

Osnova přednášky Jednoduchá lineární regrese

1. Motivace

2. Specifikace klasického modelu lineární regrese

- 2.1. Základní pojmy
- 2.2. Označení
- 2.3. Význam jednotlivých typů součtů čtverců
- 2.4. Maticový zápis klasického modelu lineární regrese
- 2.5. Odhad regresních parametrů metodou nejmenších čtverců

3. Statistická inference v klasickém modelu lineární regrese

- 3.1. Intervaly spolehlivosti pro regresní parametry
- 3.2. Celkový F-test
- 3.3. Dílčí t-testy
- 3.4. Interval spolehlivosti pro teoretickou regresní funkci
- 3.5. Predikční interval spolehlivosti

4. Kritéria pro posouzení vhodnosti zvolené regresní funkce

- 4.1. Index determinace
- 4.2. Testové kritérium celkového F-testu
- 4.3. Reziduální součet čtverců a reziduální rozptyl
- 4.4. Střední absolutní procentuální chyba predikce

5. Analýza reziduí

- 5.1. Požadavky kladené na rezidua
- 5.2. Ověřování požadavků kladených na rezidua

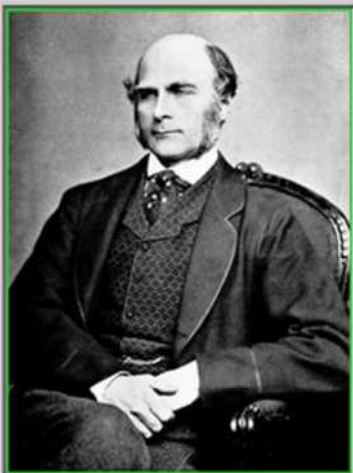
1. Motivace

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?

Základy regresní analýzy položil kolem roku 1880 Francis Galton, anglický vědec, činný ve velmi mnoha různých oborech: psychologii a antropologii, statistice, geografii a dalších. Byl to bratranec Charlese Darwina.



Sir Francis Galton F.R.S. 1822-1911

- 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. Příklad-li v úvahu více funkcí, posuzujeme jejich vhodnost pomocí různých kritérií – viz dále.

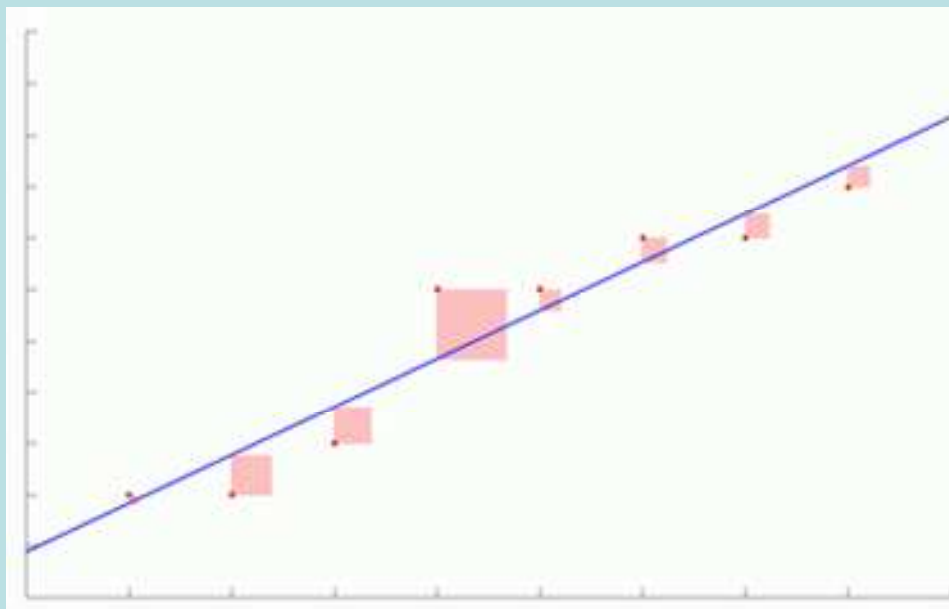
Není-li dostatek informací k provedení teoretického rozboru, snažíme se odhadnout typ funkce pomocí tečkových diagramů.

Zde se omezíme na funkce, které závisejí lineárně na parametrech $\beta_0, \beta_1, \dots, \beta_p$.

ad b) Odhady b_0, b_1, \dots, b_p neznámých parametrů $\beta_0, \beta_1, \dots, \beta_p$ 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.

