Osnova přednášky Pokročilé metody v jednoduché lineární regresi

1. Dvě nezávislé regresní přímky
- **1.1. Popis modelů a označení**
- **1.2. Test homoskedasticity náhodných odchylek**
- **1.3. Test totožnosti dvou regresních přímek**
- **1.4. Test rovnoběžnosti dvou regresních přímek**
- **1.5. Příklad**

2. Test adekvátnosti regresního modelu

3. Linearizující transformace
- **3.1. Přehled linearizujících transformací**
- **3.2. Příklad**
- **3.3. Provedení regresní analýzy pomocí modulu Jednoduchá nelineární regrese**
- **3.4. Získání odhadů parametrů modelu** $y = \beta_0 \beta_1^{x}$ **pomocí Bodových grafů**

1. Dvě nezávislé regresní přímky

1.1. Popis modelů a označení

Předpokládáme, že máme dva nezávislé regresní modely

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \ldots, n$$

$$Y_i^* = \beta_0^* + \beta_1^* x_i^* + \varepsilon_i^* \text{ , } i = 1, 2, \ldots, n^*$$

Přitom platí, že

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}),\ \boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*) \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}) \text{ ,}$$

vektory $\boldsymbol{\varepsilon}$ a $\boldsymbol{\varepsilon}^*$ jsou nezávislé.

Označíme

b_1, b_1^* odhady směrnic v 1. a 2. modelu,

S_E a S_E^* reziduální součty čtverců v 1. a 2. modelu,

S^2 a S^{*2} výběrové rozptyly nezávisle proměnných veličin v 1. a 2. modelu.

1.2. Ověření předpokladu o homoskedasticitě náhodných odchylek ε a ε*

$$H_0 : \frac{\sigma^2}{\sigma^{*2}} = 1 \text{ proti } H_0 : \frac{\sigma^2}{\sigma^{*2}} \neq 1$$

Testová statistika: $T_0 = \dfrac{\dfrac{S_E}{n-2}}{\dfrac{S_E^*}{n^*-2}}$ má rozložení F(n-2, n^*-2), pokud H_0 platí.

Kritický obor: $W = \left\langle 0, F_{\alpha/2}\left(n-2, n^*-2\right)\right\rangle \cup \left\langle F_{1-\alpha/2}\left(n-2, n^*-2\right), \infty\right)$

1.3. Test totožnosti dvou regresních přímek

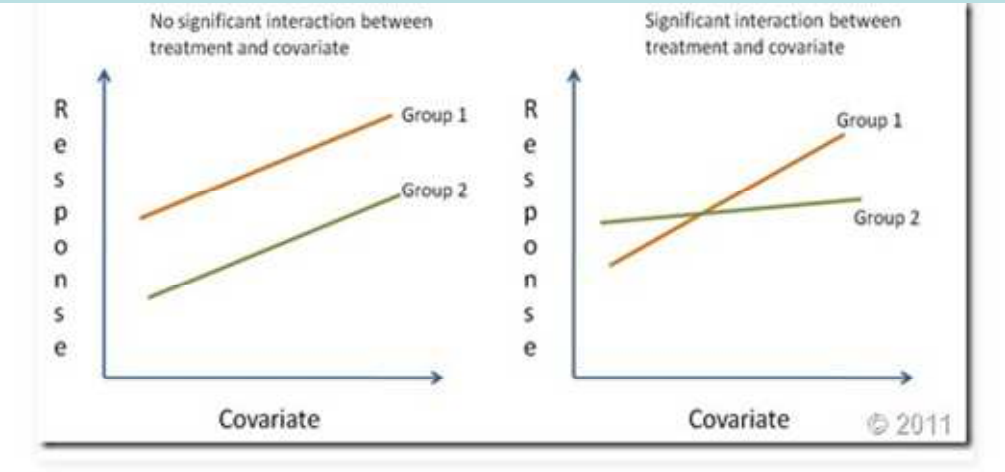
$$H_0 : \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} \text{ proti } H_1 : \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \neq \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix}$$

Sestavíme nový model regresní přímky, v němž hodnoty nezávisle proměnné veličiny vzniknou sdružením hodnot x_i a x_i^* a hodnoty závisle proměnné veličiny vzniknou sdružením hodnot y_i a y_i^*. Označíme S_{EE^*} reziduální součet čtverců v tomto sdruženém modelu.

