Pedagogická fakulta
Faculty of Education
Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích
Pedagogická fakulta
Katedra matematiky

Diplomová práce

Regresní analýza v systému STATISTICA

Vypracoval: Bc. Břetislav Roháček
Vedoucí práce: doc. RNDr. Vladimíra Petrášková, Ph.D.

České Budějovice 2018

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma Regresní analýza v systému STATISTICA jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své diplomové práce, a to v nezkrácené podobě, elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích 12. července 2018

…….

Břetislav Roháček

Poděkování

Tímto bych rád poděkoval své vedoucí práce doc. RNDr. Vladimíře Petráškové, Ph.D. za její podporu, trpělivost, spolupráci, přístup, cenné rady a čas, který mi věnovala při tvorbě této diplomové práce.

Anotace

Diplomová práce je zaměřena na zpracování regresní analýzy jednak ve statistickém software STATISTICA, jednak v tabulkovém procesoru Excel. Práce je rozdělena do dvou stěžejních částí. V teoretické části práce je uveden stručný přehled teorie (lineární regrese s jednou vysvětlující proměnnou, lineární regrese s více vysvětlujícími proměnnými, nelineární regrese). Praktická část obsahuje řešené příklady ve statistickém software STATISTICA, resp. v tabulkovém procesoru Excel, včetně návodu na řešení v těchto dvou programech. V závěru práce je shrnutí výhod a nevýhod software STATISTICA a tabulkového procesoru Excel při řešení dané problematiky.

Annotation

The thesis is focused on doing regression analysis in the STATISTICA and the spreadsheet Excel software. It is divided into two main parts. The theoretical part contains brief overview of the theory (linear regression with one explanatory variable, multiple linear regression with more than one explanatory variable, nonlinear regression). The practical part contains solved regression analysis examples in both the statistical software STATISTICA and the spreadsheet Excel and also includes step by step instructions for both these applications. The last part contains discussion about advantages and disadvantages of using STATISTICA software and Excel software when doing regression analysis.

Obsah

1 Úvod

Statistika je v dnešní době důležitým oborem, která stále více získává na důležitosti, jelikož v každém vědním oboru se vytvářejí hypotézy, které je třeba ověřit. A tak se zvyšuje poptávka po kvalitních materiálech vztahující se k této problematice a po lidech, kteří by tomuto oboru rozuměli. Proto doufám, že by tato práce mohla obojímu přispět.

Práce je zaměřena na jednu ze statistických metod, konkrétně na regresní analýzu. Práci jsem rozdělil na dvě části – na teoretickou část, kde se čtenář seznámí s teorií vztahující se k regresní analýze a na praktickou část, ve které čtenář nalezne řešené příklady, které byly inspirované praxí, ve dvou známých programech STATISTICA a Excel. Na závěr jsou uvedeny výhody a nevýhody obou těchto programů v rámci příkladů, které se dotýkají regresní analýzy.

Toto téma jsem si zvolil z důvodu mého rostoucího zájmu o statistické disciplíny a také proto, že mi dané téma není úplně cizí a již s ním mám určité zkušenosti.

2 Cíl práce a metodika

Cíl práce

Cílem této práce je:

1) Shrnutí teorie vztahující se k regresní analýze.
2) Vytvoření detailních postupů řešení příkladů ve statistickém softwaru STATISTICA a tabulkovém procesoru Excel.
3) Porovnání výhod a nevýhod programů STATISTICA a Excel

Metodika

V rámci práce je zpracována regresní analýza tak, aby text mohl sloužit studentům vysokých škol k pochopení dané problematiky. Práce je rozdělena do dvou stěžejních částí: teoretická část a praktická část. V teoretické části jsou shrnuty základní poznatky regresní analýzy. Praktická část obsahuje řešené příklady a příklady k procvičení, u kterých jsou zaznamenány výsledky. Příklady byly řešeny jednak pomocí software STATISTICA 12, jednak pomocí Excelu. V obou případech je popsán návod, jak v daném software pracovat.

Při zpracování diplomové práce bylo postupováno následovně:

1) Prostudování doporučené literatury.
2) Seznámení s programem STATISTICA 12.
3) Seznámení s programem Excel:
   - typy grafů,
   - doplněk Analýza dat.
4) Vypracování teoretického základu.
5) Podrobné řešení vzorových příkladů. Zadání příkladů k procvičení (včetně výsledků).

3 Teoretická část

Význam slova regrese má podle Slovníku spisovného jazyka českého z roku 1968 znamenat zpětný postup, návrat (opak k progresu). Velký anglicko-český slovník z roku 1984 překládá slovo regression jako ustupování, ústup, krok zpět, pokles, zpětný pohyb, návrat a jako odborný termín je také uveden pojem regrese. Francis Galton kdysi zkoumal závislost tělesné výšky synů na tělesné výšce otců a zjistil, že je zřejmá tendence návratu k celkovému průměru výšky mužů v generaci synů. Jednalo se o to, že otcové, kteří jsou například o 10 cm vyšší, než je průměrná výška mužů jejich generace, mají syny v průměru třeba jen o 5 cm vyšší, než je průměrná výška mužů v generaci synů. Hovoříme tak o tzv. **regresi k průměru**, návrat k průměru. [1]

Regresní analýza se zabývá **vztahem mezi proměnnými**. Konkrétně se jedná o jednu závislou proměnnou a jednu či více nezávislých proměnných. Vztahem rozumíme například vztah lineární, kvadratický, logaritmický, exponenciální, atp. U těchto vztahů pak můžeme provádět odhady (predikce) závisle proměnné a určit, s jakou chybou tento odhad odpovídá realitě. Také lze testovat důležitost jednotlivých parametrů vztahu ve snaze jeho zjednodušení. [2]

Předpokládá se, že čtenář této práce má znalosti základních pojmů z oblasti statistiky, ovládá diferenciální počet (derivace) a rozumí testování hypotéz.

3.1 Definice pojmů

V této kapitole se budeme zabývat vysvětlením základních pojmů, které se vztahují k regresní analýze.

Proměnné

Nezávisle proměnnou značíme X a jedná se o tzv. **regresor**, též označován jako vysvětlující proměnná. Závisle proměnnou značíme Y a jedná se o tzv. **regresand**, též označován jako vysvětlovaná proměnná nebo cílová proměnná [2]. Osobně mi tato značení přijdou logická a přesně je vystihují. Proměnná Y je závislá na proměnné X, respektive proměnných X_i (kde i udává, o kolikátou proměnnou se jedná), a je pomocí nich vysvětlována. Také se na ni zaměřujeme při výpočtech (snažíme se určit její hodnotu, respektive hodnoty), proto je pro nás cílová.

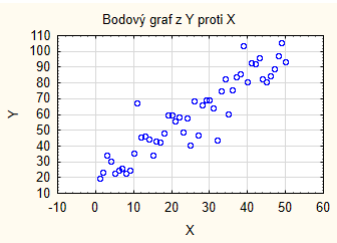
Tabulka 1: Značení proměnných

X	regresor	nezávislá proměnná	vysvětlující proměnná	-
Y	regresand	závislá proměnná	vysvětlovaná proměnná	cílová proměnná

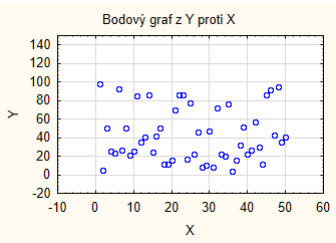
Korelační pole

Korelační pole je grafickým zobrazením dvou veličin, které tvoří dvojice hodnot, tzv. párová data $[x, y]$, kde jsou hodnoty y vysvětlovány pomocí hodnot x v rámci jedné vysvětlující proměnné. V podstatě se jedná o bodový graf, ve kterém zkoumáme vztah mezi proměnnými. V grafu se zaměřujeme na míru závislosti (jak moc data vypadají nahodile), tvar závislosti mezi proměnnými (zda je lineární, kvadratický, apod.), zda data neobsahují podezřele vychýlené body. [3]

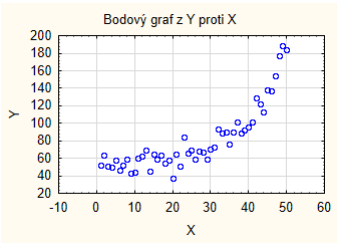
Graf 1: Příklad lineární závislosti



Graf 2: Příklad nezávislosti



Graf 3: Příklad nelineární závislosti



3.2 Lineární regrese

Lineární regrese je specifická tím, že její rovnice jsou tvořeny **lineární kombinací parametrů**. Lineárními jsou tzv. regresní koeficienty, ovšem nezávislé proměnné být lineární nemusí. Rovnice lineární regrese obsahují jednu závislou proměnnou a jednu či více nezávislých proměnných. [4]

3.2.1 Regresní rovnice

Obecný zápis regresní rovnice má tento tvar [5]:

$$Y = f(x). \tag{3.1}$$

Regresní rovnice vyjadřuje vztah mezi regresandem a regresory, vysvětluje závislost proměnné Y na hodnotách x pomocí regresní funkce f. Regresní rovnici nazýváme též **regresním modelem**. V rámci měření se předpokládá, že hodnoty cílové proměnné mohou být naměřeny s případnou chybou ε, též označována jako náhodná složka: [5]

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, 2, \dots, n, \tag{3.2}$$

kde…

y_i ..... je i-tá hodnota závisle proměnné,

x_i ..... je i-tá hodnota nezávisle proměnné,

ε_i ..... je i-tá chyba měření hodnoty y_i,

n ..... je počet pozorování.

Regresní funkce obsahuje kromě nezávislých proměnných také **regresní koeficienty** (typicky značené jako β_0, β_1, atd.) určující tvar námi zkoumaného vztahu.

Pomocí regresní rovnice jsme schopni provádět predikce závisle proměnné při zvolených hodnotách nezávisle proměnných. Tyto predikce můžeme provádět uvnitř intervalu zadaných hodnot regresorů (tzv. interpolací), ale i mimo něj (tzv. extrapolací) a například v případě času jako nezávisle proměnné určit předpověď pro budoucí hodnoty závisle proměnné. [2]

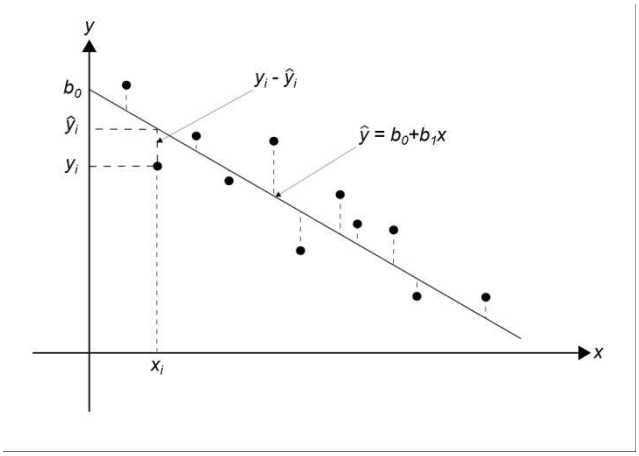
3.2.2 Jednoduchá lineární regrese

Jednoduchá lineární regrese či prostá lineární regrese je speciálním případem lineární regrese o jedné vysvětlující proměnné [4]. Vyjadřuje lineární vztah mezi závislou a nezávislou proměnnou. Regresní funkce odpovídá lineární funkci, přímce [5]:

$$f(x) = \beta_0 + \beta_1 x, \tag{3.3}$$

kde β_0 je absolutní člen, jedná se o pevnou hodnotu, která je přičtena k výpočtu cílové proměnné a β_1 je lineární člen určující sklon přímky, udává, o kolik se změní hodnota Y při navýšení hodnoty X o jednotku.

Graf 4: Lineární regresní model

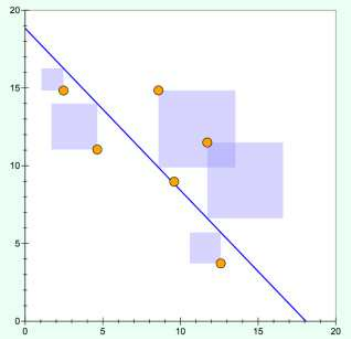


Zdroj: https://rmanual.fri.uniza.sk/?page_id=124

K získání rovnice (3.3) je nejprve třeba určit bodové odhady parametrů β_0 a β_1. Odhady parametrů najdeme pomocí tzv. **metody nejmenších čtverců** (MNČ), která je založena na minimalizaci součtu čtvercových chyb modelu [5]:

$$m(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2. \tag{3.4}$$

Graf 5: Ukázka čtvercových chyb modelu



Zdroj: https://phet.colorado.edu/en/simulation/least-squares-regression

Výraz (3.4) je konvexní funkcí o dvou proměnných a tak je minimální tehdy, jsou-li všechny parciální derivace podle regresních koeficientů β_0 a β_1 rovny nule [6]:

$$\frac{\delta m(\beta_0, \beta_1)}{\delta \beta_0} = 0, \qquad \frac{\delta m(\beta_0, \beta_1)}{\delta \beta_1} = 0. \tag{3.5}$$

Po úpravě těchto rovnic získáme bodové odhady parametrů β_0 a β_1, které značíme $\hat{\beta}_0$ a $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{3.6}$$

kde $\bar{x}$ a $\bar{y}$ udává průměr hodnot dle vzorců:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{3.7}$$

Model jednoduché lineární regrese je možné využít i v případě, že zadaná nezávislá proměnná není lineární. Například pro model $y_i = \beta_0 + \beta_1 x_i^3$ by stačilo zavést novou proměnnou $x_i' = x_i^3$ a podle toho upravit data, získali bychom pak původní lineární model:

$$y_i = \beta_0 + \beta_1 x_i'. \tag{3.8}$$

Intervalové odhady regresních koeficientů

Výše uvedené odhady jsou pouze bodovými odhady, ale mohou nastat situace, kdy by nás mohlo zajímat rozpětí (interval), ve kterém se daný parametr nachází se zvolenou spolehlivostí. Takové odhady pak nazýváme intervalové odhady.

