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 $m(x; \beta_0, \beta_1, \ldots, \beta_p)$.

Budeme testovat hypotézu, zda je tato regresní funkce vhodným modelem pro naše data.

Při testování budeme potřebovat tyto součty čtverců:

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

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

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

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 = <$F_{1-\alpha}$ (r-p-1, n-r), ∞)

F $\in$ W => na hladině významnosti α zamítáme hypotézu, že funkce $m(x;\beta_0,\beta_1,\ldots,\beta_p)$ 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: $S_R = \sum_{i=1}^{r} n_i (\hat{y}_i - M_i)^2 = 30907{,}9041$,

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

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

Testová statistika: $F = \dfrac{(S_A - S_R)/(r-p-1)}{(S_T - S_A)/(n-r)} = \dfrac{(32474{,}1087 - 30907{,}9041)/(5-2)}{(35870{,}6087 - 32474{,}1087)/(23-5)} = 2{,}768$

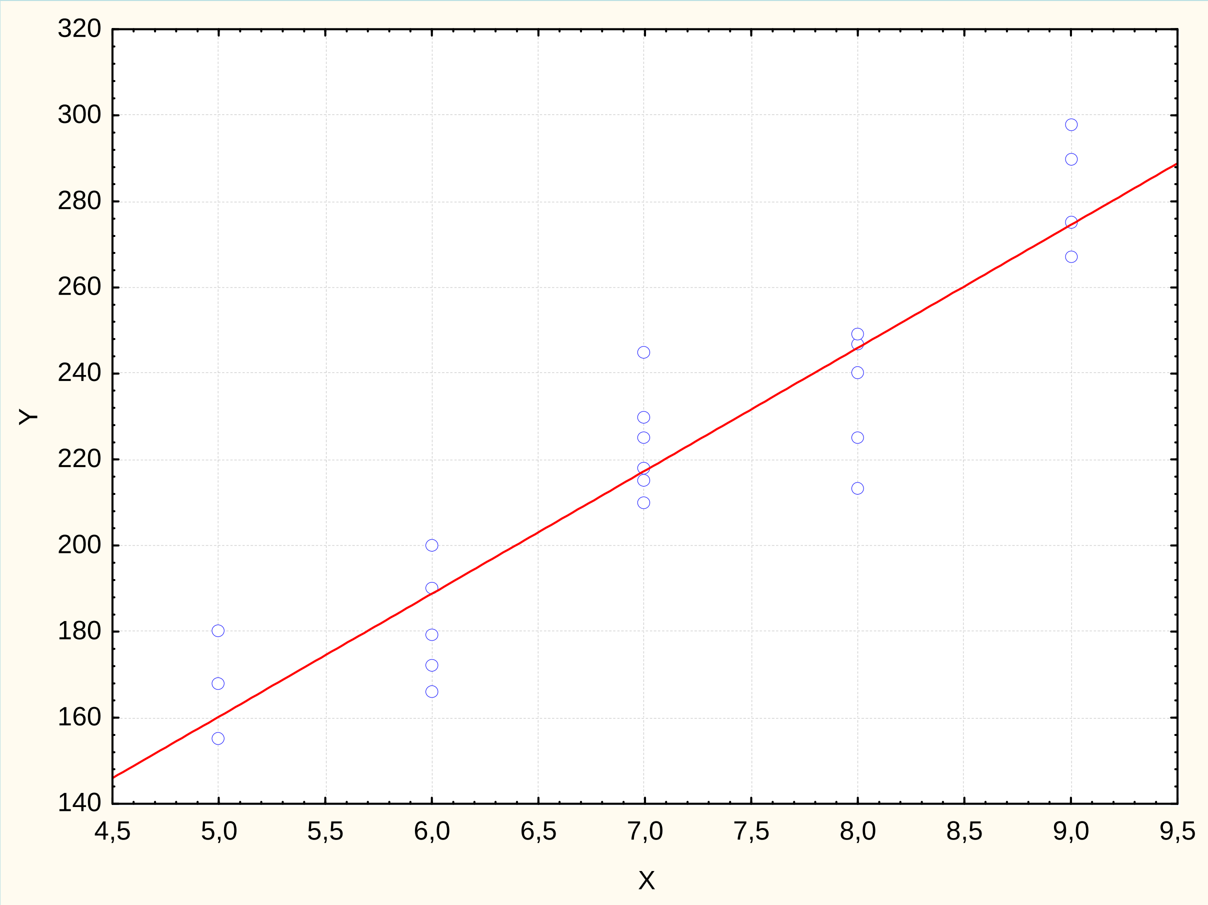
Stanovíme kritický obor W = $\langle F_{1-\alpha}(r-p-1, n-r), \infty) = \langle F_{0,95}(3, 18), \infty) = \langle 3{,}161, \infty)$.