Testová statistika: $T_0 = \dfrac{\dfrac{S_{EE^*} - S_E - S_E^*}{2}}{\dfrac{S_E + S_E^*}{n + n^* - 4}}$ se v případě platnosti H_0 řídí rozložením $F(2, n+n^*- 4)$.

Kritický obor: $W = \left\langle F_{1-\alpha}\left(2, n + n^* - 4\right), \infty\right)$

1.4. Test rovnoběžnosti dvou regresních přímek



$H_0 : \beta_1 = \beta_1^*$ proti $H_1 : \beta_1 \neq \beta_1^*$

Testová statistika: $$T_0 = \frac{\left(b_1 - b_1^*\right)\sqrt{n + n^* - 4}}{\sqrt{\left(S_E + S_E^*\right)\left(\frac{1}{(n-1)S^2} + \frac{1}{(n^*-1)S^{*2}}\right)}}$$ se v případě platnosti nulové hypotézy řídí rozložením $t(n + n^* - 4)$.

Kritický obor: $W = \left(-\infty, -t_{1-\alpha/2}\left(n + n^* - 4\right)\right\rangle \cup \left\langle t_{1-\alpha/2}\left(n + n^* - 4\right), \infty\right)$.

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

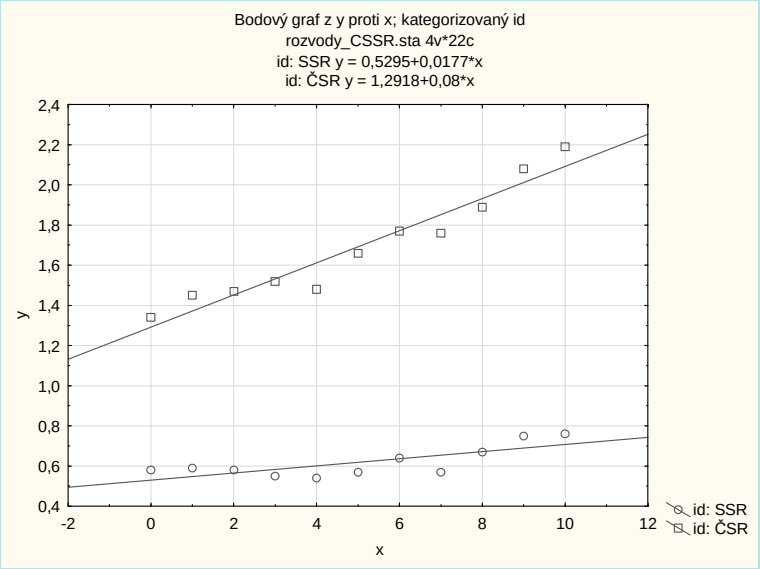
1.5. Příklad: Máme k dispozici údaje o počtu rozvodů za rok, které připadají na 100 000 obyvatel, a to v českých zemích a na Slovensku v letech 1960 – 1970. Hodnoty x_i udávají roky po odečtení 1960, y_i rozvodovost v českých zemích a y_i^* na Slovensku.

x_i	y_i	y_i^*
0	1,34	0,58
1	1,45	0,59
2	1,47	0,58
3	1,52	0,55
4	1,48	0,54
5	1,66	0,57
6	1,77	0,64
7	1,76	0,57
8	1,89	0,67
9	2,08	0,75
10	2,19	0,76

Je zapotřebí zjistit, zda průběh rozvodovosti v letech 1960 – 1970 byl stejný v českých zemích jako na Slovensku. Pokud se ukáže, že stejný nebyl, pak je třeba zjistit, zda aspoň vzestup rozvodovosti byl v obou případech stejný. Testy se mají provádět na hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

Otevřeme datový soubor rozvody_CSSR.sta o třech proměnných a 22 případech. V proměnné x jsou 2x pod sebou hodnoty 0 – 10, v proměnné y jsou pod sebou hodnoty rozvodovosti pro ČSR a pro SSR a proměnná id obsahuje 1 pro ČSR a 0 pro SSR.
Nejprve znázorníme data s proloženými regresními přímkami.
Grafy – Bodové grafy – Proměnné x,y – OK – Kategorizovaný – Kategorie X – Zapnuto – Změnit proměnnou – id – OK – Rozložení Přes sebe – OK.



Z grafického znázornění dat vyplývá, že regresní přímky se zřejmě budou lišit v obou parametrech.

Dále ověříme, zda rozptyly náhodných odchylek $\boldsymbol{\varepsilon}$ a $\boldsymbol{\varepsilon}^*$ jsou shodné.

