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. 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 $\beta_0, \beta_1, \ldots, \beta_p$.
Zvláštní pozornost budeme věnovat polynomiální funkci 1. stupně $y = \beta_0 + \beta_1 x$.

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

$m(x, \beta_1, \ldots, \beta_p)$ - teoretická regresní funkce, která lineárně závisí na neznámých regresních parametrech $\beta_1, \ldots, \beta_p$ 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: $Y_i = m(x_i, \beta_1, \ldots, \beta_p) + \varepsilon_i$.

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

a) $E(\varepsilon_i) = 0$ (odchylky nejsou systematické)

b) $D(\varepsilon_i) = \sigma^2$ (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) $\varepsilon \sim N(0, \sigma^2)$.

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

$\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} \right)^2$ 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 ($S_T = S_E + S_R$)

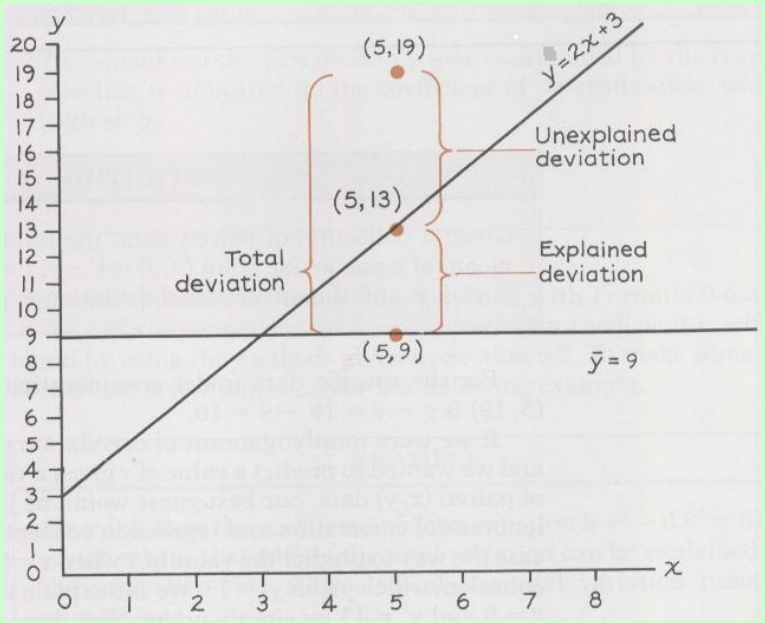
Význam jednotlivých typů součtů čtverců

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 $(y_i - \bar{y})^2$.

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 $(y_i - \hat{y}_i)^2$.

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 $(\hat{y}_i - \bar{y})^2$.



Maticový zápis klasického modelu lineární regrese

$\mathbf{y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 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 $h(\mathbf{X}) = p+1 < n$)

$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ - 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 $(\mathbf{XX})^{-}\mathbf{X}$;
- 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) $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, provedeme-li 4 měření,

b) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, 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

$s_{b_j} = \sqrt{v_{jj}}$ - 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 $T_j = \frac{b_j - \beta_j}{s_{b_j}} \sim t(n-p)$, tedy 100(1- α)% interval spolehlivosti pro β_j má meze:

$b_j \pm t(n-p) \cdot s_{b_j}$.

(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 :
R= ,84604287 R2= ,71578853 Upravené R2
F(1,6)=15,111 p<,00810 Směrod. chyba odh

N=8	b*	Sm.chy z b*	b	Sm.chy z b	t(6)	p-hod
Abs.čl			-2312	607,4	-3,80	0,008
rok	0,846	0,217	1,1	0,30	3,88	0,008

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_tec
R= ,84604287 R2= ,71578853 Upravené R2= ,66841995
F(1,6)=15,111 p<,00810 Směrod. chyba odhadu : 1,9651

N=8	b*	Sm.chy z b*	b	Sm.chy z b	t(6)	p-hod	DM =v3-v	HM =v3+v	chyb =100*
Abs.čl			-2312	607,4	-3,80	0,008	-3798	-825,	64,28
rok	0,846	0,217	1,1	0,30	3,88	0,008	0,436	1,920	62,94