Intervalový odhad koeficientu β_1 (3.9) a β_0 (3.10) se spolehlivostí $(1-\alpha)$ [6]:

$$\langle \hat{\beta}_1 - t^* \sqrt{\frac{1}{n-2} \cdot \frac{\sum(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2}};\ \hat{\beta}_1 + t^* \sqrt{\frac{1}{n-2} \cdot \frac{\sum(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2}} \rangle, \tag{3.9}$$

$$\langle \hat{\beta}_0 - t^* \sqrt{\frac{\sum x_i^2}{n(n-2)} \cdot \frac{\sum(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2}};\ \hat{\beta}_0 + t^* \sqrt{\frac{\sum x_i^2}{n(n-2)} \cdot \frac{\sum(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2}} \rangle. \tag{3.10}$$

kde t^* udává hodnotu studentova rozdělení na hladině $1 - \frac{\alpha}{2}$ při $n-2$ stupních volnosti, standardně značeno $t_{1-\frac{\alpha}{2}}(n-2)$.

Pás spolehlivosti

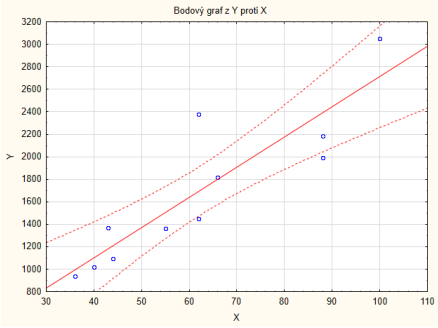
Intervalový odhad cílové proměnné Y se spolehlivostí $(1-\alpha)$:

$$\langle \hat{y} \pm t^* \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2} \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right)} \rangle. \tag{3.11}$$

kde t^* udává hodnotu studentova rozdělení na hladině $1 - \frac{\alpha}{2}$ při $n-2$ stupních volnosti, standardně značeno $t_{1-\frac{\alpha}{2}}(n-2)$.

Určíme-li intervalové odhady ze ((3.11) pro všechny hodnoty X, vznikne nám tzv. **pás spolehlivosti**, viz příklad Graf 6. Hranice pásu mají tvar hyperboly. [5]

Graf 6: Pás spolehlivosti



Předpoklady jednoduché lineární regrese

Oprávněnost využití jednotlivých sestavených modelů v lineární regresi je podmíněna splněním následujících předpokladů ([7]; [8]):

Předpoklady vztahující se k datům

- Nezávislá proměnná je měřena bez chyb.
- Rozptyl hodnot regresoru je větší než jedna (všechny hodnoty nejsou stejné).

Předpoklady vztahující se k modelu

- Model je lineární v parametrech.
- Náhodná složka je náhodnou veličinou s normálním rozdělením.
- Střední hodnota náhodné složky je nulová:

$$\mathrm{E}(\varepsilon) = 0. \tag{3.12}$$

- Rozptyl náhodné složky je konstantní:

$$\mathrm{var}(\varepsilon) = c. \tag{3.13}$$

- Chyby v modelu jsou nekorelované, tedy hodnota kovariance je nulová:

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \tag{3.14}$$

  pro každé $i, j = 1, 2, \dots, n$, kde $i \neq j$.

- Nezávislá proměnná a náhodná složka jsou nekorelované.

3.2.3 Reziduum

Reziduum, chyba odhadu či odhad hodnoty náhodné složky, značeno e, případně $\hat{\varepsilon}$, vyjadřuje rozdíl (odchylku) odhadované hodnoty $\hat{y}_i$ od naměřené hodnoty y_i, která se k odhadu vztahuje:

$$e_i = y_i - \hat{y}_i. \tag{3.15}$$

V podstatě nám to říká, jak moc se náš odhad liší od naměřené hodnoty, jak velké chyby jsme se v našem odhadu dopustili. Tato hodnota rovněž zahrnuje další faktory, které ovlivňují závislou proměnnou a nejsou vysvětleny pomocí regresorů zahrnuté v modelu. [4]

3.2.4 Součty čtverců

Mezi další důležité parametry patří součty kvadratických odchylek, které typicky slouží jako mezivýpočet k dalším výpočtům. Samostatně vyjadřují jakýsi součet všech chyb, kterých jsme se v našem modelu dopustili, je to taková celková chybovost modelu. K těmto součtům čtverců patří:

- reziduální součet čtverců,
- celkový součet čtverců,
- vysvětlený součet čtverců.

Tabulka 2: Značení součtů čtverců

	Anglicky	**Česky**
RSS	Residual sum of squares	Reziduální součet čtverců
SSR	Sum of squared residuals	Součet čtvercových reziduí
SSE	Sum of squared errors of prediction	Součet čtvercových chyb odhadu
TSS	Total sum of squares	Celkový součet čtverců
ESS	Explained sum of squares	Vysvětlený součet čtverců

RSS, SSR, SSE

Reziduální součet čtverců, značeno RSS, SSR či SSE, určuje součet kvadratických odchylek odhadů od naměřených hodnot. Jedná se o součet druhých mocnin reziduí (čtvercových chyb odhadu): [2]

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (e_i)^2. \tag{3.16}$$

Čím je tento součet větší, tím je pochopitelně model méně vhodný. Proto se snažíme, aby byl součet co nejmenší. Na tomto výpočtu je založena metoda nejmenších čtverců.

TSS

Celkový součet čtverců, značeno TSS (total sum of squares), určuje součet kvadratických odchylek naměřených hodnot od jejich průměru, hovoříme o tzv. **celkové variabilitě** [2]:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{3.17}$$

ESS

Vysvětlený součet čtverců, značeno ESS (explained sum of squares), určuje součet kvadratických odchylek odhadů od průměru naměřených hodnot, hovoříme o tzv. **vysvětlené variabilitě** [2]:

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{3.18}$$

Vztah mezi RSS, TSS a ESS

Součet reziduálního součtu čtverců a vysvětleného součtu čtverců nám dá celkový součet čtverců [9]. Tedy platí vztah:

$$TSS = RSS + ESS. \tag{3.19}$$

3.2.5 Koeficient determinace

Koeficient determinace, index determinace či index spolehlivosti, značíme R^2, nám udává, jak velká část variability cílové proměnné je vysvětlena pomocí nezávislé proměnné a daného modelu. Jedná se o poměr vysvětlené variability a celkové variability závislé proměnné. [2]

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}. \tag{3.20}$$

Koeficient determinace by měl nabývat hodnoty v intervalu $\langle 0; 1 \rangle$. Jeho vynásobením 100 je možné jej vyjádřit v procentech. Ovšem může se stát, že hodnota koeficientu nabude hodnoty mimo tento interval. To se stane, pokud je model určen nesprávně.

3.2.6 Mnohonásobná lineární regrese

Mnohonásobná lineární regrese či vícenásobná lineární regrese je druhým případem lineární regrese. Na rozdíl od jednoduché lineární regrese model obsahuje alespoň dvě nezávislé proměnné X vysvětlující jednu závislou proměnnou Y.

Regresní model o k proměnných má tvar:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \tag{3.21}$$

kde k určuje počet vysvětlujících proměnných. [5]

Maticový zápis modelu:

$$Y = \beta X + \varepsilon, \tag{3.22}$$

kde…

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

a n určuje počet pozorování [5]. Celý zápis by v rovnici vypadal takto:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.23}$$

Pomocí mnohonásobné lineární regrese můžeme nalézt například kvadratický model (který patří mezi polynomické modely):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \tag{3.24}$$

kde bychom nahrazením proměnných $x_{i1} = x_i$, $x_{i2} = x_i^2$ získali lineární model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{3.25}$$

Obecně pro kterýkoli polynomický model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k, \tag{3.26}$$

bychom nahrazením všech proměnných ($x_{i1} = x_i$, $x_{i2} = x_i^2$,… $x_{ik} = x_i^k$) získali model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \tag{3.27}$$

Předpoklady mnohonásobné lineární regrese

Předpoklady užití modelu mnohonásobné lineární regrese jsou stejné jako u jednoduché lineární regrese, ovšem obsahují navíc tyto předpoklady:

- Počet pozorování musí být větší než počet nezávisle proměnných.
- Nezávislé proměnné jsou lineárně nezávislé.
- Předpoklady, které se vztahovaly k jedné vysvětlující proměnné, zde platí pro všechny vysvětlující proměnné.

Koeficient determinace

Je třeba zmínit, že hodnota koeficientu determinace roste s přibývajícím počtem proměnných v modelu. Což může v praxi vést ke zneužití při umělém zvyšování hodnoty koeficientu přidáním irelevantních proměnných do modelu. V takovýchto případech je možné sáhnout po hodnotě takzvaného „upraveného R^2“ (viz níže), které tento problém řeší. Proto také není možné porovnávat neupravené hodnoty R^2 u modelů s odlišným počtem nezávisle proměnných. [10]

Někdy se koeficient determinace uvádí jako R_k^2, kde k určuje počet vysvětlujících proměnných včetně konstanty β_0. [10]

Upravené R^2

Upravený koeficient determinace či upravené R^2, značeno $\bar{R}^2$ či R_{adj}^2 (z anglického adjusted), se používá při řešení situací, kde přidávání nezávisle proměnných do modelu zvyšuje hodnotu klasického koeficientu determinace. Slouží k porovnávání modelů, které se právě liší v počtu proměnných zahrnutých v modelu. Hodnota upraveného R^2 může být záporná a je vždy menší nebo rovna hodnotě R^2. [10]

Upravené R^2 je definováno jako:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{3.28}$$

kde p určuje počet nezávisle proměnných (nezahrnuje konstantu) a n určující počet pozorování [10].

3.2.7 Mnohorozměrná lineární regrese

Mnohorozměrná lineární regrese či vícerozměrná lineární regrese je specifická tím, že se snaží vysvětlit více proměnných Y_r pomocí stejné sady proměnných X_{rk}, kde k určuje počet vysvětlujících proměnných a r určuje počet závislých proměnných (počet rozměrů) [12].

Obecný zápis modelu mnohorozměrné lineární regrese:

$$Y_j = \beta_{0j} + \beta_{1j}X_1 + \cdots + \beta_{kj}X_k \tag{3.29}$$

Maticový zápis:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ y_{21} & y_{22} & \cdots & y_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nr} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1r} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nr} \end{bmatrix} \tag{3.30}$$

Tato regrese je využívána v mnoha oborech, jako je například ekonometrie.

3.3 Nelineární regrese

O nelineární regresi hovoříme v případě, že regresní rovnice není lineární v parametrech, to znamená, že model není tvořen lineární kombinací parametrů. V některých případech je možné model upravit pomocí tzv. **linearizační transformace**, aby se model stal lineární, díky tomu pak můžeme postupovat klasicky podle lineární regrese.

Například exponenciální model $Y = \beta_0 \cdot e^{\beta_1 x_i}$ lze upravit pomocí logaritmu (za předpokladu, že hodnoty Y jsou kladné):

$$\begin{aligned} Y &= \beta_0 \cdot e^{\beta_1 x_i}, \\ \ln Y &= \ln \beta_0 + \ln e^{\beta_1 x_i}, \\ \ln Y &= \ln \beta_0 + \beta_1 x_i, \end{aligned} \tag{3.31}$$

pak při $Y' = \ln Y$, $\beta_0' = \ln \beta_0$ lze model přepsat do podoby:

$$Y' = \beta_0' + \beta_1 x_i \tag{3.32}$$

a získáme tak lineární model.

Další příklady linearizační transformace:

Tabulka 3: Příklady nelineárních modelů a jejich transformací

Model	**Regresní rovnice**	**Transformace**	**Výsledná rovnice**
Hyperbolický	$Y = \beta_0 + \frac{\beta_1}{x_i}$	$x_i' = \frac{1}{x_i}$	$Y = \beta_0 + \beta_1 x_i'$
Mocninný	$Y = \beta_0 \cdot x_i^{\beta_1}$	$Y' = \ln Y,$ $\beta_0' = \ln \beta_0,$ $x_i' = \ln x_i$	$Y' = \beta_0' + \beta_1 x_i'$
Logaritmický	$Y = \beta_0 + \beta_1 \ln x_i$	$x_i' = \ln x_i$	$Y = \beta_0 + \beta_1 x_i'$
Exponenciální	$Y = \beta_0 \cdot e^{\beta_1 x_i}$	$Y' = \ln Y,$ $\beta_0' = \ln \beta_0$	$Y' = \beta_0' + \beta_1 x_i$

3.4 Ověřování vhodnosti modelu

Pomocí techniky testování hypotéz můžeme ověřit vhodnost modelu, eventuálně významnost jednotlivých regresních koeficientů pro případné zjednodušení modelu.

Model testujeme pomocí analýzy rozptylu, též označována jako ANOVA (z anglického analysis of variance). Metodický postup pro aplikaci této metody můžeme nalézt v publikaci J. Anděla [13]. Aby model byl užitečný (významný, vhodný), je třeba, aby alespoň jeden regresní koeficient byl nenulový. Naopak model je neužitečný (nevýznamný, nevhodný), pokud jsou všechny regresní koeficienty rovny nule (z rovnice by totiž nic nezbylo). Při k regresních koeficientech bychom stanovili hypotézy:

H_0: $\beta_0 = \beta_1 = \cdots = \beta_k = 0$ ..... model není významný,

H_A: non H_0 ..... model je významný.