Statistiky – Vícenásobná regrese – Select cases – Zapnout filtr – zadáme id = 1 – OK – Proměnné y, x – OK – OK – Detailní výsledky – ANOVA.
Analogicky pro 2. model zadáme id = 0.

Efekt	Analýza rozptylu (rozvody_CSSR.sta) Zhrnout podmínku: id=1				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	0,704000	1	0,704000	122,4025	0,000002
Rezid.	0,051764	9	0,005752		
Celk.	0,755764				

Efekt	Analýza rozptylu (rozvody_CSSR.sta) Zhrnout podmínku: id=0				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	0,034568	1	0,034568	12,34801	0,006577
Rezid.	0,025195	9	0,002799		
Celk.	0,059764				

Vypočteme testovou statistiku $T_0 = \dfrac{\dfrac{S_E}{n-2}}{\dfrac{S_E^*}{n^*-2}} = \dfrac{\dfrac{0{,}051764}{9}}{\dfrac{0{,}025195}{9}} = 2{,}0623$.

Kritický obor:

$$W = \left\langle 0, F_{\alpha/2}(n-2, n^*-2)\right\rangle \cup \left\langle F_{1-\alpha/2}(n-2, n^*-2), \infty\right) = \left\langle 0, F_{0,025}(9,9)\right\rangle \cup \left\langle F_{0,975}(9,9), \infty\right) =$$

$$= \left\langle 0;0{,}2484\right\rangle \cup \left\langle 4{,}026;\infty\right)$$

Testová statistika nepatří do kritického oboru, hypotézu o homogenitě rozptylů nezamítáme na hladině významnosti 0,05.

Nyní provedeme **test totožnosti dvou regresních přímek**. Testová statistika má tvar $T_0 = \frac{\left(S_{EE^*} - S_E - S_E^*\right)/2}{\left(S_E + S_E^*\right)/\left(n + n^* - 4\right)}$. Reziduální součty čtverců S_E a S_E^* již známe, S_E = 0,051764 a S_E^* = 0,025195. Stanovíme reziduální součet čtverců S_{EE}^* ve sdruženém modelu.

Statistiky – Vícenásobná regrese – OK – Proměnné – Závislá y, Nezávislé x – OK – OK – Detailní výsledky – ANOVA.

	Analýza rozptylu (rozvody_CSSR.sta)				
Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	0,525284	1	0,525284	1,584552	0,222602
Rezid.	6,630066	20	0,331503		
Celk.	7,155350				

$$T_0 = \frac{\left(6{,}630066 - 0{,}051764 - 0{,}025195\right)/2}{\left(0{,}051764 + 0{,}025195\right)/18} = 766{,}356$$

Kritický obor:

$$W = \left\langle F_{1-\alpha}\left(2, n + n^* - 4\right), \infty\right) = \left\langle F_{0{,}95}\left(2{,}18\right), \infty\right) = \left\langle 3{,}5546; \infty\right)$$

Testová statistika patří do kritického oboru, hypotézu o totožnosti regresních přímek zamítáme na hladině významnosti 0,05. Průběh rozvodovosti v letech 1960 – 1970 byl jiný v ČSR a SSR.

Nakonec budeme **testovat hypotézu o rovnoběžnosti dvou regresních přímek**.
K datovému souboru přidáme novou proměnnou id*x, která vznikne jako součin proměnných id a x.
Statistiky – Vícenásobná regrese – OK – Proměnné – Závislá y, Nezávislé x, id, id*x – OK – OK – Výpočet: výsledky regrese.

N=22	Výsledky regrese se závislou proměnnou : y (rozvody_CSSR.sta) R= ,99460773 R2= ,98924454 Upravené R2= ,98745196 F(3,18)=551,86 p<,00000 Směrod. chyba odhadu : ,06539					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(18)	p-hodn.
Abs.člen			0,529545	0,036883	14,35727	0,000000
x	0,098296	0,034570	0,017727	0,006234	2,84344	0,010783
id	0,668307	0,045731	0,762273	0,052161	14,61383	0,000000
id*x	0,366244	0,051854	0,062273	0,008817	7,06294	0,000001

Testovou statistiku najdeme na řádku id*x, ve sloupci t(18): $t_0 = 7{,}06294$. Odpovídající p-hodnota je velmi blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že vzestup rozvodovosti v letech 1960 – 1970 je stejný v ČSR a SSR.

