Pokročilé metody v jednoduché lineární regresi

Test adekvátnosti regresního modelu

Hodnoty veličiny Y jsou roztříděny do $r \geq 3$ skupin podle variant $x_{[1]}, \ldots, x_{[r]}$ veličiny X.
Označme n_i počet pozorování v i-té skupině, $i = 1, \ldots, r$, přičemž aspoň jedna skupina má více než jedno pozorování.
Budeme předpokládat, že každá skupina hodnot má normální rozložení a že všechny skupiny mají týž rozptyl.
Všech pozorování je n.
Průměr hodnot v i-té skupině označme M_i a průměr všech hodnot označme M.
Charakter závislosti Y na X popíšeme regresní funkcí $m(x; \beta_0, \ldots, \beta_p)$.
Budeme testovat hypotézu, zda je tato regresní funkce vhodným modelem pro naše data. Při testování budeme potřebovat tyto součty čtverců:

celkový součet čtverců $S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - M)^2$,

skupinový součet čtverců $S_A = \sum_{i=1}^{r} n_i (M_i - M)^2$,

regresní součet čtverců $S_R = \sum_{i=1}^{r} n_i (\hat{y}_i - M)^2$.

Testová statistika: $F = \dfrac{(S_A - S_R)/(r-p-1)}{(S_T - S_A)/(n-r)}$ se řídí rozložením $F(r-p-1, n-r)$, jestliže H_0 platí.

Kritický obor: $W = \langle F_{1-\alpha}(r-p-1, n-r), \infty)$
$F \in W$ => na hladině významnosti α zamítáme hypotézu, že funkce $m(x; \beta_0, \ldots, \beta_p)$ je vhodným regresním modelem závislosti Y na X.

Těsnost závislosti Y na X vyjádřenou skupinovými průměry měří poměr determinace $P^2 = S_A/S_T$.
Nabývá hodnot z intervalu <0,1>. Čím je poměr determinace bližší jedné, tím je závislost silnější, čím je bližší nule, tím je závislost slabší.

Příklad: Máme k dispozici údaje o cenách 23 náhodně vybraných domů (veličina Y - v tisících $) a počtu jejich pokojů (veličina X) v jednom americkém městě.

počet pokojů	cena
5	155,168,180
6	166,172,179,190,200
7	210,215,218,225,230,245
8	213,225,240,247,249
9	267,275,290,298

Závislost ceny domu na počtu pokojů popište regresní přímkou.

Na hladině významnosti 0,05 testujte hypotézu, že přímka je vhodným regresním modelem pro tato data.

Těsnost závislosti vyjádřete poměrem determinace.

Znázorněte data s proloženou regresní přímkou.

Řešení: MNČ odhadneme parametry regresní přímky. Má tvar y = 17,2885 + 28,5851 x.

Vypočítáme regresní součet čtverců: $S_R = \sum_{i=1}^{r} n_i (\hat{y}_i - M)^2 = 30907{,}9041$,

celkový součet čtverců: $S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - M)^2 = 35870{,}6087$,

skupinový součet čtverců: $S_A = \sum_{i=1}^{r} n_i (M_i - M)^2 = 32474{,}1087$.

Dosadíme do vzorce pro testovou statistiku: $F = \frac{(S_A - S_R)/(r-p-1)}{(S_T - S_A)/(n-r)} = \frac{(32474{,}1087 - 30907{,}9041)/3}{(35870{,}6087 - 32474{,}1087)/18} = 2{,}76$

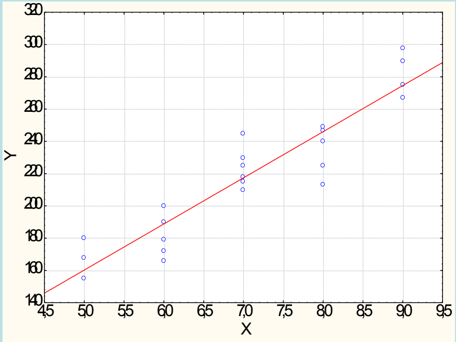
Stanovíme kritický obor W = $\langle F_{1-\alpha}(r-p-1, n-r), \infty) = \langle F_{0,95}(3, 18), \infty) = \langle 3{,}161, \infty)$.

Jelikož $F \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem.

Poměr determinace: $P^2 = S_A/S_T = 32474{,}1087/35870{,}6087 = 0{,}9053$,

tedy závislost ceny domu na počtu pokojů je v daném datovém souboru značně silná.