Hodnotu testového kritéria získáme pomocí vzorce:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)}, \tag{3.33}$$

kde k určuje počet nezávislých proměnných a n určuje počet hodnot ze všech regresorů dohromady. Hodnota (3.32) bývá také uváděna v tabulce s výsledky regresní analýzy (např. v programu Excel) v této podobě:

Tabulka 4: Tabulka analýzy rozptylu jednoduchého třídění

Zdroj	Součet čtverců	Stupně volnosti	Podíl	F
Skupiny	ESS	$k-1$	$\frac{ESS}{k-1}$	$\frac{ESS/(k-1)}{RSS/(n-k)}$
Reziduální	RSS	$n-k$	$\frac{RSS}{n-k}$	–
Celkový	TSS	$n-1$	–	–

Zdroj: Anděl [13], str. 154

Hodnota testového kritéria je převedena na p-value a pokud je p-value menší než zvolená hladina významnosti, zamítneme nulovou hypotézu ve prospěch alternativní hypotézy, čímž prokážeme významnost modelu jako celku. V opačném případě bychom významnost modelu neprokázali.

Regresní koeficienty testujeme samostatně každý zvlášť pomocí t-testů. Významnost daného koeficientu prokážeme tehdy, když se jeho hodnota bude významně lišit od nuly. Pokud by totiž byl koeficient nulový, vyrušil by proměnnou, která by s ním byla v součinu. Hypotézy pro koeficient β_i bychom stanovili takto:

H_0: $\beta_i = 0$ ..... koeficient β_i není významný,

H_A: $\beta_i \neq 0$ ..... koeficient β_i je významný.

Hodnoty testových kritérií nalezneme v knize J. Anděla [13]. Jestliže testem určená p-value je menší než zvolená hladina významnosti, zamítneme nulovou hypotézu ve prospěch alternativní hypotézy, čímž prokážeme významnost daného koeficientu. V případě, že bychom významnost koeficientu neprokázali, bylo by možné příslušný koeficient (spolu s proměnnou v součinu) z modelu vypustit a model tak zjednodušit.

Aplikace analýzy rozptylu i t-testů vyžaduje splnění předpokladu normality dat.

4 Praktická část

Praktická část je tvořena řešenými příklady ve statistickém softwaru STATISTICA a v tabulkovém procesoru Excel. Budou zde uvedeny podrobné návody a postupy řešení dvou příkladů v obou těchto programech. Jeden vzorový příklad bude zadán z jednoduché lineární regrese a druhý z mnohonásobné lineární regrese. Dále bude uvedena sada příkladů včetně výsledků zpracované v programu STATISTICA a Excel.

4.1 Software STATISTICA a Excel

Příklady budou řešeny v programech STATISTICA a Excel, konkrétně se jedná o verze STATISTICA Cz 12 (dále jen STATISTICA) a Microsoft Office Excel 2007 (dále jen Excel).

Obrázek 1: Logo – STATISTICA

Zdroj: https://verify.wiki/wiki/STATISTICA

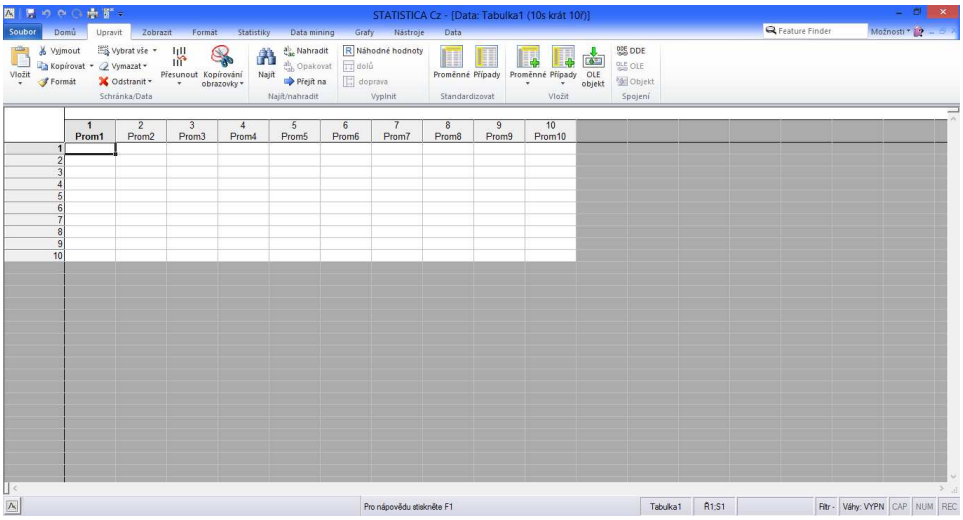
Obrázek 2: Logo – Excel 2007

Zdroj: http://logos.wikia.com/wiki/Microsoft_Excel

STATISTICA

STATISTICA je statistický software vyvinutý společností StatSoft. Jedná se o placený program, který je také k dispozici zdarma ve zkušební (trial) třicetidenní verzi. Plnou verzi poskytuje Jihočeská univerzita v Českých Budějovicích pro výukové a výzkumné aktivity univerzity. Náhled na uživatelské rozhraní můžete vidět níže na obrázku (Obrázek 3).

Obrázek 3: Prostředí STATISTICA



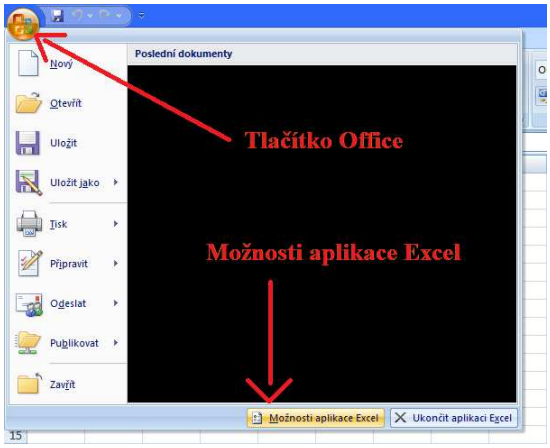
Excel

Microsoft Excel je tabulkový procesor, který je součástí kancelářského balíku Microsoft Office, který má většina uživatelů k dispozici. Je také cenově dostupnější než program STATISTICA. Zatímco za STATISTICU zaplatíte řádově několik desetitisíců, s Microsoft Office se vejdete do 5 000 Kč [11].

K řešení příkladů z regresní analýzy budeme v Excelu potřebovat mít doinstalovaný doplněk nazývaný „Analýza dat“. Postup k přidání doplňku:

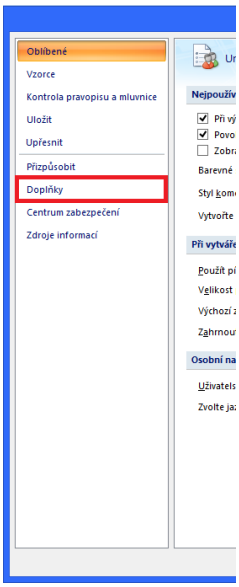
1) V levém horním rohu klikneme na „Tlačítko Office“ a dole z nabídky vybereme „Možnosti aplikace Excel“.

Obrázek 4: Excel – Tlačítko Office a Možnosti aplikace Excel

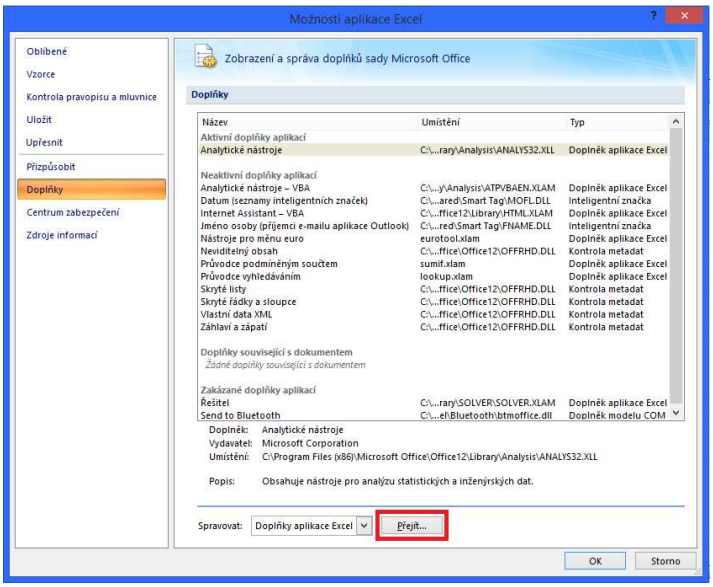


2) V menu možnostech vybereme „Doplňky“ a stiskneme tlačítko „Přejít…“.

Obrázek 5: Excel – Doplňky

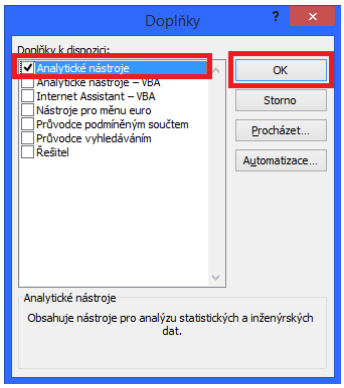


Obrázek 6: Excel – Doplňky – Přejít…



3) Z nabídky doplňků vybereme „Analytické nástroje“ a stiskneme tlačítko „OK“.

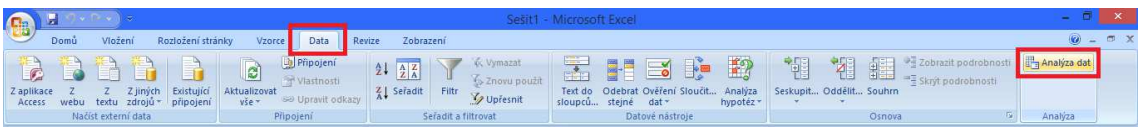
Obrázek 7: Excel – Doplňky – Analytické nástroje



;

4) Jakmile bude doplněk doinstalován, v hlavním menu v kategorii „Data“ vpravo naleznete položku „Analýza dat“.

Obrázek 8: Excel – Menu – Data – Analýza dat



4.2 Příklad jednoduché lineární regrese

V této kapitole si ukážeme řešení vzorového příkladu jednoduché lineární regrese v programech STATISTICA a Excel s podrobným postupem.

4.2.1 Vzorový příklad

Máme k dispozici následující data týkající se kapacit akumulátorových baterií a jejich cen, viz Tabulka 5.

Tabulka 5: Data k vzorovému příkladu

Kapacita [Ah]	36	40	43	44	55	62	62	66	88	88	100
Cena [Kč]	939	1016	1371	1092	1364	2378	1452	1818	1991	2184	3050

Zdroj: EF JU, přednášky prof. RNDr. Anny Čermákové, CSc.

Zadání

1. Nalezněte model, který by vyjadřoval závislost ceny baterie na její kapacitě, a ověřte spolehlivost nalezeného modelu.
2. Nalezněte bodovou předpověď ceny baterie s kapacitou 75 Ah.
3. Určete, o kolik se zvýší cena, pokud se kapacita baterie zvýší o 5 Ah.

Řešení

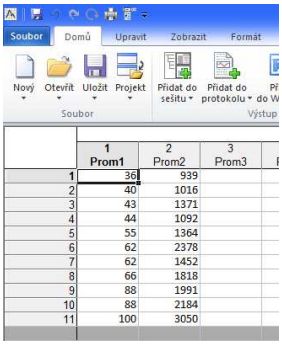
Při řešení této úlohy budeme postupovat následujícím způsobem:

1. a) Vytvoříme korelační pole. Z korelačního pole posoudíme míru a vztah závislosti mezi proměnnými.
   b) Nalezneme vhodný regresní model, nalezneme regresní rovnici a určíme koeficient determinace.
   c) Provedeme test analýzy rozptylu k určení spolehlivosti modelu.
   d) Provedeme dílčí t-testy k určení důležitosti jednotlivých regresních koeficientů.
   e) Ověříme předpoklady užití nalezeného regresního modelu.
2. Nalezneme bodovou předpověď pro kapacitu 75 Ah.
3. Určíme změnu ceny při zvýšení kapacity baterie o 5 Ah.

4.2.2 Řešení v softwaru STATISTICA

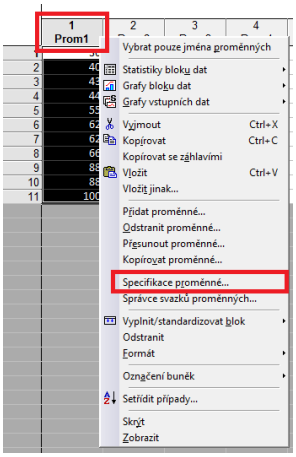
1.a) Nejprve přepíšeme data do programu STATISTICA. Pozor! Proměnné je ve STATISTICE třeba psát do sloupců, nikoliv do řádků, viz Obrázek 9. Pro přehlednost si proměnné vhodně pojmenujeme. Pravým myšítkem klikneme na název „Prom1“ a z nabídky zvolíme „Specifikace proměnné…“, viz Obrázek 10. Přepíšeme jméno a potvrdíme tlačítkem „OK“, viz Obrázek 11. To samé u proměnné „Prom2“.