2. Test adekvátnosti regresního modelu

Hodnoty veličiny Y jsou roztříděny do $r \geq 3$ skupin podle variant $x_{[1]}, ..., x_{[r]}$ veličiny X.

Označme n_i počet pozorování v i-té skupině, i = 1, …, 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: $F = \frac{\left(S_A - S_R\right)/\left(r-p-1\right)}{\left(S_T - S_A\right)/\left(n-r\right)}$ 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 $m\left(x;\beta_0,\beta_1,\ldots,\beta_p\right)$ je vhodným regresním modelem závislosti Y na X.

Těsnost závislosti Y na X vyjádřenou skupinovými průměry měří poměr determinace

$P^2 = \frac{S_A}{S_T}$.

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.

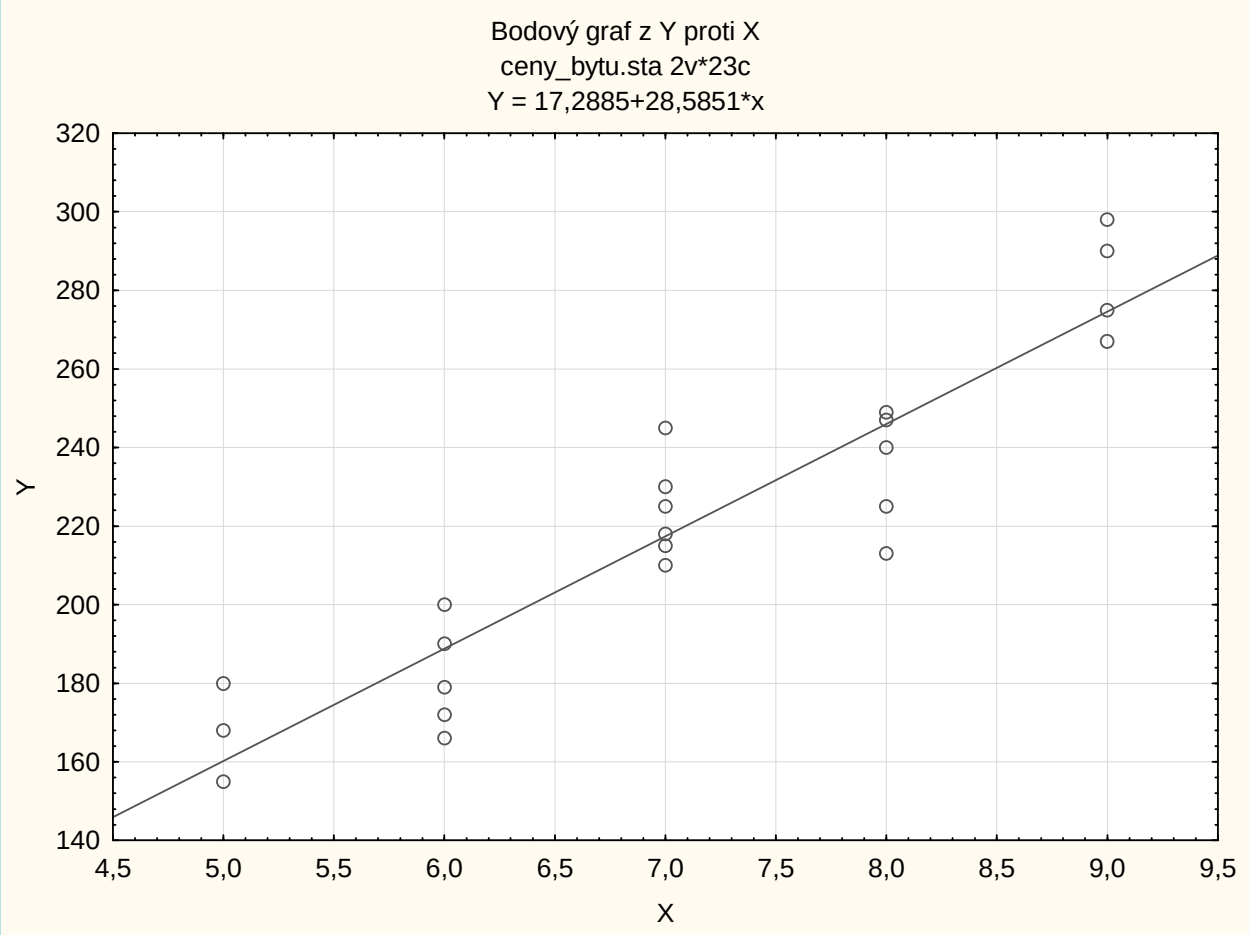
Těsnost závislosti vyjádřete poměrem determinace.