Znázorníme data s proloženou regresní přímkou.



Řešení v systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 23 případy:

	1 X	2 Y
1	5	15
2	5	16
3	5	18
4	6	16
5	6	17
6	6	17
7	6	19
8	6	20
9	7	21
10	7	21
11	7	21
12	7	22
13	7	23
14	7	24
15	8	21
16	8	22
17	8	24
18	8	24
19	8	24
20	9	26
21	9	27
22	9	29
23	9	29

Odhadneme parametry regresní přímky:

N=23	Výsledky regrese se závislou proměnnou R= ,92825096 R2= ,86164984 Upravené F(1,21)=130,79 p<,00000 Směrod. chyba					
	Beta	Sm.chy beta	B	Sm.chy B	t(21)	Úrove
Abs.čl			17,28	18,00	0,960	0,347
X	0,928	0,081	28,58	2,499	11,43	0,000

Cena = 17,28851 + 28,5806*počet pokojů

Sestavíme tabulku ANOVA:
Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (ceny_bytu Souč čtverc	sv	Průmě čtverc	F	Úroveň
Regre	30907	1	30907	130,7	0,000
Rezid	4962	2	236,		
Celk.	3587(				

Vidíme, že $S_R = 30907,9$, $S_T = 35870,61$

Provedeme jednofaktorovou analýzu rozptylu, abychom získali skupinový součet čtverců:
Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – Y, Grupovací - X – OK – OK – Analýza rozptylu.

Proměn	Analýza rozptylu (ceny_bytu.sta) Označ. efekty jsou význ. na hlad. p < ,05000 SČ efekt	SV efek	PČ efekt	SČ chyba	SV chyb	PČ chyba	F	p
Y	32474	4	8118,	3396,	1	188,6	43,02	0,000

Zde najdeme $S_A = 32474,11$.

Vypočteme testovou statistiku $F = \frac{S_A - \dots}{S_T - \dots} = \frac{32474,1 \dots}{35870,6 \dots} =$ [illegible] 76

a najdeme kritický obor W = <3,161, ∞). Jelikož F ∉ W, nezamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem.

Test adekvátnosti modelu pomocí Obecných regresních modelů

Zadáme data a použijeme cestu:

Statistiky – Pokročilé lineární/nelineární modely – Obecné regresní modely – Jednorozměrná regrese - OK – na záložce Možnosti zaškrtneme Kvalita proložení – OK – Závislá Y, Spoj. nezáv. prom. X – OK – Více výsledků – Celkové R – ve stromové struktuře vlevo vybereme Test kvality modelu.

Závislá Proměi	Test kvality modelu (ceny_bytu.sta) SC Rezid	sv Rezid	PC Rezid	SC Chyb	sv Chyl	PC Chyb	SC Kv prolož	SV Kv prolož	PC Kv prolož	ta F	p
Y	4962,	2	236,3	3396,	1	188,6	1566,	3	522,0	2,766	0,071

Čitatel testové statistiky F je roven 1566,205 a je uveden ve sloupci Kvalita proložení.

Jmenovatel testové statistiky F je roven 3396,5 a je uveden ve sloupci SČ Chyba.

Hodnota testové statistiky je 2,767 a odpovídající p-hodnota je 0,0717. Na hladině významnosti 0,05 tedy nemůžeme zamítnout hypotézu, že přímka je vhodným modelem k popisu závislosti ceny domu na počtu pokojů.

Problém autokorelovaných reziduí a jeho odstranění

Předpokládejme, že náhodná odchylka ε_i je lineárně závislá na předešlé náhodné odchylce ε_{i-1}, tj. jde o autokorelaci 1. řádu (v praxi nejčastější případ):

$\varepsilon_i = \rho\varepsilon_{i-1} + u_i$, i = 2, …, n ($u_i$ je náhodná odchylka od modelu lineární závislosti a ρ je koeficient korelace dvou sousedních náhodných odchylek ε_i, ε_{i-1}).

Předpoklad o existenci autokorelace 1. řádu můžeme ověřit pomocí Durbinova – Watsonova testu, který je založen na Durbinově – Watsonově statistice: $D = \dfrac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$.