Ilustrace metody nejmenších čtverců pro model přímky



2. Specifikace klasického modelu lineární regrese

2.1. Základní pojmy

Model má tvar $Y = m(x; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon$, kde

$m(x; \beta_0, \beta_1, \dots, \beta_p)$ - **teoretická regresní funkce**

- lineárně závisí na neznámých regresních parametrech $\beta_0, \beta_1, \dots, \beta_p$,
- lineárně závisí na známých funkcích $f_1(x), \dots, f_p(x)$, které již neobsahují neznámé parametry,

tj. $m(x; \beta_0, \beta_1, \dots, \beta_p) = \sum_{j=0}^p \beta_j f_j(x)$, 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 $(x_1, y_1), \dots, (x_n, y_n)$, tj. dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$.

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

O náhodných odchylkách $\varepsilon_1, \dots, \varepsilon_n$ 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**.

2.2. Označení

b_0, b_1, \dots, b_p - odhady regresních parametrů $\beta_0, \beta_1, \dots, \beta_p$ (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 $\beta_j = b_j, j = 0, 1, \dots, p$)

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

$\hat{y}_i = \hat{m}(x_i; b_0, \dots, b_p) = \sum_{j=0}^p b_j f_j(x_i)$ - 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, $f_E = n - p - 1$... reziduální počet stupňů volnosti

$s^2 = \frac{S_E}{n - p - 1}$ - odhad rozptylu σ^2 (reziduální rozptyl, průměrný reziduální čtverec)

$S_R = \sum_{i=1}^n (\hat{y}_i - m_2)^2$ - regresní součet čtverců ($m_2 = \frac{1}{n} \sum_{i=1}^n y_i$), $f_R = p$... regresní počet stupňů volnosti

$\frac{S_R}{p}$... průměrný regresní čtverec

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

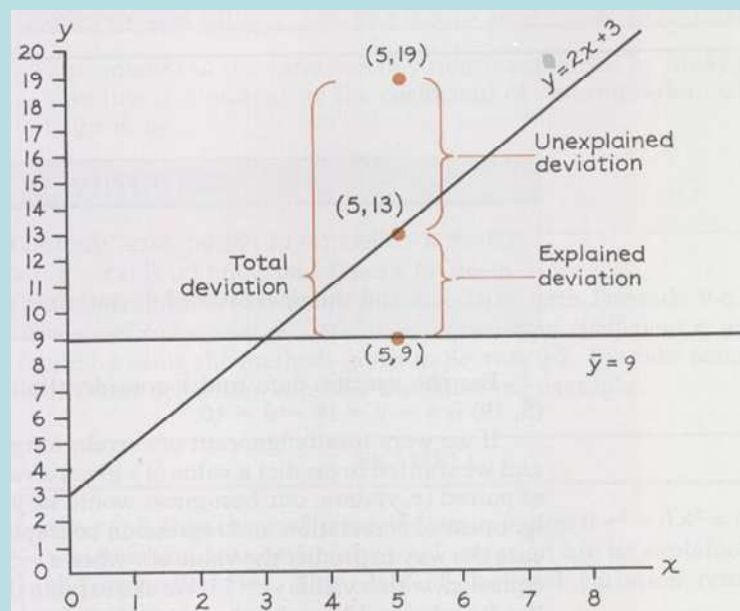
2.3 Význam jednotlivých typů součtů čtverců

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 $\hat{y}_i - m_2$.



2.4. Maticový zápis klasického modelu lineární regrese

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & f_1(x_1) & \dots & f_p(x_1) \\ \dots & \dots & \dots & \dots \\ 1 & f_1(x_n) & \dots & f_p(x_n) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \dots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}, \text{ tj } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ kde}$$

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

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

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

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

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

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

Příklad

Sestrojte regresní matici \mathbf{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:

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

2.5. Odhad regresních parametrů metodou nejmenších čtverců

Maticově zapsaná metoda nejmenších čtverců $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \rightarrow \min$ vede na rovnice

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

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

$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ - vektor regresních odhadů (vektor predikce)

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

Vlastnosti odhadu $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$:

- odhad \mathbf{b} je lineární, je to lineární kombinace pozorování y_1, \dots, y_n s maticí vah $(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$;
- odhad \mathbf{b} je nestranný, neboť $E(\mathbf{b}) = \boldsymbol{\beta}$;
- odhad \mathbf{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 $\varepsilon_i \sim N(0, \sigma^2)$;
- $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ je nejlepší nestranný lineární odhad vektoru $\boldsymbol{\beta}$.

3. Statistická inference v klasickém modelu lineární regrese

3.1. 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 $j = 0, 1, \dots, p$ statistika $T_j = \frac{b_j - \beta_j}{s_{b_j}} \sim t(n - p - 1)$, tedy $100(1 - \alpha)\%$ interval spolehlivosti

pro β_j má meze: $b_j \pm t_{1-\alpha/2}(n - p - 1)s_{b_j}$.