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

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

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

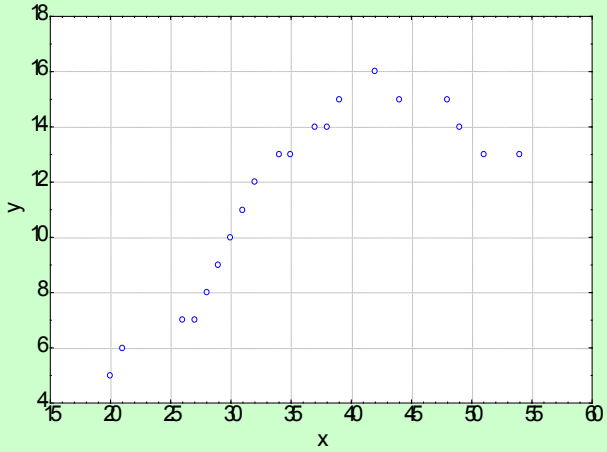
Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

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	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.

N=20	Výsledky regrese se závislou proměnnou : R= ,95519276 R2= ,91239322 Upravené F F(2,17)=88,524 p<,00000 Směrod. chyba c					
	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-20,7	3,373	-6,15	0,000
x	4,526	0,548	1,56	0,189	8,256	0,000
xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

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

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

Podrobnější výsledky získáme v tabulce analýzy rozptylu:

Aktivujeme Výsledky–vícenásobná regrese – Detailní výsledky – ANOVA

Efekt	Analýza rozptylu (prodejna_ Souč čtverc	sv	Prům čtverc	F	p-hod
Regre	199,8	2	99,90	88,52	0,000
Rezid	19,18	17	1,128		
Celk.	219,0				

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

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

$t_j \in W$ 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 :
R= ,95519276 R2= ,91239322 Upravené F
F(2,17)=88,524 p<,00000 Směrod. chyba o

N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-20,7	3,373	-6,15	0,000
x	4,52(	0,548	1,56	0,189	8,25(	0,000
xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

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

$ID^2 = \frac{S}{S} = 1 - \frac{S}{S}$ - 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:

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

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

Výsledky regrese se závislou proměnnou :
R= ,95519276 R2= ,91239322 Upravené R
F(2,17)=88,524 p<,00000 Směrod. chyba o

N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-20,7	3,373	-6,15	0,000
x	4,52(	0,548	1,56	0,189	8,25(	0,000
xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

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 : R= ,95519276 R2= ,91239322 Upravené R F(2,17)=88,524 p<,00000 Směrod. chyba o					
	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-20,7	3,373	-6,15	0,000
x	4,526	0,548	1,56	0,189	8,256	0,000
xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

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: $s^2 = \frac{S_E}{\quad}$

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_				
Efekt	Souč čtverc	sv	Prům čtverc	F	p-hod
Regre	199,8	2	99,90	88,52	0,000
Rezid	19,18	1	1,128		
Celk.	219,0				

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 rok, nezávisle proměnné Y - OK – OK – zvolíme Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua & předpovědi – vybereme proměnnou y - OK. K vzniklému datovému souboru přidáme jedni 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:

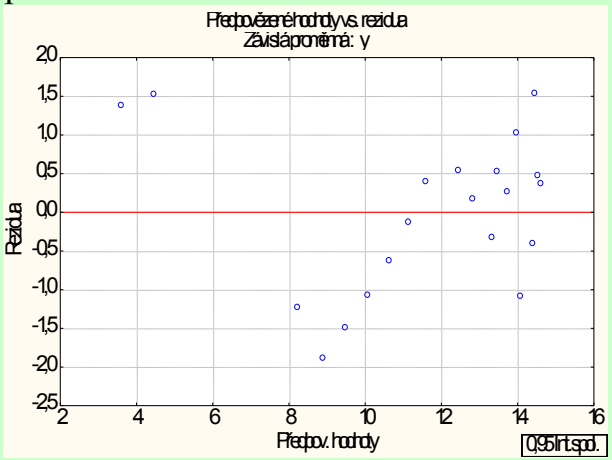
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x, xkv – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbi Watso	Sério korela
Odha	0,702	0,599

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:

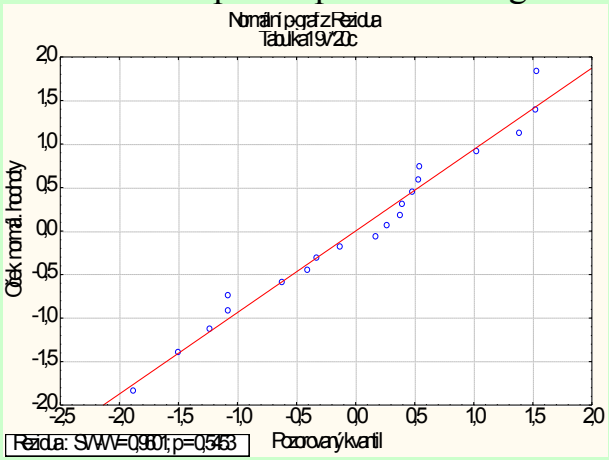
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměn	Prům	Sm.od	N	Sm.chy	Referen konsta	t	SV	p
Rezidu	-0,000	1,004	2(	0,224	0,(	-0,000	1(	1,000

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.

Problém autokorelovaných reziduí a jeho odstranění

Předpokládejme, že náhodná odchylka ε je lineárně závislá na předešlé náhodné odchylce ε, tj. jde o autokorelaci 1. řádu (v praxi nejčastější případ): $\varepsilon = \_ +$, i = 2, …, n (u_i je náhodná odchylka od modelu lineární závislosti a ρ je koeficient korelace dvou sousedních náhodných odchylek ε, ε).

Předpoklad o existenci autokorelace 1. řádu můžeme ověřit pomocí Durbinova – Watsonova testu, který je založen na Durbinově – Watsonově statistice: $D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_{i-1}^2} \rightarrow$, jejíž hodnoty se nacházejí v intervalu $\langle 0, 4 \rangle$. Pro nezamítnutí nulové hypotézy se hodnoty testového kritéria musejí pohybovat kolem hodnoty 2.

Testujeme vlastně hypotézu $H_0: \rho$ proti alternativní hypotéze $H_1: \rho$ resp. $H_1: \rho$ s tím, že zamítnutí H_0 se interpretuje jako tvrzení o existenci pozitivní resp. negativní autokorelace 1. řádu (na dané hladině významnosti α). Pro dané α, daný rozsah n náhodného výběru a daný počet p regresních koeficientů jsou tabelovány kritické hodnoty d_L a d_U.

Testujeme-li existenci pozitivní autokorelace, pak při $D > d_U$ se nezamítá H_0 a při $D < d_L$ se přijímá H_1.

Je-li $d_L \leq D \leq d_U$, pak nelze přijmout žádné rozhodnutí (říkáme, že test mlčí).

Testujeme-li existenci negativní autokorelace, pak při $D < 4 - d_U$ se nezamítá H_0 a při $D > 4 - d_L$ se přijímá H_1.

Je-li $4 - d_U \leq D \leq 4 - d_L$, pak nelze přijmout žádné rozhodnutí.

Prokážeme-li na dané hladině významnosti α existenci autokorelace 1. řádu, měli bychom ji eliminovat.

Nejprve odhadneme koeficient korelace ρ: $\hat{\rho} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_{i-1}^2}$.

Pak už můžeme vypočítat odhady náhodných odchylek (tj. rezidua) v autokorelaci): $\hat{u}_i = \_$, i = 2, …, n.

Získané odhady $\hat{u}_i$ přičteme k predikovaným hodnotám $\hat{y}_i$ získaným z regresního modelu a znovu provedeme regresní analýzu, kde roli závisle proměnné veličiny bude hrát součet $\hat{y}_i +$.

Postup v systému STATISTICA

(Použijeme data z příkladu o závislosti tržeb na počtu zákazníků.)

Rezidua z modelu $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ jsou uložena v proměnné Rezidua. Pro tato rezidua je hodnota D-W statistiky D = 0,702506 a kritické hodnoty pro $\alpha = 0{,}0$ . . . . . $_L = 1{,}1$, $d_U = 1{,}54$. Protože $D < d_L$, zamítáme na hladině významnosti 0,05 hypotézu o nekorelovanosti reziduí ve prospěch alternativy o pozitivní autokorelaci 1. řádu.