Obrázek 9: STATISTICA – Data

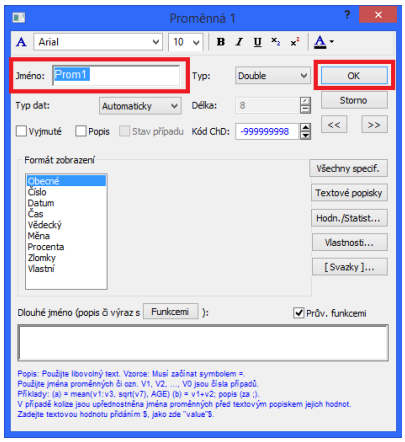


	1 Prom1	2 Prom2	3 Prom3
1	36	939	
2	40	1016	
3	43	1371	
4	44	1092	
5	55	1364	
6	62	2378	
7	62	1452	
8	66	1818	
9	88	1991	
10	88	2184	
11	100	3050	

Obrázek 10: STATISTICA – Specifikace proměnné

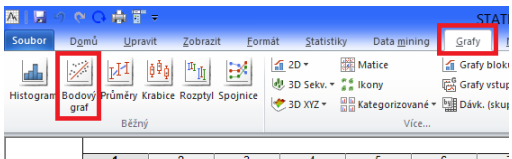


Obrázek 11: STATISTICA – Pojmenování proměnné



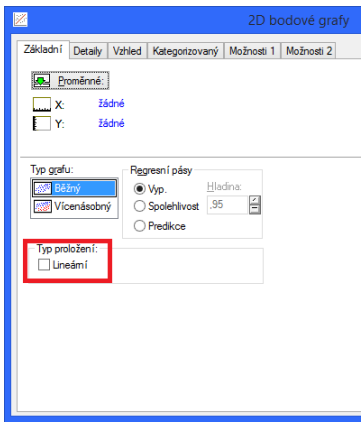
Nyní vytvoříme korelační pole ze zadaných dat, abychom viděli grafické znázornění dat a mohli jsme posoudit vztah závislosti mezi proměnnými. V menu zvolíme položku „Grafy“ a v levé části vybereme „Bodový graf“, viz Obrázek 12. Zobrazí se okno, ve kterém odškrtneme „lineární typ proložení“, viz Obrázek 13.

Obrázek 12: STATISTICA – Menu – Bodový graf



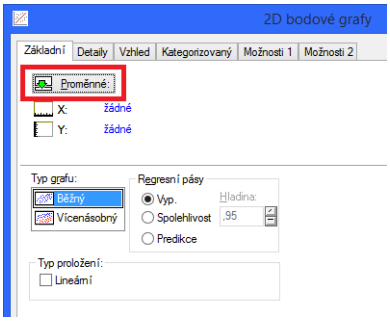
	1 Kapacita	2 Cena	3 Prom3	4 Prom4	5 Prom5	6 Prom6
1	36	939				
2	40	1016				
3	43	1371				
4	44	1092				
5	55	1364				
6	62	2378				
7	62	1452				
8	66	1818				
9	88	1991				
10	88	2184				
11	100	3050				

Obrázek 13: STATISTICA – Nastavení grafu

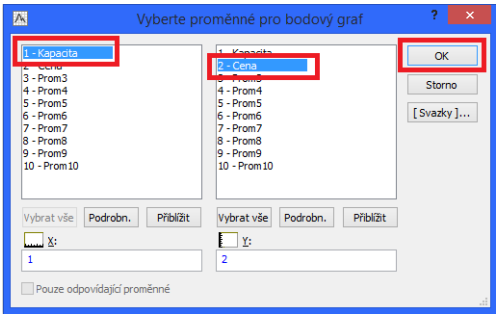


Vybereme proměnné, ze kterých chceme vytvořit graf. Klikneme na položku „Proměnné“ (viz Obrázek 14) a v novém okně zvolíme proměnnou X (Kapacita) a Y (Cena), tlačítkem „OK“ potvrdíme výběr proměnných, viz Obrázek 15.

Obrázek 14: STATISTICA – Proměnné

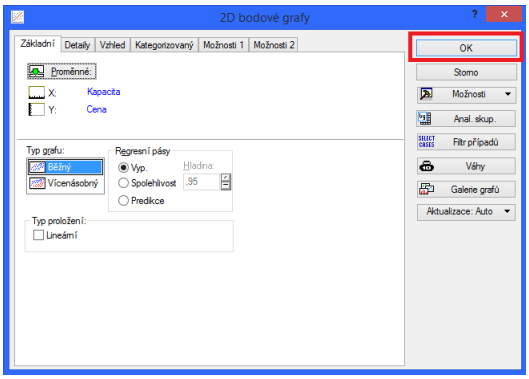


Obrázek 15: STATISTICA – Výběr proměnných

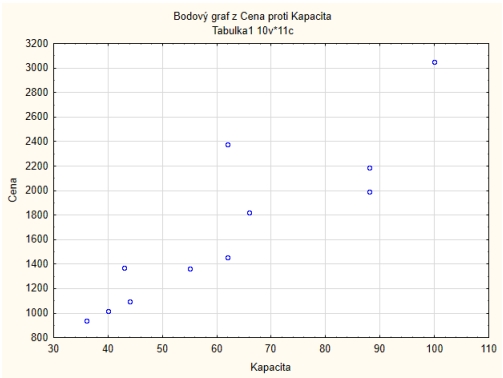


Potvrzením výběru proměnných se vrátíme do předchozího okna. Stisknutím tlačítka „OK“ (viz Obrázek 16) potvrdíme vytvoření grafu.

Obrázek 16: STATISTICA – Vytvořit graf



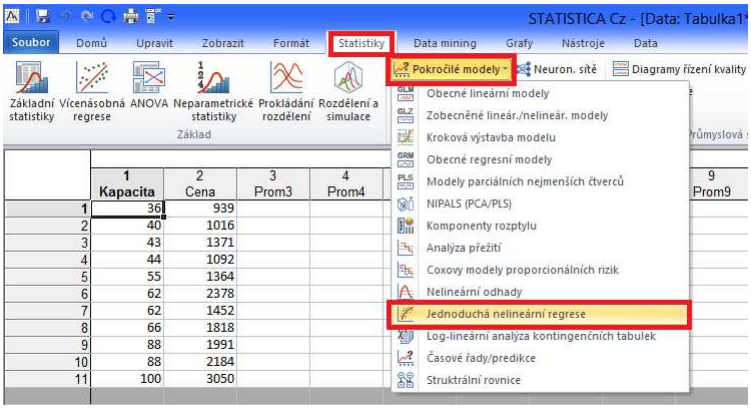
Graf 7: STATISTICA – Korelační pole



Z grafu se zdá, že se jedná o lineární závislost. Nalezneme lineární model ve tvaru $Y = \beta_0 + \beta_1 X$ a určíme jeho vhodnost.

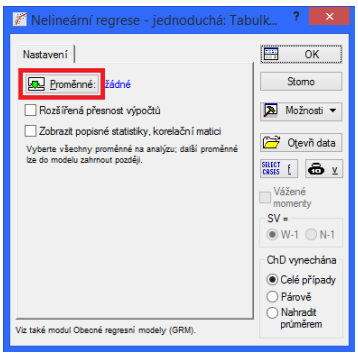
1.b) V menu „Statistiky“ zvolíme „Pokročilé modely“ a vybereme položku „Jednoduchá nelineární regrese“ (slouží pro lineární i nelineární modely), viz Obrázek 17.

Obrázek 17: STATISTICA – Regresní modely

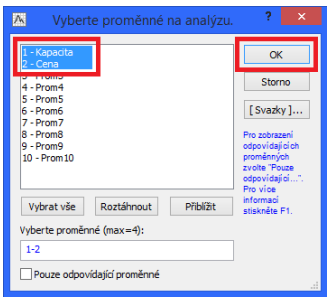


Zobrazí se okno, pomocí kterého zvolíme proměnné, z nichž bude model vytvořen. Klikneme na „Proměnné“, viz Obrázek 18. Označíme obě naše proměnné Kapacita a Cena a stiskneme „OK“, viz Obrázek 19. Pokračujeme stisknutím tlačítka „OK“, viz Obrázek 20.

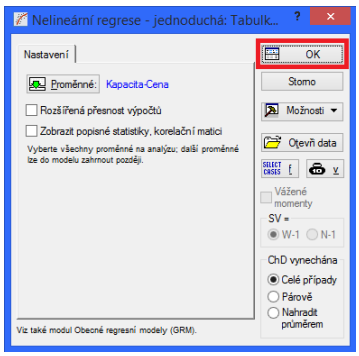
Obrázek 18: STATISTICA – Model – Proměnné



Obrázek 19: STATISTICA – Model – Výběr proměnných

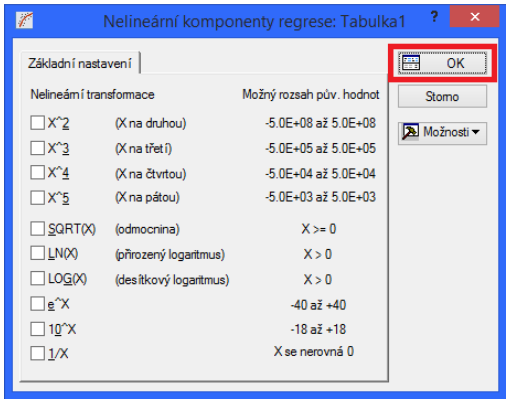


Obrázek 20: STATISTICA – Model – Potvrzení proměnných



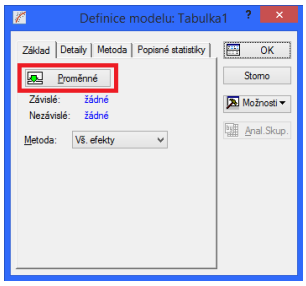
V základním nastavení (viz Obrázek 21) bychom mohli vybrat další nelineární podoby proměnných, které bychom chtěli zahrnout do modelu, což nyní nechceme, jde nám o lineární model, takže jen potvrdíme tlačítkem „OK“.

Obrázek 21: STATISTICA – Model – Základní nastavení

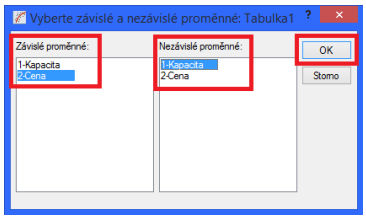


Nyní je třeba specifikovat, které proměnné jsou závislé a které nezávislé. Klikneme na „Proměnné“ (viz Obrázek 22), jako závislou zvolíme Cenu a jako nezávislou Kapacitu, potvrdíme tlačítkem „OK“, viz Obrázek 23. Pokračujeme stiskem „OK“, viz Obrázek 24.

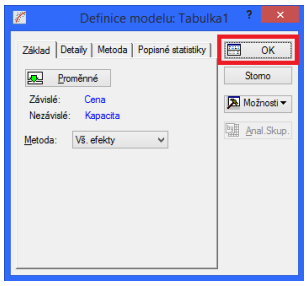
Obrázek 22: STATISTICA – Model – Proměnné



Obrázek 23: STATISTICA – Model – Výběr proměnných

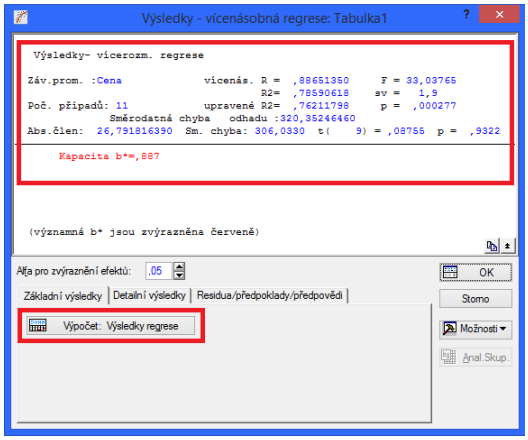


Obrázek 24: STATISTICA – Model – Potvrzení proměnných



V následujícím okně již můžeme v horní části vidět částečné výsledky, viz Obrázek 25. Pomocí tlačítka „Výpočet: Výsledky regrese“ zobrazíme kompletní výsledky, viz Obrázek 26.

Obrázek 25: STATISTICA – Předběžné výsledky



Obrázek 26: STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Cena (Tabulka1)
R= ,88651350 R2= ,78590618 Upravené R2= ,76211798
F(1,9)=33,038 p<,00028 Směrod. chyba odhadu : 320,35

N=11	b*	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			26,79182	306,0330	0,087546	0,932155
Kapacita	0,886513	0,154234	26,84253	4,6700	5,747838	0,000277

Ve sloupci „b“ nalezneme odhady parametrů β. Naše regresní rovnice je $Y = 26{,}792 + 26{,}843X$ s indexem determinace $R^2 = 0{,}7859$. Lze tedy říci, že 78,59 % variability ceny je vysvětlena pomocí kapacity a této rovnice.

1.c) Vhodnost modelu lze také dokázat pomocí analýzy rozptylu. Hypotézy:

H_0: $\beta_0 = \beta_1 = 0$ ..... model není vhodný,

H_A: non H_0 ..... model je vhodný,

hodnotu testového kritéria F i p-value nalezneme v horní části obrázku (Obrázek 26):

$F(1, 9) = 33{,}038,$

$p = 0{,}00028.$

Při zvolení typické hladiny významnosti 5 % se nám podařilo zamítnout nulovou hypotézu ve prospěch alternativní hypotézy (jelikož $p < \alpha$), tudíž jsme prokázali významnost našeho modelu.

1.d) Dále pomocí t-testu můžeme otestovat důležitost jednotlivých regresních koeficientů. Hypotézy pro absolutní člen:

H_0: $\beta_0 = 0$ ..... absolutní člen není důležitý,

H_A: $\beta_0 \neq 0$ ..... absolutní člen je důležitý,

hodnota testového kritéria a p-value (nalezneme v prvním řádku v pravé části tabulky):

$t = 0{,}088$,

$p = 0{,}9322$.

Při zvolené hladině významnosti 5 % se nám nepodařilo zamítnout nulovou hypotézu ve prospěch alternativní hypotézy (jelikož $p > \alpha$), tudíž jsme neprokázali významnost absolutního členu.