Pokud interval spolehlivosti obsahu nulu (tj. dolní mez je záporná a horní mez je kladná), pak regresní parametr je s rizikem omylu nejvýše $100\alpha\%$ nevýznamný.

Příklad:

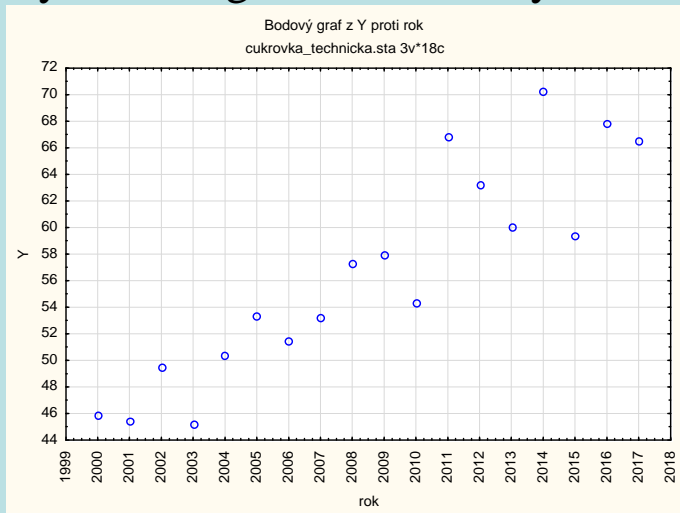
V tabulce jsou výnosy technické cukrovky v tunách na ha od roku 2000 do roku 2017.

i	rok	výnos
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
13	2012	63,26
14	2013	60,00
15	2014	70,28
16	2015	59,38
17	2016	67,81
18	2017	66,56

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

- MNČ najděte odhady neznámých regresních parametrů β_0 , β_1 .
- Sestrojte 95% intervaly spolehlivosti pro regresní parametry β_0 , β_1 .

Vytvoříme graf závislosti výnosu na roku slizně:



Vzhled grafu svědčí o tom, že přímka by mohla být vhodným regresním modelem. Kvůli jednodušší interpretaci výsledků bude roli nezávisle proměnné hrát nikoliv rok samotný, ale jeho pořadí, které označíme X . Proměnná X tedy nabývá hodnot 1, 2, ..., 18.

Získáme odhady b_0 , b_1 společně s 95% intervaly spolehlivosti pro β_0 , β_1 :

	b	d	h
N=18			
Abs.člen	-24,7891	-33,0121	-16,5661
Y	0,6063	0,4623	0,7504

Rovnice regresní přímky: $y = -24,7891 + 0,6063x$.

S pravděpodobností 95 % se bude úsek β_0 nacházet v intervalu (-33,0121; -16,5661).

S pravděpodobností 95 % se bude směrnice β_1 nacházet v intervalu (0,4623; 0,7504).

3.2. Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme

$$H_0: (\beta_1, \dots, \beta_p)' = (0, \dots, 0)' \text{ proti } H_1: (\beta_1, \dots, \beta_p)' \neq (0, \dots, 0)'.$$

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

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

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

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

Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

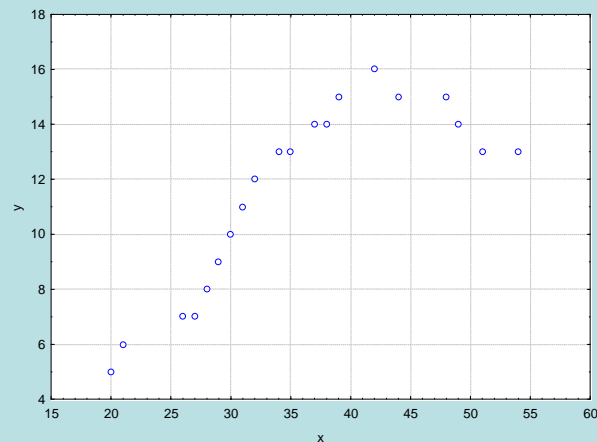
zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R	p	S_R/p	$\frac{S_R/p}{S_E/(n-p-1)}$
reziduální	S_E	$n-p-1$	$S_E/(n-p-1)$	-
celkový	S_T	$n-1$	-	-

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:

Tabulka s odhady b_0 , b_1 , b_2 a mezemi 95% intervalů spolehlivosti pro β_0 , β_1 , β_2 :

	b	d	h
N=20			
Abs.člen	-20,7723	-27,88920	-13,655307
x	1,5651	1,16517	1,965036
x ²	-0,0173	-0,02264	-0,011940

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

Výsledky celkového F-testu:

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				

Regresní součet čtverců je 199,8141, reziduální 19,1859, 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.