Získání odhadů reziduí v autokorelaci: $\hat{u}_i = \_ \_$, i = 2, …, n:

Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Rezidua – ARIMA & autokorelační funkce – v Parametrech modelu ARIMA zvolíme p-Autoregresní 1 – OK (Zahájit odhady parametrů) – Souhrn: Odhady parametrů.

Vstup: REZIDUA (Tabulka39)
Transformace: žádná
Model:(1,0,0) PC Rezid. = ,64920

Param	Param	Asym SmC	Asym t(1	p	Doln 95% s	Horn 95% s
p(1)	0,599	0,189	3,161	0,005	0,202	0,995

Vidíme, že odhad koeficientu korelace dvou po sobě následujících reziduí je 0,6 a na hladině 0,05 je významný (p-hodnota 0,005134 < 0,05).

Uložíme rezidua z autokorelace: Přehled & rezidua – Přehled reziduí. Vzniklou proměnnou okopírujeme do původního datového souboru a k tomuto datovému souboru přidáme ještě proměnnou s predikovanými hodnotami z původního modelu. Do nové proměnné nazvané nove y uložíme součet reziduí a predikovaných hodnot. Pak znovu provedeme regresní analýzu:

N=20	Výsledky regrese se závislou proměnnou : nc R= ,96958525 R2= ,94009556 Upravené R2= F(2,17)=133,39 p<,00000 Směrod. chyba odh					
	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-20,1	2,696	-7,46	0,000
x	4,583	0,453	1,53	0,151	10,11	0,000
xkv	-3,78	0,453	-0,01	0,002	-8,34	0,000

Nová regresní parabola má tvar: $y = -20{,}1238 + 1{,}5323x - 0{,}0169x^2$.

Porovnáme výslednou tabulku regrese s původní tabulkou:

N=20	Výsledky regrese se závislou proměnnou : R= ,95519276 R2= ,91239322 Upravené F F(2,17)=88,524 p<,00000 Směrod. chyba o					
	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-20,7	3,373	-6,15	0,000
x	4,526	0,548	1,56	0,189	8,256	0,000
xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

Získali jsme vyšší hodnotu testové statistiky F (a tedy i vyšší adjustovaný index determinace) a menší směrodatné chyby odhadů regresních parametrů (tudíž také vyšší hodnoty testových statistik pro dílčí t-testy).

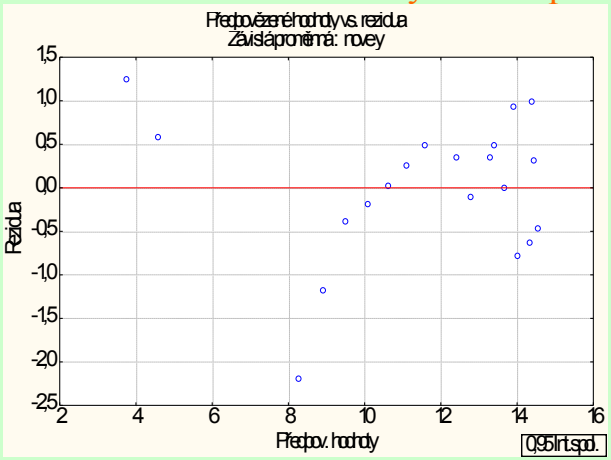
Nyní prozkoumáme chování reziduí v novém regresním modelu.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

	Durbi Watso	Sério korela
Odha	1,356	0,256

Hodnota D-W statistiky D = 1,35663 a kritické hodnoty pro $\alpha = 0,05$, n = 20, p = 2 jsou: $d_L = 1,1$, $d_U = 1,54$.
Protože $d_L \leq D \leq d_U$, nelze přijmout žádné rozhodnutí.

Posouzení homoskedasticity reziduí pomocí grafu závislosti reziduí na predikovaných hodnotách



Opět vidíme, že rezidua nejsou kolem 0 rozmístěna náhodně. Bylo by tedy vhodné celý postup zopakovat znovu.

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