Hypotézy pro lineární člen:

H_0: $\beta_1 = 0$ ..... lineární člen není důležitý,

H_A: $\beta_1 \neq 0$ ..... lineární člen je důležitý,

hodnota testového kritéria a p-value (opět nalezneme v tabulce, ve druhém řádku):

$t = 5{,}748$,

$p = 0{,}00028$.

Při zvolené hladině významnosti 5 % se nám podařilo zamítnout nulovou hypotézu ve prospěch alternativní hypotézy (jelikož $p < \alpha$), tudíž jsme prokázali významnost lineárního členu.

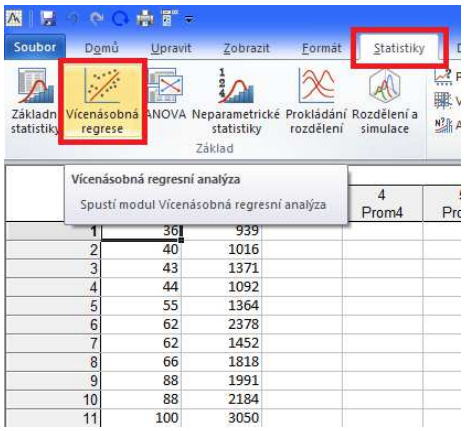
Za zmínku stojí, že p-value lineárního členu je rovna p-value modelu (což platí pouze v případě jednoduché lineární regrese), viz Obrázek 27.

Obrázek 27: STATISTICA – Porovnání p-value

N=11	Výsledky regrese se závislou proměnnou : Cena (Tabulka1) R= ,88651350 R2= ,78590618 Upravené R2= ,76211798 F(1,9)=33,038 p<,00028 Směrod. chyba odhadu : 320,35					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			26,79182	306,0330	0,087546	0,932155
Kapacita	0,886513	0,154234	26,84253	4,6700	5,747838	0,000277

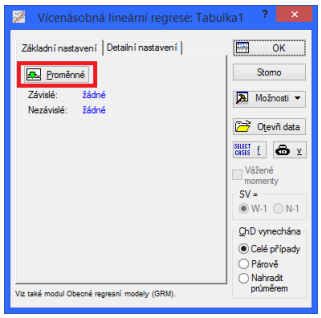
ad 1.b) Druhým způsobem výpočtu jednoduché lineární regrese v softwaru STATISTICA je způsob přes vícenásobnou regresi. Ukážeme si tento postup od korelačního pole, viz Graf 7. V menu „Statistiky“ zvolíme v levé části nabídky „Vícenásobná regrese“, viz Obrázek 28.

Obrázek 28: STATISTICA – Menu – Vícenásobná regrese

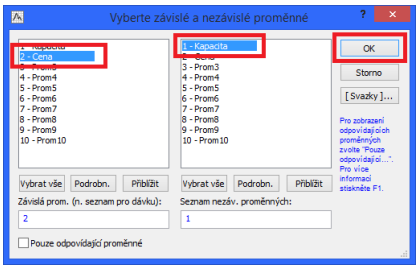


Klikneme na tlačítko pro „Proměnné“, viz Obrázek 29. Jako závislou proměnnou zvolíme Cenu a jako nezávislou proměnnou zvolíme Kapacitu, obojí potvrdíme tlačítkem „OK“, viz Obrázek 30. Pokračujeme stisknutím tlačítka „OK“.

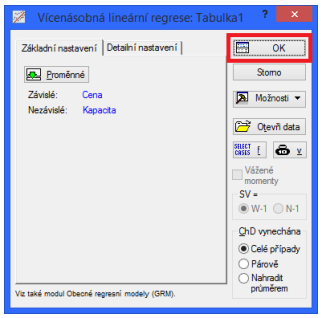
Obrázek 29: STATISTICA – Proměnné



Obrázek 30: STATISTICA – Výběr proměnných



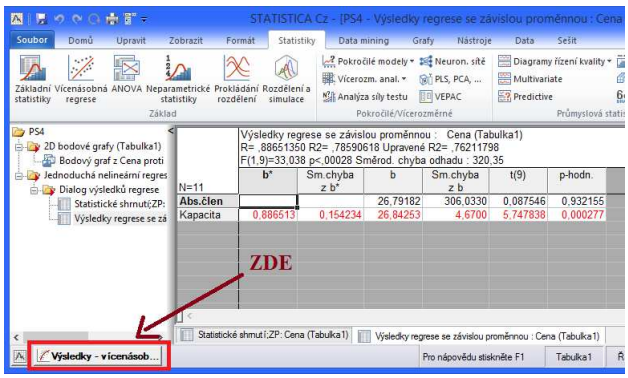
Obrázek 31: STATISTICA – Potvrzení



Zobrazí se identické okno s předběžnými výsledky, jako v předchozím postupu, viz Obrázek 25. Dále pokračujeme stejným způsobem.

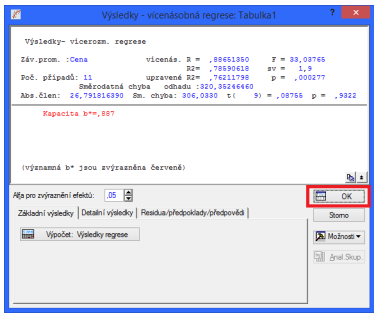
1.e) Nyní ověříme předpoklady užití lineárního regresního modelu. Vrátíme se zpět do dialogového okna s lineární regresí kliknutím na tlačítko s popiskem „Výsledky – vícenásobná regrese“ v levém dolním rohu, viz Obrázek 32.

Obrázek 32: STATISTICA – Návrat do dialogového okna

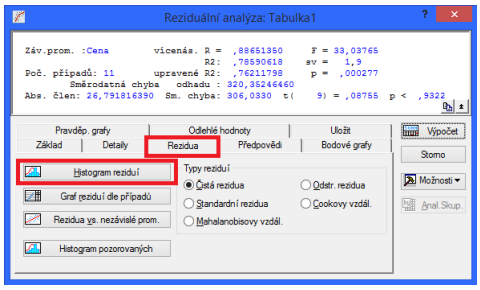


Stisknutím tlačítka „OK“ se přesuneme do reziduální analýzy, viz Obrázek 33. Zvolíme záložku „Rezidua“ a klikneme na „Histogram reziduí“, viz Obrázek 34.

Obrázek 33: STATISTICA – Výsledky

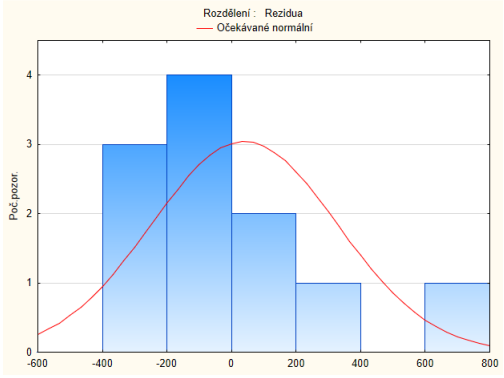


Obrázek 34: STATISTICA – Reziduální analýza



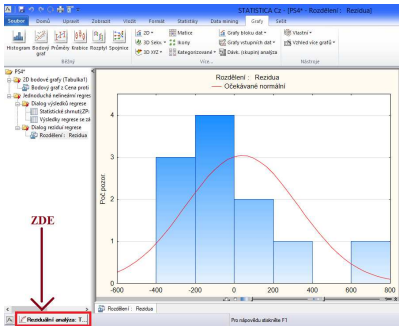
Z grafu se zdá, že rozložení reziduí odpovídá normálnímu rozdělení se střední hodnotou nula, viz Graf 8.

Graf 8: STATISTICA – Histogram reziduí

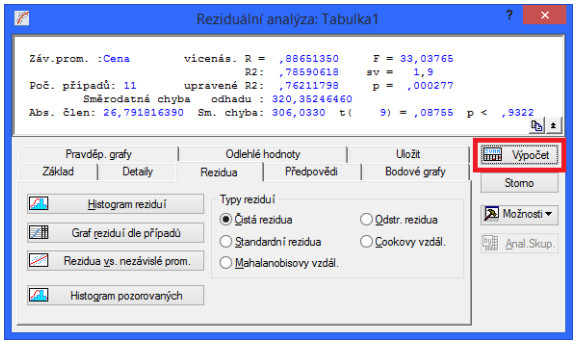


Předpoklad normality reziduí ověříme technikou testování hypotéz pomocí Shapiro-Wilkova testu. Nejprve ovšem musíme získat hodnoty reziduí. Vrátíme se zpět do reziduální analýzy (viz Obrázek 35) a klikneme na „Výpočet“ (viz Obrázek 36).

Obrázek 35: STATISTICA – Návrat do reziduální analýzy



Obrázek 36: STATISTICA – Reziduální analýza



Získané hodnoty reziduí (viz Obrázek 37) zkopírujeme do tabulky s původními daty a proměnnou pojmenujeme jako „Rezidua“ (viz Obrázek 38).

Obrázek 37: STATISTICA – Výpočet reziduí

Předpovědi a rezidua (Tabulka1)
Závislá proměnná : Cena

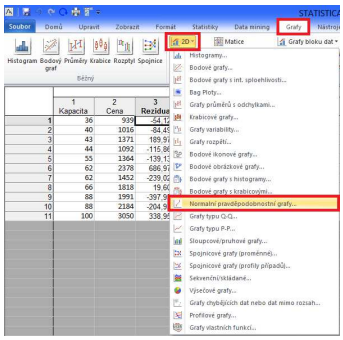
Případ	Pozorovaná hodnota	Předpovězená hodnota	Reziduum	Stand. předpov.	Stand. Rezid.
1	939,000	993,123	-54,123	-1,20695	-0,16895
2	1016,000	1100,493	-84,493	-1,02256	-0,26375
3	1371,000	1181,021	189,979	-0,88426	0,59303
4	1092,000	1207,863	-115,863	-0,83816	-0,36167
5	1364,000	1503,131	-139,131	-0,33107	-0,43431
6	2378,000	1691,029	686,971	-0,00838	2,14442
7	1452,000	1691,029	-239,029	-0,00838	-0,74614
8	1818,000	1798,399	19,601	0,17601	0,06119
9	1991,000	2388,934	-397,934	1,19019	-1,24218
10	2184,000	2388,934	-204,934	1,19019	-0,63972
11	3050,000	2711,045	338,955	1,74338	1,05807
Minimum	939,000	993,123	-397,934	-1,20695	-1,24218
Maximum	3050,000	2711,045	686,971	1,74338	2,14442
Průměr	1695,909	1695,909	-0,000	0,00000	-0,00000
Medián	1452,000	1691,029	-84,493	-0,00838	-0,26375

Obrázek 38: STATISTICA – Rezidua

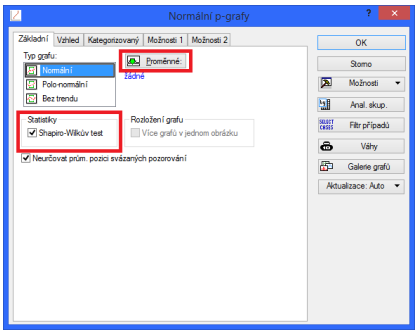
	1 Kapacita	2 Cena	3 Rezidua
1	36	939	-54,123
2	40	1016	-84,493
3	43	1371	189,979
4	44	1092	-115,863
5	55	1364	-139,131
6	62	2378	686,971
7	62	1452	-239,029
8	66	1818	19,601
9	88	1991	-397,934
10	88	2184	-204,934
11	100	3050	338,955

Nyní v menu zvolíme kategorii „Grafy“, vybereme položku „2D“ a zde vybereme „Normální pravděpodobnostní grafy…“, viz Obrázek 39. V nastavení zaškrtneme provedení Shapiro-Wilkovo testu a klikneme na proměnné, viz Obrázek 40.

Obrázek 39: STATISTICA – Grafy

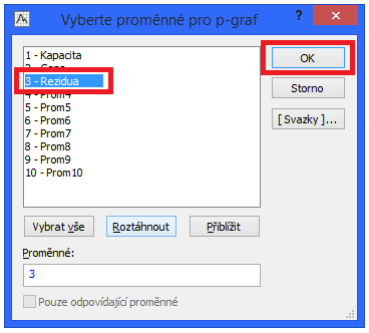


Obrázek 40: STATISTICA – Shapiro-Wilkův test

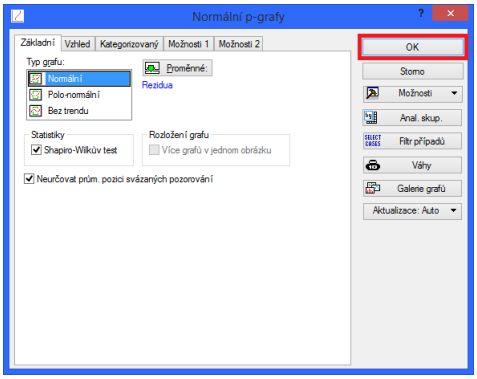


Jako proměnné označíme „Rezidua“ a výběr potvrdíme tlačítkem „OK“, viz Obrázek 41. Vytvoření grafu provedeme stisknutím tlačítka „OK“, viz Obrázek 42.

