta_j f_j(x_0)$ teoretické regresní funkce v bodě x_0.

100(1- α)% interval spolehlivosti pro $\mathbf{x}_0'\beta$, tj. pro hodnotu regresní funkce $m(x_0; \beta_0, \beta_1, \ldots, \beta_p)$ má meze

$$\mathbf{x}_0'\mathbf{b} \pm t_{1-\alpha/2}(n-p-1)s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\,.$$

Při spojité změně argumentu x_0 mezní hodnoty tohoto 100(1-α)% empirického intervalu spolehlivosti pro teoretickou regresní funkci vytvoří 100(1-α)% pás spolehlivosti kolem regresní funkce.

Tento pás spolehlivosti však nelze interpretovat tak, že pokrývá celou regresní funkci s pravděpodobností 1-α, pouze ukazuje na šířku intervalu spolehlivosti pro vypočtenou hodnotu z modelu při pevně zvolené hodnotě argumentu x_0.)

Příklad: U automobilu Škoda 120 byla změřena spotřeba benzínu (v l/100 km) v závislosti na rychlosti (v km/h).

rychlost X	40	50	60	70	80	90	100	110
spotřeba Y	5,7	5,4	5,2	5,2	5,8	6,0	7,5	8,1

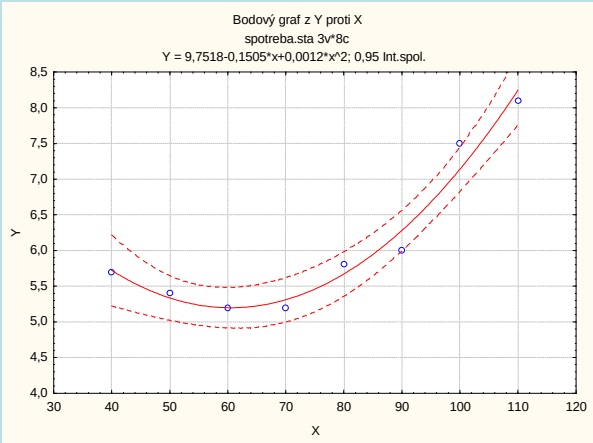
Vhodným modelem je regresní parabola $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Odhadněte její parametry a najděte 95% pás spolehlivosti kolem regresní funkce.

Řešení:

N=8	Výsledky regrese se závislou proměnnou : Y (spotreba.sta) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(5)	p-hodn.
Abs.člen			9,751786	0,945689	10,31183	0,000148
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905

Spotřeba = 9,751786 – 0,150536*rychlost + 0,001244*rychlost^2

Získání 95% pásu spolehlivosti kolem regresní funkce: Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zvolíme Proložení Polynomiální (implicitně je nastaveno na polynom 2. stupně, lze měnit na záložce Možnosti 2) – zapneme Regresní pásy Spolehl. – OK.



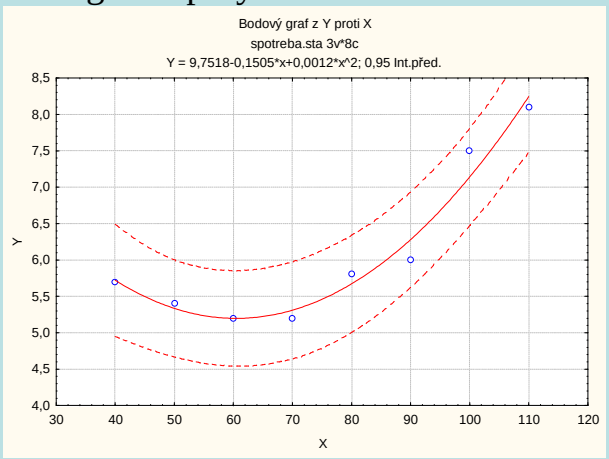
Predikční interval spolehlivosti

V případě, kdy chceme zkonstruovat 100(1- α)% interval spolehlivosti nikoli pro hodnotu regresní funkce, ale pro i-tou predikovanou hodnotu $\hat{y}_i$ (tzv. predikční interval), dostaneme meze

$$\mathbf{x}_0'\mathbf{b} \pm t_{1-\alpha/2}(n-p-1)s\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\ .$$

Vidíme, že tento predikční interval je širší než předešlý interval spolehlivosti.
Je to interval, který nás informuje o tom, v jakém rozsahu můžeme očekávat jedno další pozorování s pravděpodobností aspoň 1- α.
Při spojitě se měnícím $\mathbf{x}_0$ vytvoří meze tohoto predikčního intervalu spolehlivosti tzv. predikční pás spolehlivosti kolem regresní funkce.

Příklad: Pro regresní parabolu z předešlého příkladu sestrojte 95% predikční pás spolehlivosti kolem regresní funkce.
Řešení: Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zvolíme Proložení Polynomiální (implicitně je nastaveno na polynom 2. stupně) – zapneme Regresní pásy Predikce – OK.



Chceme-li mít v jednom obrázku zakresleny oba typy pásů, postupujeme takto: ve vytvořeném grafu 2x klikneme na pozadí – vybereme Regresní pásy – Přidat nový pár pásů – OK.

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 $(0 \leq ID^2 \leq 1)$

- 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:

N=20	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					
	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: $s^2 = \frac{S_E}{n - p - 1}$

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:

	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				

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

d) Střední absolutní procentuální chyba predikce (MAPE)

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

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 Kolmogorova – Smirnovova testu nebo Shapirovým – Wilkovým 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:

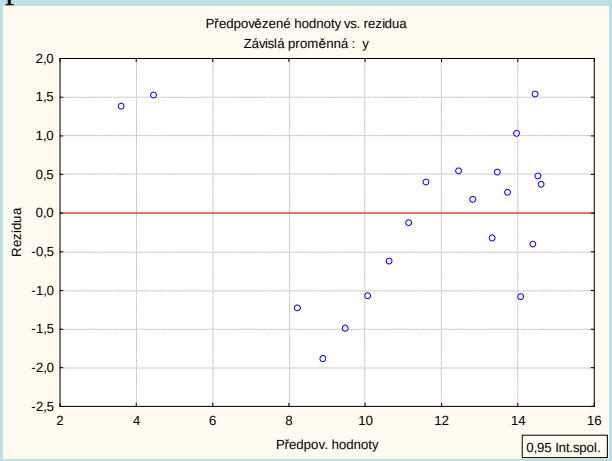
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x, xkv – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	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:

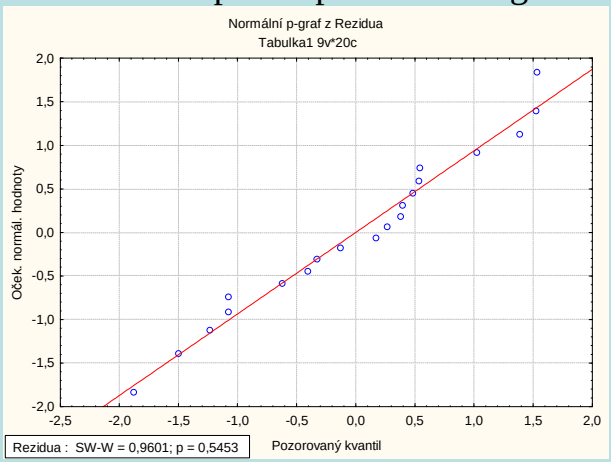
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

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.