Znázorněte data s proloženou regresní přímkou.

Řešení v systému STATISTICA:

Otevřeme datový soubor ceny_bytu.sta se dvěma proměnnými X a Y a 23 případy:

Parametry regresní přímky získáme pomocí bodových grafů s lineárním proložením:



Pro test adekvátnosti modelu použijeme modul Pokročilé lineární/nelineární modely. 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.

Pro výpočet poměru determinace použijeme cestu:
Statistiky – Základní statistiky/tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné Y, X – OK – OK – Analýza rozptylu. Dostaneme tabulku

	Analýza rozptylu (ceny_bytu.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	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

V 1. sloupci je skupinový součet čtverců S_A a ve 4. sloupci reziduální součet čtverců S_E. Poměr determinace počítáme podle vzorce $P^2 = S_A/S_T = S_A/(S_A + S_E)$. K této tabulce tedy přidáme novou proměnnou P2 a do jejího dlouhého jména napíšeme =v1/(v1+v4).

	Analýza rozptylu (ceny_bytu.sta) Označ. efekty jsou význ. na hlad. p < ,05000								
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p	P2 =v1/(v1+v4)
Y	32474,11	4	8118,527	3396,500	18	188,6944	43,02473	0,000000	0,90531245

Vidíme, že poměr determinace nabývá hodnoty 0,905.

3. 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 $\ln y = \ln \beta_0 + x \ln \beta_1$, 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.

3.1. Přehled linearizujících transformací

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

3.2. 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:

Otevřeme datový soubor hotely.sta se dvěma proměnnými a 12 případy.
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:

	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					
N=12	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).

	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						
N=12	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).

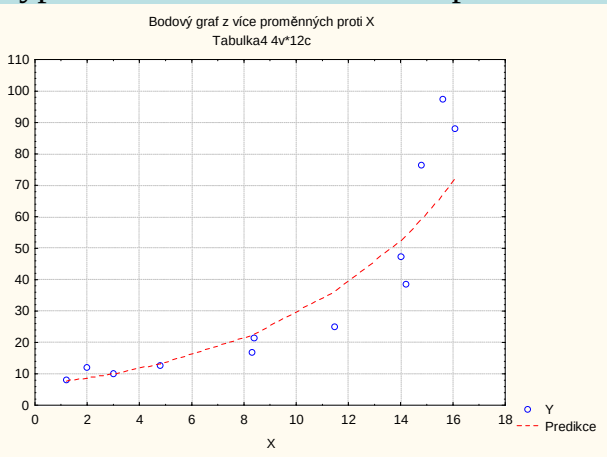
Proměnná	Předpovězené hodnoty (hotely.sta) proměnné: ln y			
	b-váha	Hodnota	b-váha * Hodnot	predikce =exp(v3)
X	0,150428	10,00000	1,504281	4,500918
Abs. člen			1,855881	6,397333
Předpověď			3,360163	28,79387
-95,0%LS			3,189835	24,28441
+95,0%LS			3,530490	34,14071

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.



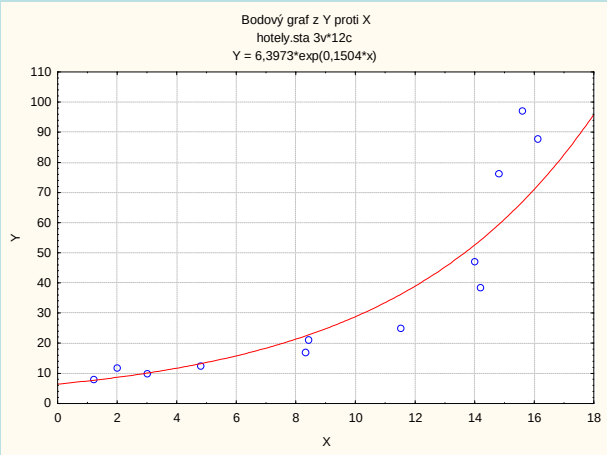
3.3. Provedení regresní analýzy pomocí modulu Jednoduchá nelineární regrese

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.

3.4. 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
$b_0 = 6{,}3973$, $b_1 = e^{0{,}1504} = 1{,}1623$.