Obrázek 41: STATISTICA – Výběr proměnných

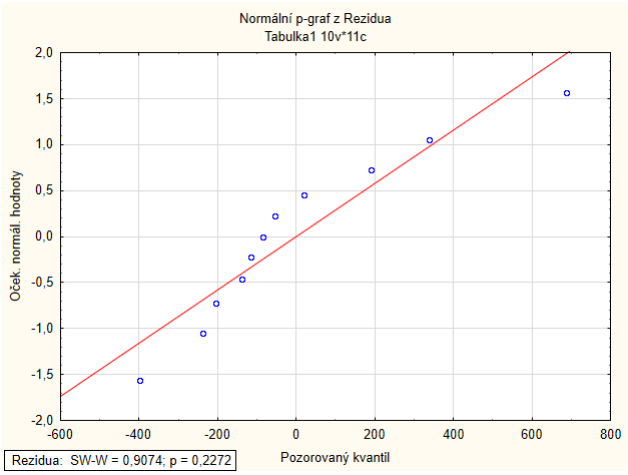


Obrázek 42: STATISTICA – Potvrzení



Výsledný pravděpodobnostní graf reziduí můžeme vidět v grafu Graf 9. Rovněž zde můžeme vidět hodnotu testové statistiky SW-W = 0,9074 a hodnotu p-value p = 0,2272.

Graf 9: STATISTICA – Shapiro-Wilkův test



Při stanovených hypotézách:

H_0: rezidua sledují normální rozdělení,

H_A: rezidua nesledují normální rozdělení

a hladině významnosti 5 % bychom nezamítli nulovou hypotézu ve prospěch alternativní hypotézy (jelikož $p > \alpha$), tedy lze říci, že rozdělení reziduí přibližně odpovídá normálnímu rozdělení.

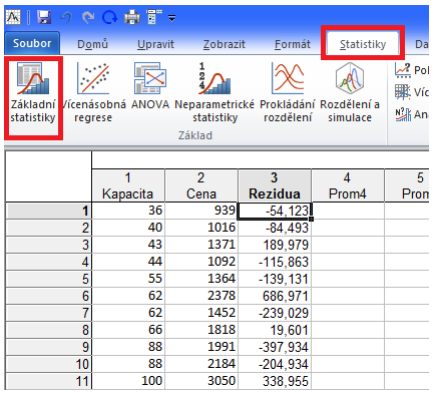
Dále ověříme hypotézu, že střední hodnota reziduí odpovídá nule, pomocí jednovýběrového t-testu:

H_0: $\mu = 0$ ..... střední hodnota je rovna nule,

H_A: $\mu \neq 0$ ..... střední hodnota není rovna nule,

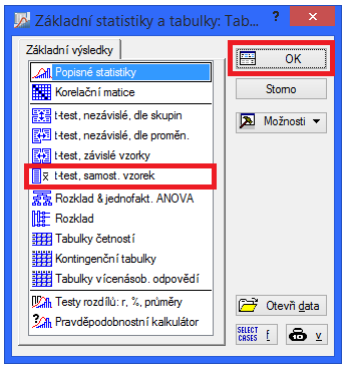
testové kritérium a p-value získáme z programu STATISTICA. V menu zvolíme kategorii „Statistiky“ a v levé části klikneme na „Základní statistiky“, viz Obrázek 43. Z nabídky vybereme „t-test, samost. vzorek“ a výběr potvrdíme tlačítkem „OK“, viz Obrázek 44.

Obrázek 43: STATISTICA – Základní statistiky



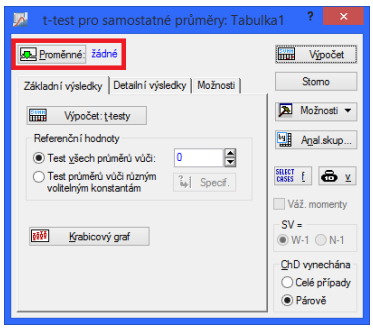
	1 Kapacita	2 Cena	3 Rezidua	4 Prom4	5 Pron
1	36	939	-54,123		
2	40	1016	-84,493		
3	43	1371	189,979		
4	44	1092	-115,863		
5	55	1364	-139,131		
6	62	2378	686,971		
7	62	1452	-239,029		
8	66	1818	19,601		
9	88	1991	-397,934		
10	88	2184	-204,934		
11	100	3050	338,955		

Obrázek 44: STATISTICA – T-test

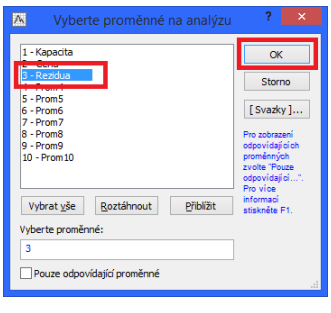


V nastavení nejprve klikneme na „Proměnné“, viz Obrázek 45. Jako proměnné zvolíme opět pouze „Rezidua“ a výběr potvrdíme tlačítkem „OK“, viz Obrázek 46. Rezidua chceme porovnat s hodnotou nula, proto „Test všech průměrů vůči:“ necháme na nule, viz Obrázek 47. Výsledky testu získáme kliknutím na tlačítko „Výpočet.

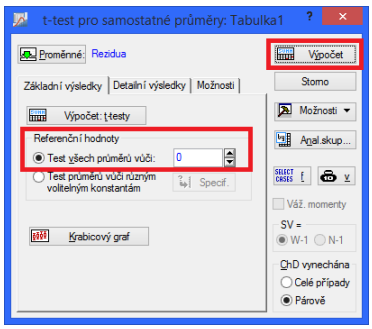
Obrázek 45: STATISTICA – Proměnné



Obrázek 46: STATISTICA – Výběr proměnných



Obrázek 47: STATISTICA – Výpočet t-testu



Výsledky t-testu vyšly jednoznačně, viz Obrázek 48. Testové kritérium je prakticky nulové a tím pádem je p-value přibližně rovna jedné.

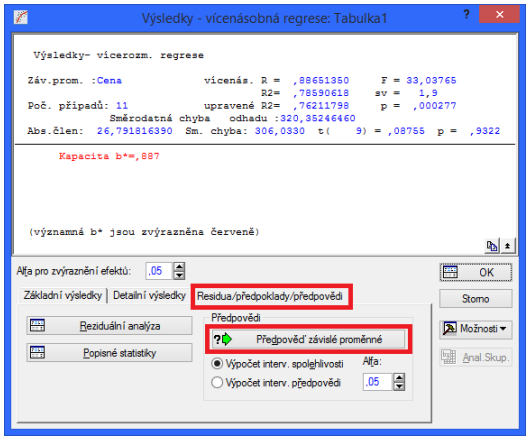
Obrázek 48: STATISTICA – Výsledky t-testu

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka1)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000018	303,9130	11	91,63322	0,00	-0,000000	10	1,000000

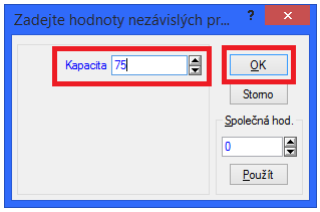
Při zvolené hladině významnosti 5 % nemůžeme zamítnout nulovou hypotézu ve prospěch alternativní hypotézy, naše hypotéza o nulové střední hodnotě reziduí byla potvrzena.

2) V poslední části provedeme výpočet bodové předpovědi pro hodnotu kapacity 75 Ah. Vrátíme se do okna s výsledky lineární regrese, viz Obrázek 32. Zvolíme záložku „Residua/předpoklady/předpovědi“ a klikneme na tlačítko „Předpověď závislé proměnné“, viz Obrázek 49. Vyplníme hodnotu kapacity na 75 a potvrdíme výpočet tlačítkem „OK“, viz Obrázek 50.

Obrázek 49: STATISTICA – Předpověď



Obrázek 50: STATISTICA – Zadání hodnoty



Získáme tabulku, ve které nalezneme předpověď ceny baterie pro kapacitu 75 Ah dle námi nalezeného lineárního modelu, která je 2 039,98 Kč, viz Obrázek 51.

Obrázek 51: STATISTICA – Výsledek předpovědi

Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Cena		
	b-váha	Hodnota	b-váha * Hodnot
Kapacita	26,84253	75,00000	2013,190
Abs. člen			26,792
Předpověď			2039,982
-95,0%LS			1782,921
+95,0%LS			2297,042

Tento odhad můžeme provést i prostým dosazením do nalezené regresní rovnice:

$$Y = 26{,}792 + 26{,}843 \cdot 75 = 2040{,}017,$$

kdybychom zpřesnili lineární regresní koeficient na více desetinných míst, výsledek bude více odpovídat hodnotě ze STATISTICY:

$$Y = 26{,}792 + 26{,}84253 \cdot 75 = 2039{,}98175.$$

3) Poslední otázka směřovala k určení změny ceny při zvýšení kapacity o 5 Ah. Opět využijeme regresní rovnici a jen za proměnnou X dosadíme $(X + 5)$ a rovnici upravíme:

$$Y = 26{,}792 + 26{,}843(X + 5),$$

$$Y = 26{,}792 + 26{,}843X + 134{,}215,$$

porovnáním získané rovnice s původní rovnicí zjistíme, k jaké změně došlo:

Původní rovnice: $Y = \boxed{26{,}792 + 26{,}843X}$,

Upravená rovnice: $Y = \boxed{26{,}792 + 26{,}843X} + 134{,}215.$

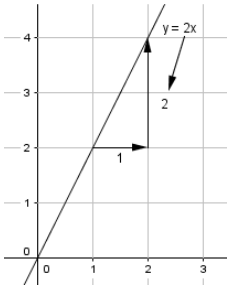
Zvýrazněná část se shoduje. Ve druhé rovnici je navíc 134,215, došlo tedy k navýšení ceny (závislé proměnné) o 134,215 Kč. Jinými slovy každých 5 Ah kapacity baterie navíc by nás cenově přišlo přibližně o 134,215 Kč dráž.

K tomuto číslu lze také snadno dojít prostým vynásobením hodnoty vyjadřující změnu a lineárního regresního koeficientu: $\beta_1 k$, kde k určuje onu změnu:

$$26{,}843 \cdot 5 = 134{,}215.$$

Protože změna mezi hodnotami Y je v podstatě dána lineárním členem a změnou hodnoty X. Lineární člen udává sklon přímky, tj. určuje, o kolik se změní hodnota Y při navýšení hodnoty X o 1, viz Graf 10.

Graf 10: Příklad směrnice přímky



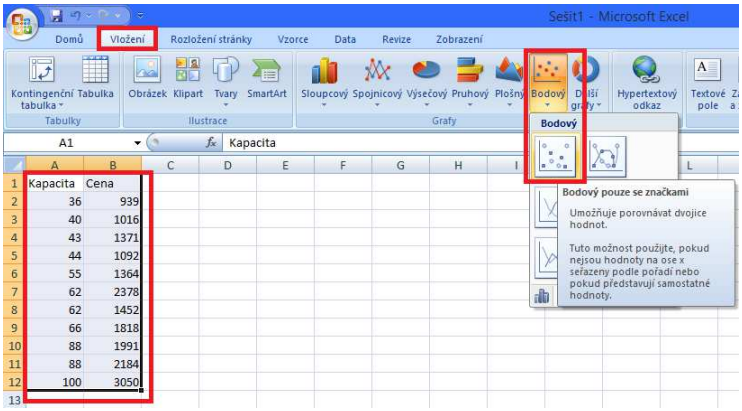
4.2.3 Řešení v softwaru Excel

1.a) Stejně jako ve STATISTICE si nejprve připravíme data a rovněž je třeba, aby proměnné byly ve sloupcích, viz Obrázek 52. Poté vytvoříme bodový graf, který bude reprezentovat námi hledané korelační pole. Označíme si data (včetně popisků, ale není to podmínkou), přepneme se do záložky „Vložení“ a vybereme „Bodový graf“, konkrétně „Bodový pouze se značkami“, viz Obrázek 53.

Obrázek 52: Excel – Data

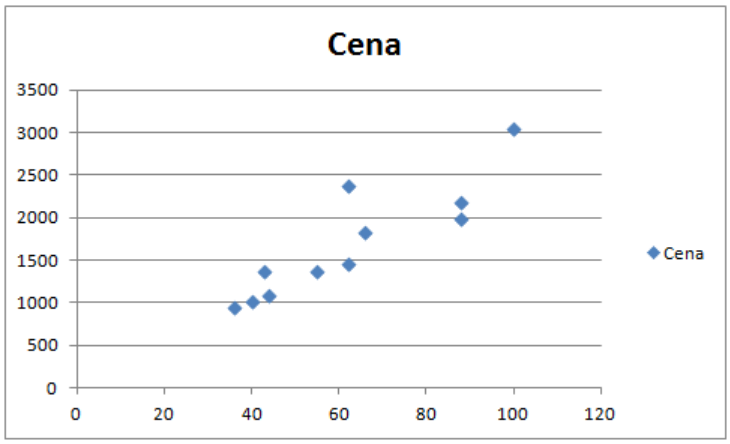
	A	B	C
1	Kapacita	Cena	
2	36	939	
3	40	1016	
4	43	1371	
5	44	1092	
6	55	1364	
7	62	2378	
8	62	1452	
9	66	1818	
10	88	1991	
11	88	2184	
12	100	3050	
13			

Obrázek 53: Excel – Bodový graf, postup



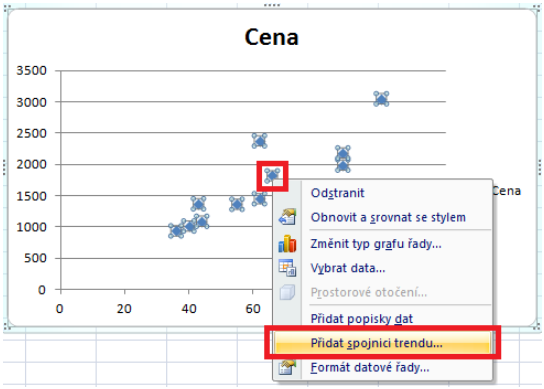
Opět v grafu vidíme náznak lineární závislosti (viz Graf 11), použijeme model $Y = \beta_0 + \beta_1 X$ a určíme koeficient determinace.