Odhad rozptylu náhodných odchylek (reziduální rozptyl, průměrný reziduální čtverec) $s^2 = 1,12858$.

3.3. Testování významnosti regresních parametrů (dílní t-testy)

Na hladině významnosti α pro $j = 0, 1, \dots, p$ 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 .

3.3. Testování významnosti regresních parametrů (dílní t-testy)

Na hladině významnosti α pro $j = 0, 1, \dots, p$ 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:

Tabulka odhadu regresních parametrů s výsledky dílčích t-testů:

	b	t(17)	p-hodn.
N=20			
Abs.člen	-20,7723	-6,15792	0,000011
x	1,5651	8,25655	0,000000
x ²	-0,0173	-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 .

Interpretace výsledků F-testu a t-testů

Mohou nastat tyto situace:

- a) F-test je významný, všechny t-testy jsou významné \Rightarrow vhodný model
- b) F-test je nevýznamný, všechny t-testy jsou nevýznamné \Rightarrow nevhodný model
- c) F-test je významný, ale některé t-testy jsou nevýznamné \Rightarrow vypustíme odpovídající vysvětlující proměnné
- d) F-test je významný, ale všechny t-testy jsou nevýznamné \Rightarrow model formálně vyhovuje jako celek, ale žádná vysvětlující proměnná není významná. Je to důsledek silného lineárního vztahu mezi proměnnými (tzv. multikolinearita). Je nutné upravit nebo zcela změnit model.

3.4. Interval spolehlivosti pro teoretickou regresní funkci

V uvažovaném lineárním modelu $Y = m(x_0; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon$ neboli $Y = \sum_{j=0}^p \beta_j f_j(x_0) + \varepsilon$ můžeme na zá-

kladě n dvojic pozorování (x_i, y_i) , $i = 1, \dots, n$ získat jak bodové, tak intervalové odhady neznámých regresních parametrů $\beta_0, \beta_1, \dots, \beta_p$.

Lze však spočítat též meze $100(1-\alpha)$ % intervalu spolehlivosti pro teoretickou regresní funkci při zadané

hodnotě x_0 , tj. pro hodnotou $m(x_0; \beta_0, \beta_1, \dots, \beta_p) = \sum_{j=0}^p \beta_j f_j(x_0)$.

$100(1-\alpha)$ % interval spolehlivosti pro $m(x_0; \beta_0, \beta_1, \dots, \beta_p)$ 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 spojitě změně argumentu x_0 mezní hodnoty tohoto $100(1-\alpha)$ % empirického intervalu spolehlivosti pro teoretickou regresní funkci vytvoří **100(1- α)% pás spolehlivosti kolem regresní funkce**.

Pás spolehlivosti je nejužší v bodě $m_1 = \frac{1}{n} \sum_{i=1}^n x_i$, směrem od tohoto bodu se na obě strany rozšiřuje.

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

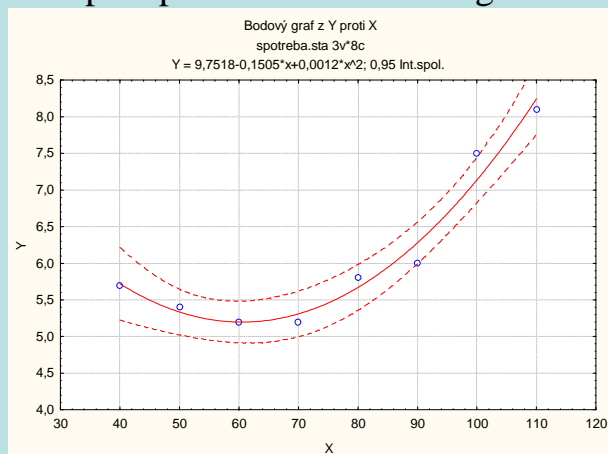
Vhodným modelem je regresní parabola $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Odhadněte její parametry a najděte 95% pás spolehlivosti kolem regresní funkce.