Tato statistika nabývá hodnot z intervalu $\langle 0;4 \rangle$ a má střední hodnotu 2. Nízké hodnoty statistiky D znamenají, že sousední rezidua jsou spíše podobná, což svědčí ve prospech kladné autokorelace. Naopak, vysoké hodnoty statistiky D jsou způsobeny negativní autokorelací, avšak s tou se v praxi příliš často nesetkáváme.

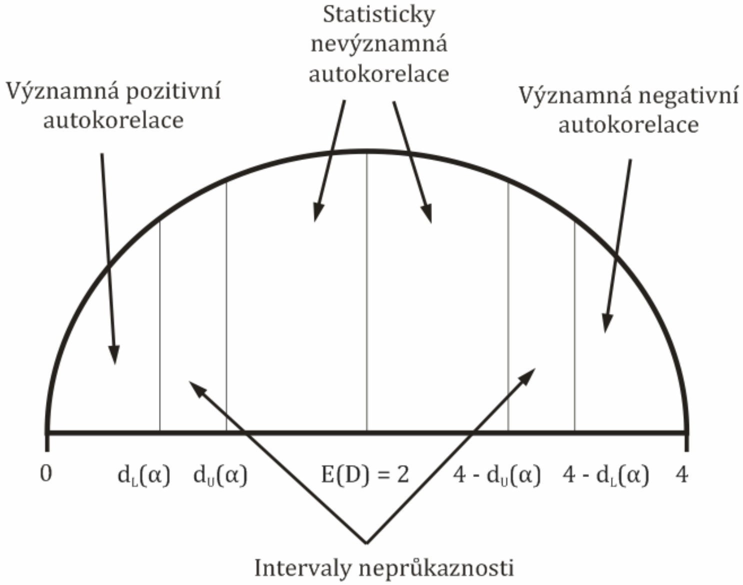
Pro dané α, daný počet pozorování n a daný počet p regresních parametrů – bez konstanty (v případě regresní přímky p = 1) jsou tabelovány kritické hodnoty $d_L(\alpha), d_U(\alpha)$.

Testujeme-li na hladině významnosti α existenci pozitivní autokorelace, pak při $D \in \langle d_U(\alpha); 2\rangle$ se nezamítá H_0 a při $D \in \langle 0; d_L(\alpha))$ se přijímá H_1.

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

Testujeme-li na hladině významnosti α existenci negativní autokorelace, pak při $D \in \langle 2; 4 - d_U(\alpha)\rangle$ se nezamítá H_0 a při $D \in (4 - d_L(\alpha); 4\rangle$ se přijímá H_1.

Je-li $4 - d_U(\alpha) < D \leq 4 - d_L(\alpha)$, pak nelze učinit žádné rozhodnutí.



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

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

Pak už můžeme vypočítat odhady náhodných odchylek (tj. rezidua) v autokorelaci: $\hat{u}_i = e_i - \hat{\rho} e_{i-1}$, 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 $\hat{y}_i + \hat{u}_i$.

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 α = 0,05, n = 20, p = 2 jsou: $d_L = 1{,}1$, $d_U = 1{,}54$. 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: $\hat{u}_i = e_i - \hat{\rho} e_{i-1}$, 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) PČ Rezid. = ,64920					
Param	Parar	Asym SmC	Asym t(19	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 autokoreace: Přehled & rezidua – Přehled reziduí. Vzniklou proměnnou okopírujeme do původního datového souboru a k tomuto datovému souboru přidáme ještě proměnnou s predikovanými hodnotami z parabolického 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 : no 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	4,58	0,453	1,53	0,151	10,11	0,000
xkv	-3,78	0,453	-0,0	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 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	4,52	0,548	1,56	0,189	8,25	0,000
xkv	-3,73	0,548	-0,0	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 pomocí Durbinovy – Watsonovy statistiky:

	Durbi Watso	Serio 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.

Pokud celý postup zopakujeme ještě jednou, dostaneme hodnotu D-W statistiky 1,6885. Nyní je D $> d_U$, tudíž nelze zamítnout hypotézu, že rezidua nejsou kladně korelovaná.
Parametry výsledného modelu jsou:

	Výsledky regrese se závislou proměnnou R= ,97136268 R2= ,94354546 Upravené F(2,17)=142,06 p<,00000 Směrod. chyba					
N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hod
Abs.čl			-19,7	2,605	-7,58	0,000
x	4,53	0,440	1,51	0,146	10,31	0,000
xkv	-3,73	0,440	-0,0	0,001	-8,48	0,000

Regresní parabola má tedy rovnici: $y = -19{,}7523 + 1{,}5103x - 0{,}0166x^2$.