Graf 11: Excel – Korelační pole

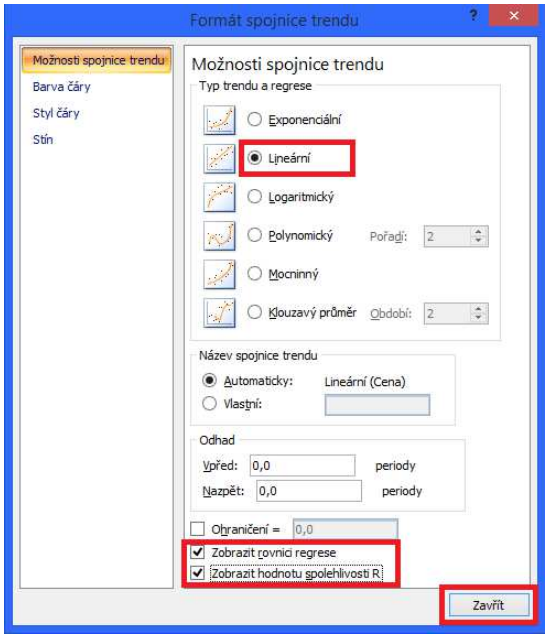


1.b) Regresní rovnici spolu s indexem determinace můžeme získat přímo z grafu. Pravým myšítkem klikneme na libovolný bod v grafu a zvolíme možnost „Přidat spojnici trendu“, viz Obrázek 54. V dialogovém okně zvolíme lineární trend a v dolní části zaškrtneme „Zobrazit rovnici regrese“ a „Zobrazit hodnotu spolehlivosti R“ (viz Obrázek 55), tlačítkem „Zavřít“ skryjeme dialogové okno.

Obrázek 54: Excel – Přidání spojnice trendu

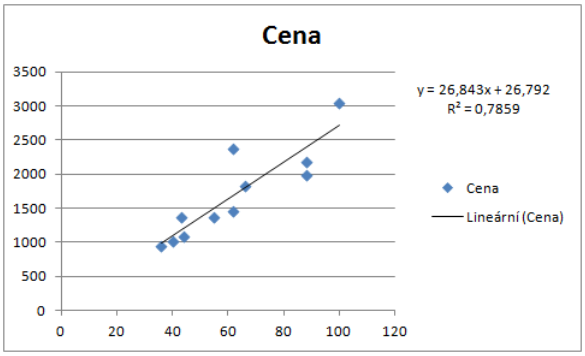


Obrázek 55: Excel – Formát spojnice trendu



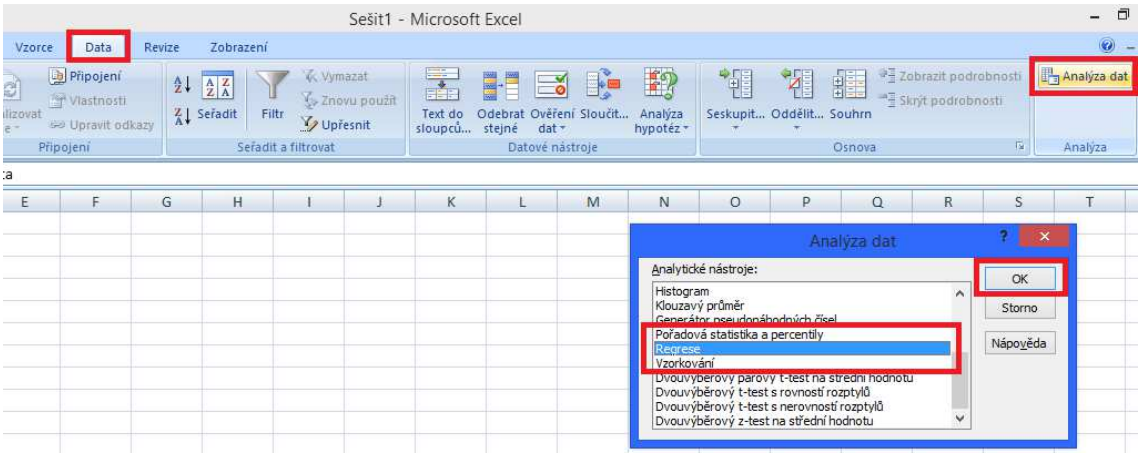
Regresní rovnice $Y = 26{,}843X + 26{,}792$ je srovnatelná s rovnicí z programu STATISTICA, jen parametry jsou v opačném pořadí, viz Graf 12. Index determinace je $R^2 = 0{,}7859$.

Graf 12: Excel – Model a index determinace



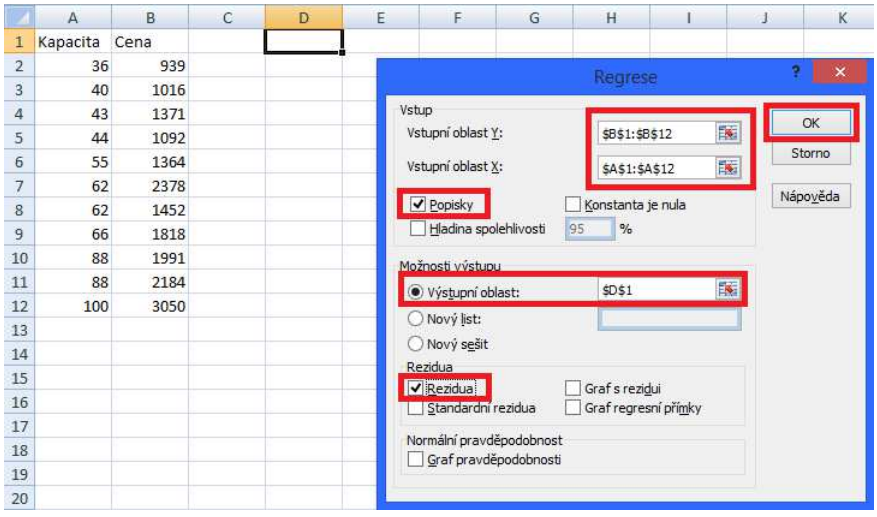
Nyní si v programu Excel ukážeme hlubší analýzu. V menu zvolíme položku „Data“ a v pravé části klikneme na „Analýza dat“, viz Obrázek 56. V seznamu vybereme „Regrese“ a výběr potvrdíme tlačítkem „OK“.

Obrázek 56: Excel – Analýza dat



V dialogovém okně vybereme proměnnou *Y* a *X* (pokud vyberete data jako já i s popisky, je třeba zaškrtnout položku „Popisky“), v možnostech pro výstup zvolíme výstupní oblast (abychom měli výsledky na stejném listě spolu s daty a grafem), zároveň zaškrtneme, že chceme vidět „Rezidua“, viz Obrázek 57. Vše potvrdíme tlačítkem „OK“.

Obrázek 57: Excel – Regrese

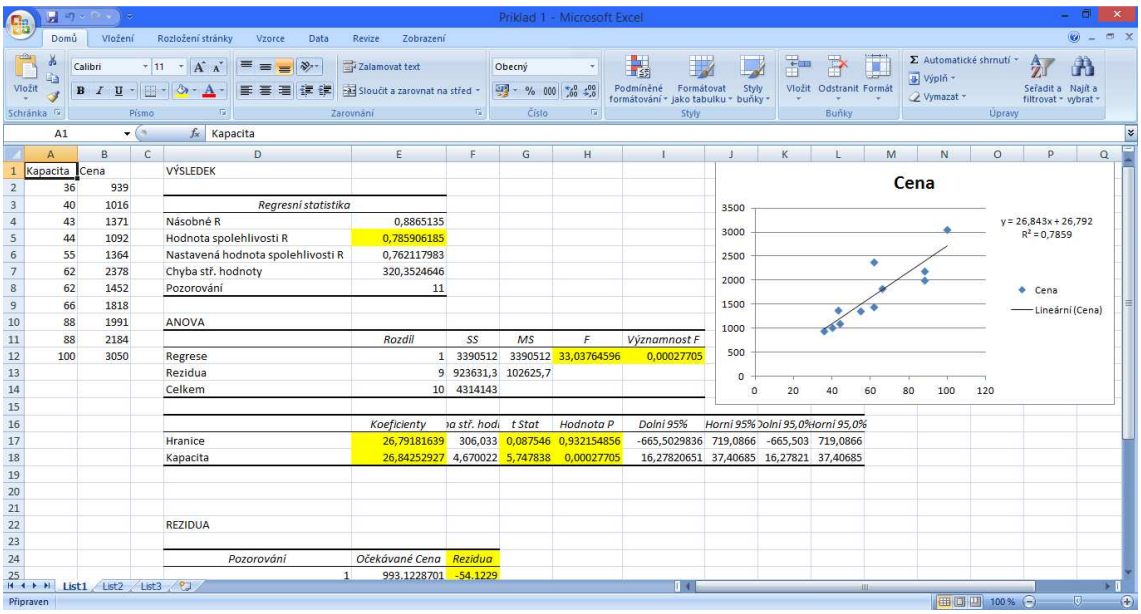


Výsledky regrese můžeme vidět níže, viz Obrázek 58. Žlutě jsem zvýraznil buňky, které nás zajímají především. V první tabulce se jedná o index determinace alias „Hodnota spolehlivosti R“.

1.c) Ve druhé tabulce, jak již nadpis napovídá, se jedná o test analýzy rozptylu (ANOVA), zvýrazněna je hodnota testového kritéria F a p-value (Významnost F).

1.d) Ve třetí tabulce nalezneme odhady regresních koeficientů β, pod názvem „Koeficienty“ a vedle nich příslušné hodnoty testového kritéria (t Stat) a p-value (Hodnota P) dílčích t-testů zkoumajících významnost těchto koeficientů.

Obrázek 58: Excel – Výsledky regrese



Kapacita	Cena
36	939
40	1016
43	1371
44	1092
55	1364
62	2378
62	1452
66	1818
88	1991
88	2184
100	3050

VÝSLEDEK

Regresní statistika

Násobné R	0,8865135
Hodnota spolehlivosti R	0,785906185
Nastavená hodnota spolehlivosti R	0,762117983
Chyba stř. hodnoty	320,3524646
Pozorování	11

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	3390512	3390512	33,03764596	0,00027705
Rezidua	9	923631,3	102625,7		
Celkem	10	4314143			

	Koeficienty	*…a stř. hodn*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
Hranice	26,79181639	306,033	0,087546	0,932154856	-665,5029836	719,0866	-665,503	719,0866
Kapacita	26,84252927	4,670022	5,747838	0,00027705	16,27820651	37,40685	16,27821	37,40685

REZIDUA

Pozorování	*Očekávané Cena*	*Rezidua*
1	993,1228701	-54,1229

1.e) Testy normality nejsou součástí doplňku Analýzy dat a tak nelze jednoduchým způsobem ověřit normalitu reziduí (pokud nechceme provádět výpočty ručně).

2), 3) Excel nenabízí výpočet bodové předpovědi pro daný regresní model a tak bychom sami museli dosadit zadanou hodnotu vysvětlující proměnné do nalezené regresní rovnice. Tento postup je popsán v předchozí kapitole v řešení pro program STATISTICA, viz str. 40.

4.3 Příklad mnohonásobné lineární regrese

Tato kapitola je věnována řešení vzorového příkladu z mnohonásobné lineární regrese. Postupy si opět ukážeme v programech STATISTICA a Excel.

4.3.1 Vzorový příklad

Ve druhém příkladě máme zadané údaje o tržbách stravovacích úseků (X_1, v mil. Kč) ve dvanácti hotelech určitého řetězce, počet „lůžkonocí“ (X_2, jedná se o měsíční kapacitu hotelů danou součinem celkového počtu lůžek a počtu dnů v měsíci) a celkové měsíční tržby (Y, v mil. Kč) hotelů [6].

Tabulka 6: Data k vzorovému příkladu

i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	12	8	76,4	17	21,3	10	12,5	97,3	88	25	38,6	47,3
x_{1i}	2	1,2	14,8	8,3	8,4	3	4,8	15,6	16,1	11,5	14,2	14
x_{2i}	150	94	811	254	399	95	149	312	952	247	400	312

Zdroj: Hindls [6], str. 218

Zadání

1) Nalezněte model vícenásobné lineární regrese, ve kterém bude celková měsíční tržba závislá na tržbách stravovacího úseku a počtu „lůžkonocí“.
2) Nalezněte bodovou předpověď celkové měsíční tržby pro 200 lůžkonocí a při 7,5 mil. Kč tržby stravovacího úseku.
3) Pokud to bude vhodné, model zjednodušte a porovnejte výsledky.

Řešení

1. a) Nalezneme lineární regresní rovnici a určíme koeficient determinace.
   b) Provedeme test analýzy rozptylu k určení spolehlivosti modelu.
   c) Provedeme dílčí t-testy k určení důležitosti jednotlivých regresních koeficientů a vysvětlujících proměnných.
   d) Ověříme předpoklady užití nalezeného regresního modelu.
2. Nalezneme bodovou předpověď celkové měsíční tržby pro zadané hodnoty vysvětlujících proměnných.

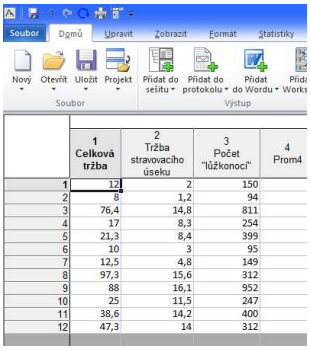
3. a) Určíme, zda je vhodné některou proměnnou vypustit a nalezneme zjednodušený model.

3. b) Porovnáme nalezené modely, jejich spolehlivost a bodovou předpověď.