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	155
2	5	168
3	5	180
4	6	166
5	6	172
6	6	179
7	6	190
8	6	200
9	7	210
10	7	215
11	7	218
12	7	225
13	7	230
14	7	245
15	8	213
16	8	225
17	8	240
18	8	247
19	8	249
20	9	267
21	9	275
22	9	290
23	9	298

Odhadneme parametry regresní přímky:

	Výsledky regrese se závislou proměnnou : Y (ceny_bytu.sta) R= ,92825096 R2= ,86164984 Upravené R2= ,85506173 F(1,21)=130,79 p<,00000 Směrod. chyba odhadu : 15,373					
N=23	Beta	Sm.chyba beta	B	Sm.chyba B	t(21)	Úroveň p
Abs.člen			17,28851	18,00156	0,96039	0,347788
X	0,928251	0,081167	28,58506	2,49950	11,43629	0,000000

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

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

Efekt	Analýza rozptylu (ceny_bytu.sta) Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	30907,90	1	30907,90	130,7888	0,000000
Rezid.	4962,70	21	236,32		
Celk.	35870,61				

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.

Proměnná	Analýza rozptylu (ceny_bytu.sta) Označ. efekty jsou význ. na hlad. p < ,05000 SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Y	32474,11	4	8118,527	3396,500	18	188,6944	43,02473	0,000000

Zde najdeme $S_A = 32474{,}11$.

Vypočteme testovou statistiku $F = \dfrac{(S_A - S_R)/(r-p-1)}{(S_T - S_A)/(n-r)} = \dfrac{(32474{,}1087 - 30907{,}9041)/(5-2)}{(35870{,}6087 - 32474{,}1087)/(23-5)} = 2{,}768$

a najdeme kritický obor $W = \langle 3{,}161, \infty)$. Jelikož $F \notin W$, 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.

	Test kvality modelu (ceny_bytu.sta)										
Závislá Proměnná	SČ Rezidua	sv Rezidua	PČ Rezidua	SČ Chyba	sv Chyba	PČ Chyba	SČ Kvali proložen	SV Kvali proložen	PČ Kvali proložen	ta F	p
Y	4962,705	21	236,3193	3396,500	18	188,6944	1566,205	3	522,0682	2,766739	0,071737

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: $$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_{i-1}^{\;2}}.$$

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.

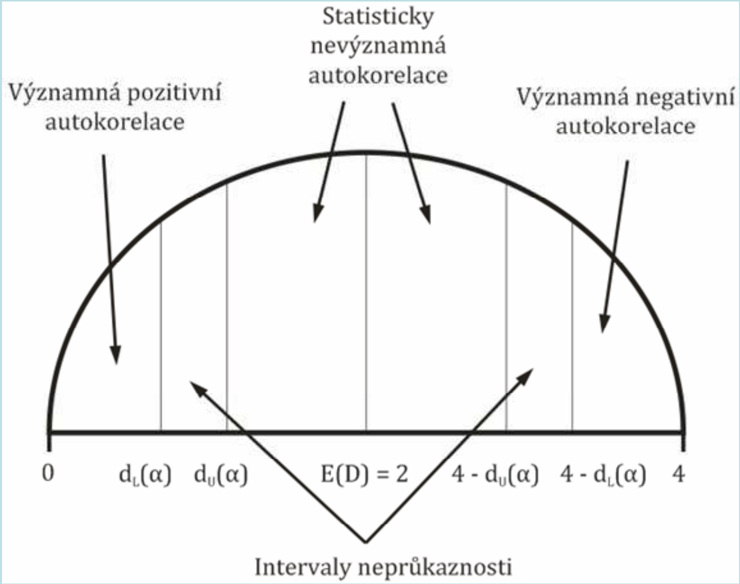
Pro dané α, daný počet pozorování n a daný počet p regresních parametrů – bez konstanty (v případě regresní přímky p = 1) jsou tabelovány kritické hodnoty $d_L(\alpha), d_U(\alpha)$.

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

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

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

Je-li $4 - d_U(\alpha) \leq D \leq 4 - d_L(\alpha)$, 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					
Paramet.	Param.	Asympt. SmCh	Asympt. t(19)	p	Dolní 95% spol	Horní 95% spol
p(1)	0,599248	0,189523	3,161883	0,005134	0,202573	0,995924

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 : nove y (prodejna_software.sta) R= ,96958525 R2= ,94009556 Upravené R2= ,93304798 F(2,17)=133,39 p<,00000 Směrod. chyba odhadu : ,84933					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,1238	2,696857	-7,46194	0,000001
x	4,58359	0,453331	1,5323	0,151549	10,11091	0,000000
xkv	-3,78264	0,453331	-0,0169	0,002027	-8,34410	0,000000

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 : 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

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:

	Durbin-Watson.d	Sériové korelace
Odhad	1,356630	0,256281

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 : nove y2 (Tabulka12) R= ,97136268 R2= ,94354546 Upravené R2= ,93690375 F(2,17)=142,06 p<,00000 Směrod. chyba odhadu : ,82061					
N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-19,7523	2,605683	-7,58046	0,000001
x	4,53932	0,440084	1,5103	0,146425	10,31467	0,000000
xkv	-3,73197	0,440084	-0,0166	0,001958	-8,48013	0,000000

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

Přehled linearizujících transformací

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

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 $y = \beta_0 \beta_1^{\,x}$. 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	12	2
2	8	1,2
3	76,4	14,8
4	17	8,3
5	21,3	8,4
6	10	3
7	12,5	4,8
8	97,3	15,6
9	88	16,1
10	25	11,5
11	38,6	14,2
12	47,3	14

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 : ln y (hotely.sta) R= ,95851605 R2= ,91875303 Upravené R2= ,91062833 F(1,10)=113,08 p<,00000 Směrod. chyba odhadu : ,26364					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(10)	Úroveň p
Abs.člen			1,855881	0,154338	12,02480	0,000000
X	0,958516	0,090137	0,150428	0,014146	10,63398	0,000001

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

N=12	Výsledky regrese se závislou proměnnou : ln y (hotely.sta) R= ,95851605 R2= ,91875303 Upravené R2= ,91062833 F(1,10)=113,08 p<,00000 Směrod. chyba odhadu : ,26364						
	Beta	Sm.chyba beta	B	Sm.chyba B	t(10)	Úroveň p	b =exp(B)
Abs.člen			1,855881	0,154338	12,02480	0,000000	6,397333
X	0,958516	0,090137	0,150428	0,014146	10,63398	0,000001	1,162332

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).

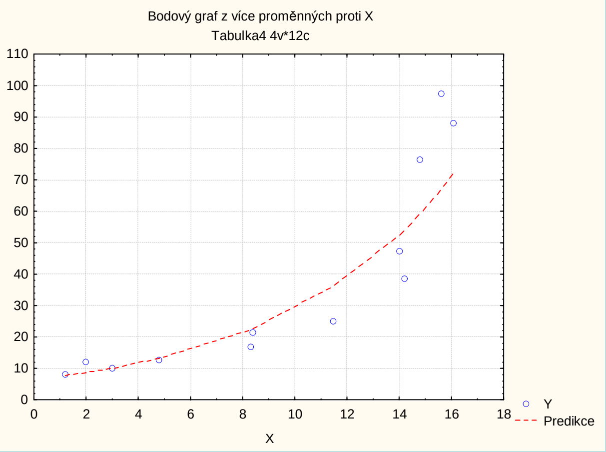
Proměnná	Předpovězené hodnoty (hotely.sta) proměnné: ln y			
	b-váha	Hodnota	b-váha * Hodnot	predikce =exp(v3)
X	0,150428	10,00000	1,504281	4,500918
Abs. člen			1,855881	6,397333
Předpověď			3,360163	28,79387
-95,0%LS			3,189835	24,28441
+95,0%LS			3,530490	34,14071

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 Předpovědi	4 Predikce
1	8	1,2	2,04	7,66
1	12	2	2,16	8,64
3	10	3	2,31	10,05
4	12,5	4,8	2,58	13,17
5	17	8,3	3,10	22,30
6	21,3	8,4	3,12	22,63
7	25	11,5	3,59	36,08
8	47,3	14	3,96	52,56
9	38,6	14,2	3,99	54,16
10	76,4	14,8	4,08	59,28
11	97,3	15,6	4,20	66,86
12	88	16,1	4,28	72,08

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.



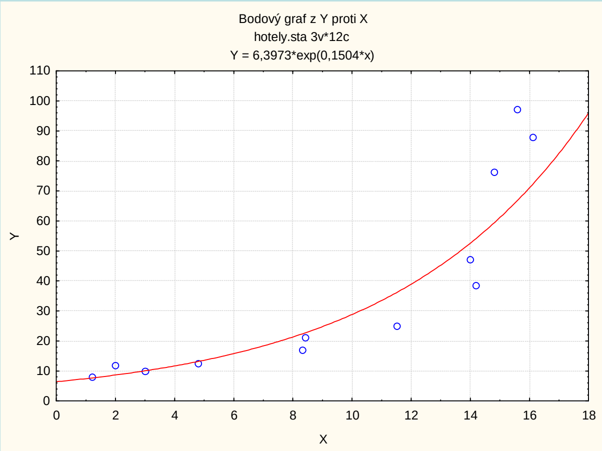
Provedení regresní analýzy pomocí modulu Jednoduchá nelineární regrese

Pro data z předešlého příkladu najdeme odhady parametrů modelu $y = \beta_0 \beta_1^{x}$ 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 $y = \beta_0 \beta_1^{x}$ 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 $\alpha = 0{,}05$, 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