Linearizující transformace

Odhad parametrů regresních funkcí, které nejsou lineární z hlediska parametrů, se neprovádí metodou nejmenších čtverců přímo, protože její použití vede k soustavě nelineárních rovnic. V některých speciálních případech však nelineární regresní funkci můžeme vhodnou transformací převést na lineární.

Např. máme exponenciální regresní funkci $y =$. Provedeme logaritmickou transformaci $\ln y = \ln \beta_0 + x \ln \beta_1$, čímž získáme regresní funkci lineární v parametrech. Parametry $\ln \beta_0$ a $\ln \beta_1$ odhadneme metodou nejmenších čtverců a odlogaritmováním získáme odhady původních regresních koeficientů β_0, β_1.

Přehled linearizujících transformací

Funkce	Linearizující transformace
$y =$	$\ln y = \ln \beta_0 + x \ln \beta_1$
$y =$	$\ln y = \ln \beta_0 + \beta_1 \ln x$
$y =$	$\ln y = \ln \beta_0 - \beta_1 \ln x$
$y = \frac{1}{\ }$	$\frac{1}{y} =$
$y = \frac{x}{\ }$	$\frac{x}{y} =$

Příklad: Hotelová společnost vlastnící 12 hotelů analyzuje vztah mezi celkovými měsíčními tržbami (veličina Y) a tržbami vyprodukovanými stravovacími úseky (veličina X).

č. h.	1	2	3	4	5	6	7	8	9	10	11	12
x	2,0	1,2	14,8	8,3	8,4	3,0	4,8	15,6	16,1	11,5	14,2	14,0
y	12,0	8,0	76,4	17,0	21,3	10,0	12,5	97,3	88,0	25,0	38,6	47,3

Popište tuto závislost exponenciální regresní funkcí $y =$. Najděte odhady parametrů β_0, β_1 a vypočtěte predikovanou hodnotu celkových měsíčních tržeb pro x = 10.

Řešení: Provedeme logaritmickou transformaci $\ln y = \ln \beta_0 + x \ln \beta_1$. Metodou nejmenších čtverců získáme odhady $\ln b_0 = 1{,}8559$, $\ln b_1 = 0{,}1504$.
Odlogaritmováním dostaneme $b_0 = 6{,}3973$, $b_1 = 1{,}1623$. Predikovaná hodnota y pro x = 10 je $6{,}3973 . 1{,}1623^{10} = 28{,}7859$.

Řešení v systému STATISTICA:
Vytvoříme datový soubor se dvěma proměnnými a 12 případy:

	1 Y	2 X
1	1	2
2	8	1,
3	76	14
4	1	8,
5	21	8,
6	1	3
7	12	4,
8	97	15
9	8	16
10	2	11
11	38	14
12	47	1

Přidáme novou proměnnou ln y. Do jejího Dlouhého jména napíšeme =log(y).
Pak provedeme regresní analýzu se závisle proměnnou ln y a nezávisle proměnnou X:

N=12	Výsledky regrese se závislou proměnnou R= ,95851605 R2= ,91875303 Upravené F(1,10)=113,08 p<,00000 Směrod. chyba					
	Beta	Sm.chy beta	B	Sm.chy B	t(10)	Úrover
Abs.čl			1,855	0,154	12,02	0,000
X	0,958	0,090	0,150	0,014	10,63	0,000

K výsledné tabulce přidáme novou proměnnou b, do jejíhož Dlouhého jména napíšeme =exp(B).

N=12	Výsledky regrese se závislou proměnnou : ln y R= ,95851605 R2= ,91875303 Upravené R2= , F(1,10)=113,08 p<,00000 Směrod. chyba odhad						
	Beta	Sm.chy beta	B	Sm.chy B	t(10)	Úrover	b =exp(
Abs.čl			1,855	0,154	12,02	0,000	6,397
X	0,958	0,090	0,150	0,014	10,63	0,000	1,162

Model má tedy tvar: $y = 6{,}397333 . 1{,}162332^{x}$.

Získání predikované hodnoty pro x = 10:
Vrátíme se do Výsledky – vícenásobná regrese – na záložce Rezidua/předpoklady/předpovědi vybereme Předpověď závisle proměnné – X = 10 – OK. K výsledné tabulce přidáme proměnnou predikce a do jejího Dlouhého jména napíšeme =exp(v3).