Řešení:

N=8	b	d	h	t(5)	p-hodn.
Abs.člen	9,751786	7,32081487	12,1827566	10,31183	0,000148
x	-0,150536	-0,2194809	-0,0815906	-5,61264	0,002483
x ²	0,001244	0,00078845	0,00169965	7,01912	0,000905

$$\text{Spotřeba} = 9,751786 - 0,150536 \cdot \text{rychlost} + 0,001244 \cdot \text{rychlost}^2$$

95% pás spolehlivosti kolem regresní funkce:



3.5. Predikční interval spolehlivosti

V případě, kdy chceme zkonstruovat $100(1 - \alpha)\%$ interval spolehlivosti nikoli pro hodnotu regresní funkce, ale pro jednu novou pozorovanou hodnotu závisle proměnné veličiny Y (tzv. predikční interval), dostaneme meze

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

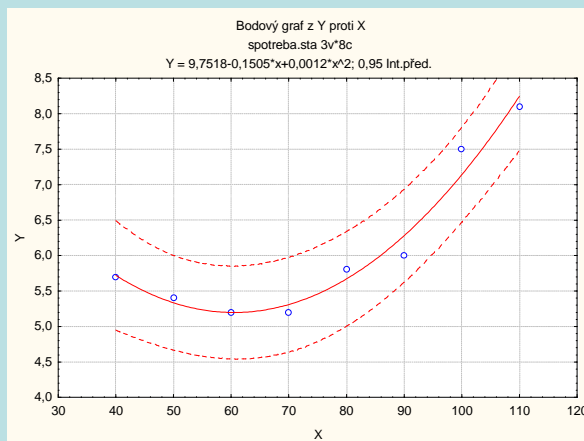
Vidíme, že tento predikční interval je širší než předešlý interval spolehlivosti.

Je to interval, který nás informuje o tom, v jakém rozsahu můžeme očekávat jedno další pozorování s pravděpodobností aspoň $1 - \alpha$.

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:



4. Kritéria pro posouzení vhodnosti zvolené regresní funkce

4.1. Index determinace

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

- 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:

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

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.

4.2. Testové kritérium celkového F-testu

Za vhodnější je považována ta regresní funkce, u níž je hodnota testové statistiky

$$F = \frac{S_R/p}{S_E/(n-p-1)} \text{ pro test významnosti modelu jako celku vyšší.}$$

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.

4.3. 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.

4.4. Střední absolutní procentuální chyba predikce (MAPE)

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

Za vhodnější považujeme tu funkci, která má MAPE nižší.

V příkladě s prodejem software je MAPE 9,31%.

5. Analýza reziduí

5.1. Požadavky kladené na rezidua

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).

5.2. Ověřování požadavků kladených na rezidua

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 – 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.

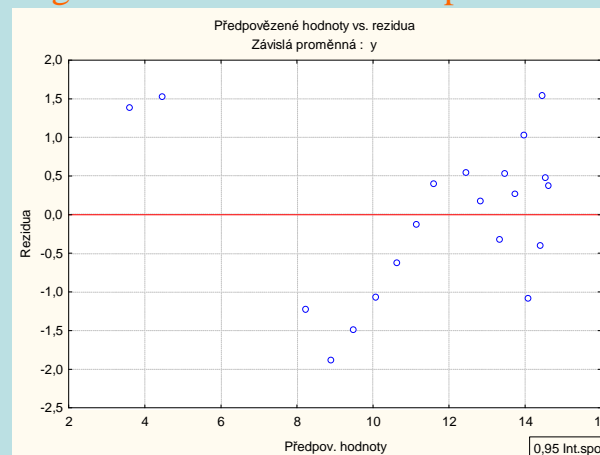
Příklad: Proveďte analýzu reziduí pro příklad s modelováním závislosti tržby na počtu zákazníků.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

	Durbin-Watson.d
Odhad	0,702506

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

Posouzení homoskedasticity reziduí pomocí grafu závislosti reziduí na předikovaných hodnotách



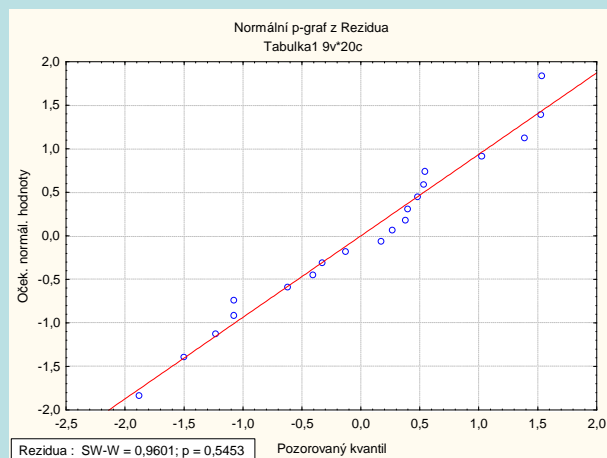
Je vidět, že rezidua nejsou kolem 0 rozmístěna náhodně. Model s regresní parabolou tedy není úplně vhodný.

Testování nulovosti střední hodnoty reziduí:

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 pomocí N-P plotu:



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.