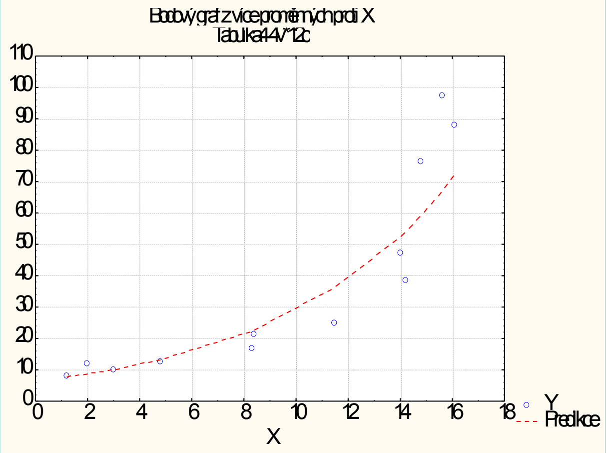
Proměn	Předpovězené hodnoty (h proměnné: ln y			
	b-vár	Hodn	b-vár * Hod	predik =exp(
X	0,150	10,00	1,504	4,500
Abs. čle			1,855	6,397
Předpo			3,360	28,79
-95,0%L			3,189	24,28
+95,0%			3,530	34,14

Vidíme, že predikovaná hodnota je 28,79.

Vytvoříme ještě dvourozměrný tečkový diagram s proloženou exponenciálou. Na záložce Rezidua/předpoklady/předpovědi vybereme reziduální analýza – Uložit – Uložit rezidua & předpovědi – vybereme X, Y – OK.
Ve vzniklé tabulce odstraníme proměnné č. 5 až 10 a proměnnou rezidua přejmenujeme na Predikce. Do Dlouhého jména této proměnné napíšeme =exp(v3).
Tento datový soubor uspořádáme podle velikosti hodnot proměnné X: Data - Setřídit – Proměnná X – OK.

	hotely.sta			
	1 Y	2 X	3 Předpov	4 Predik
1	8	1,	2,0	7,6
1	1	2	2,1	8,6
3	1	3	2,3	10,0
4	12	4,	2,5	13,
5	1	8,	3,1	22,
6	21	8,	3,1	22,
7	2	11	3,5	36,
8	47	1	3,9	52,
9	38	14	3,9	54,
1	76	14	4,0	59,
1	97	15	4,2	66,
1	8	16	4,2	72,

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



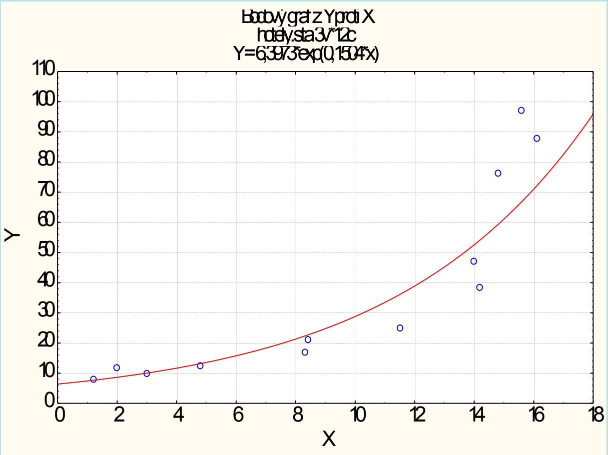
Provedení regresní analýzy pomocí modulu Jednoduchá nelineární regrese

Pro data z předešlého příkladu najdeme odhady parametrů modelu y= pomocí modulu Jednoduchá nelineární regrese.

Statistiky - Pokročilé lineární/nelineární odhady - Jednoduchá nelineární regrese – Proměnné X, Y – OK – OK – zaškrtneme LN(X) – OK – Proměnné – Závislé LN-V1, Nezávislé X – OK. Dostaneme stejnou tabulku jako předešlým postupem a výsledné hodnoty odhadů regresních parametrů získáme exponenciální transformací.

Získání odhadů parametrů modelu y= pomocí Bodových grafů

Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zaškrtneme Proložení Exponenciální – OK.



V záhlaví grafu je uvedena regresní rovnice y = 6,3973*exp(0,1504*x), tedy
$b_0 = 6{,}3973$, $b_1 = e^{0{,}1504} = 1{,}1623$.

Kritické hodnoty Durbinova-Watsonova testu pro autokorelaci 1. řádu pro α = 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