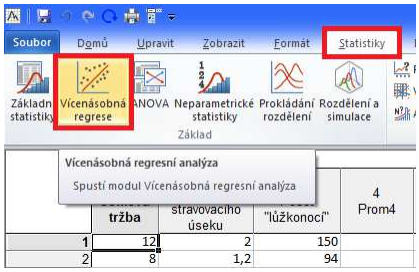
4.3.2 Řešení v softwaru STATISTICA

1.a) Data přepíšeme do STATISTICY a jednotlivé proměnné pojmenujeme dle jejich specifikace, viz Obrázek 59. K získání modelu $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ provedeme vícenásobnou lineární regresi, v menu zvolíme položku „Statistiky“ a v levé části klikneme na „Vícenásobná regrese“, viz Obrázek 60.

Obrázek 59: STATISTICA – Data

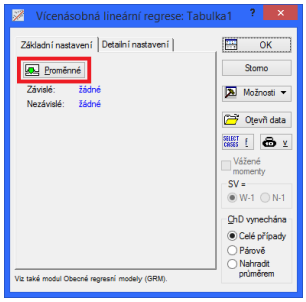


Obrázek 60: STATISTICA – Vícenásobná regrese

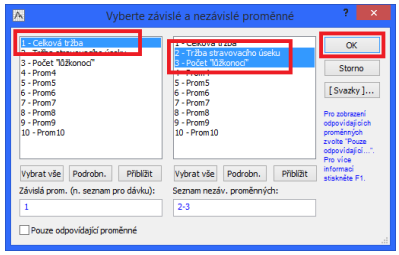


V následujícím okně zvolíme proměnné, které budou zahrnuty do výpočtu. Klikneme na „Proměnné“, viz Obrázek 61. Jako závislou proměnnou zvolíme Celkovou tržbu a jako nezávislé proměnné Tržbu stravovacího úseku a Počet „lůžkonocí“, výběr uložíme tlačítkem „OK“, viz Obrázek 62. Nastavení potvrdíme stisknutím tlačítka „OK“, viz Obrázek 63.

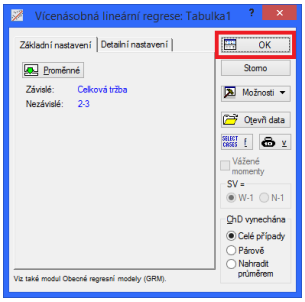
Obrázek 61: STATISTICA – Proměnné



Obrázek 62: STATISTICA – Výběr proměnných

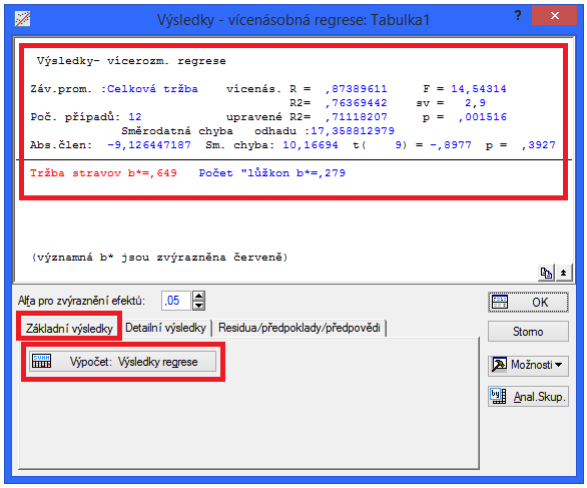


Obrázek 63: STATISTICA – Potvrzení



Nyní se dostáváme k předběžným výsledkům, viz Obrázek 64 horní část okna. V dolní části zvolíme záložku „Základní výsledky“ a stiskneme tlačítko „Výpočet: Výsledky regrese“.

Obrázek 64: STATISTICA – Předběžné výsledky



Obrázek 65: STATISTICA – Výsledky

Výsledky regrese se závislou proměnnou : Celková tržba (Tabulka1)
R= ,87389611 R2= ,76369442 Upravené R2= ,71118207
F(2,9)=14,543 p<,00152 Směrod. chyba odhadu : 17,359

N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			-9,12645	10,16694	-0,897659	0,392747
Tržba stravovacího úseku	0,648770	0,238435	3,72927	1,37058	2,720954	0,023571
Počet "lůžkonocí"	0,278590	0,238435	0,03309	0,02832	1,168411	0,272658

Ve sloupci „b“ opět nalezneme odhady regresních koeficientů β. Nyní můžeme vyjádřit regresní rovnici $Y = -9{,}126 + 3{,}729X_1 + 0{,}033X_2$, její index spolehlivosti je 0,7637.

1.b) Hypotézy vztahující se k vhodnosti modelu jsou:

H_0: $\beta_0 = \beta_1 = \beta_2 = 0$ ..... model není vhodný,

H_A: non H_0 ..... model je vhodný.

Testové kritérium analýzy rozptylu F a p-value vidíme v horní části obrázku, viz Obrázek 65:

$F(2, 9) = 14{,}543,$

$p = 0{,}00152.$

Na hladině významnosti 5 % zamítneme nulovou hypotézu ve prospěch alternativní hypotézy (jelikož $p < \alpha$), model je jako celek významný.

1.c) Na řadě jsou dílčí t-testy ověřující důležitost jednotlivých regresních koeficientů. Hypotézy pro absolutní člen:

H_0: $\beta_0 = 0$ ..... absolutní člen není důležitý,

H_A: $\beta_0 \neq 0$ ..... absolutní člen je důležitý,

testové kritérium i p-value t-testu nalezneme v prvním řádku, viz Obrázek 65:

$t = -0{,}898,$

$p = 0{,}3927.$

Na hladině významnosti 5 % je $p > \alpha$, čímž nezamítneme nulovou hypotézu ve prospěch alternativní hypotézy, neprokázali jsme významnost absolutního členu.

Hypotézy vztahující se k proměnné X_1 (tržby stravovacích úseků):

H_0: $\beta_1 = 0$ ..... lineární člen proměnné X_1 není důležitý,

H_A: $\beta_1 \neq 0$ ..... lineární člen proměnné X_1 je důležitý,

testové kritérium a p-value (viz druhý řádek s výsledky):

$t = 2{,}721,$

$p = 0{,}0236.$

Na hladině významnosti 5 % je $p < \alpha$, čímž zamítneme nulovou hypotézu ve prospěch alternativní hypotézy, prokázali jsme významnost proměnné X_1.

Hypotézy vztahující se k proměnné X_2 (počet lůžkonocí):

H_0: $\beta_2 = 0$ ..... lineární člen proměnné X_2 není důležitý,

H_A: $\beta_2 \neq 0$ ..... lineární člen proměnné X_2 je důležitý,

testové kritérium a p-value (viz třetí řádek s výpočty):

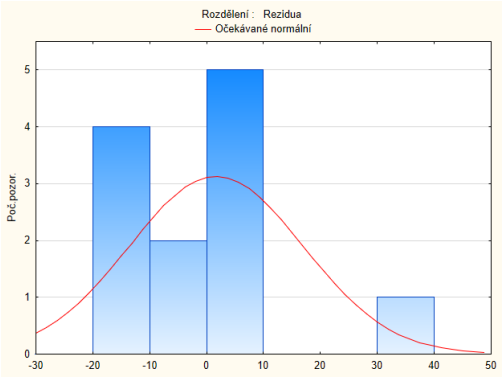
$t = 1{,}168,$

$p = 0{,}2727.$

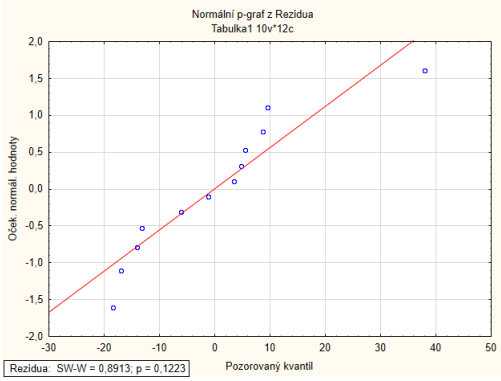
Na hladině významnosti 5 % je $p > \alpha$, čímž nezamítneme nulovou hypotézu ve prospěch alternativní hypotézy, neprokázali jsme významnost proměnné X_2.

1.d) Nyní ověříme předpoklady užití nalezeného lineárního regresního modelu. Nejprve opět vytvoříme histogram reziduí k odhadu platnosti předpokladů, viz Graf 13. Střední hodnota se zdá, že opravdu bude odpovídat nule, ovšem platnost normálního rozdělení reziduí až tak jistá není. Provedeme tedy Shapiro-Wilkův test na ověření normality, viz Graf 14. Jelikož $p = 0{,}1223$ a $p > 0{,}05$, nezamítneme hypotézu o normalitě reziduí.

Graf 13: STATISTICA – Histogram reziduí



Graf 14: STATISTICA – Shapiro-Wilkův test



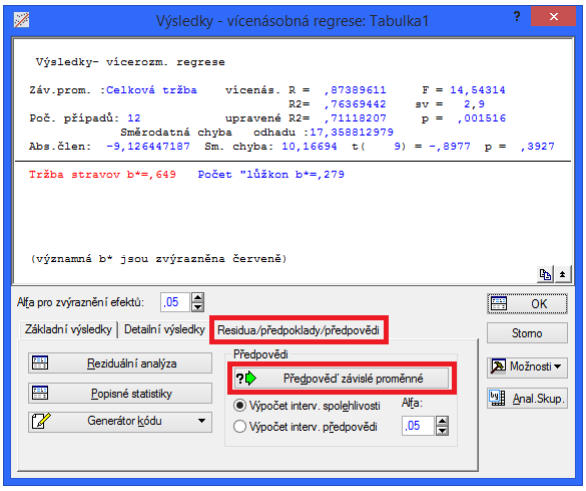
Výsledky t-testu o střední hodnotě reziduí (viz Obrázek 66) se shodují s naším odhadem, že střední hodnota je skutečně blízká nule ($p \doteq 1$, nezamítáme nulovou hypotézu).

Obrázek 66: STATISTICA – Výsledky t-testu

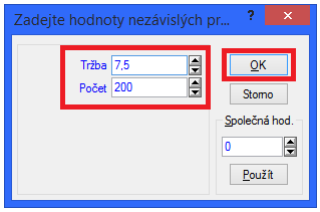
	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka1)							
Proměnná	**Průměr**	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000000	15,70164	12	4,532673	0,00	-0,000000	11	1,000000

2) Dále provedeme bodovou předpověď celkové měsíční tržby pro 200 lůžkonocí a při 7,5 mil. Kč tržby stravovacího úseku. Vrátíme se do dialogového okna s výsledky vícenásobné regrese (analogicky viz Obrázek 32). Přepneme se do záložky „Rezidua/předpoklady/předpovědi“ a klikneme na tlačítko „Předpověď závislé proměnné“, viz Obrázek 67. Vyplníme hodnoty vysvětlujících proměnných a potvrdíme je tlačítkem „OK“, viz Obrázek 68.

Obrázek 67: STATISTICA – Předpověď



Obrázek 68: STATISTICA – Zadání hodnot



Výsledek nalezneme v tabulce na obrázku (Obrázek 69). Předpověď celkové tržby je rovna 25,46 mil. Kč.

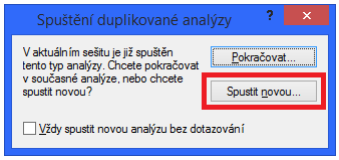
Obrázek 69: STATISTICA – Výsledek předpovědi

Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Celková tržba b-váha	Hodnota	b-váha * Hodnot
Tržba stravovacího úseku	3,729273	7,5000	27,96955
Počet "lůžkonocí"	0,033091	200,0000	6,61812
Abs. člen			-9,12645
Předpověď			25,46122
-95,0%LS			12,40062
+95,0%LS			38,52181

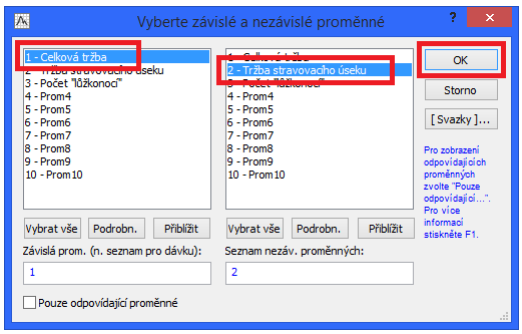
3.a) Na závěr bych se vrátil k výsledkům dílčích t-testů, určující vhodnost jednotlivých regresních koeficientů. Jelikož jsme neprokázali významnost proměnné X_2, můžeme zkusit model zjednodušit do tvaru $Y = \beta_0 + \beta_1 X_1$ a proměnnou X_2 spolu s jejím koeficientem vypustit. Znovu v menu zvolíme „Statistiky“ a v levé části „Vícenásobná regrese“, viz Obrázek 60. Protože jsme již tuto analýzu prováděli, STATISTICA se nás zeptá, zda chceme pokračovat v předchozí analýze, nebo zda chceme spustit novou analýzu, viz Obrázek 70. Zvolíme tedy možnost „Spustit

novou…“. Následně se zobrazí dialogové okno pro výběr proměnných, viz Obrázek 61, tentokrát ovšem zvolíme pouze jednu nezávislou proměnnou – Tržba stravovacího úseku, viz Obrázek 71.

Obrázek 70: STATISTICA – Spustit novou analýzu



Obrázek 71: STATISTICA – Výběr proměnných



Výsledky můžeme vidět na obrázku Obrázek 72. Regresní rovnice má tvar $Y = -8{,}764 + 4{,}904X_1$ a její index spolehlivosti je $0{,}7278$. Model jako celek je stále významný ($p = 0{,}0042$, $p < 0{,}05$), stejně tak lineární člen proměnné X_1 ($p = 0{,}0004$, $p < 0{,}05$).

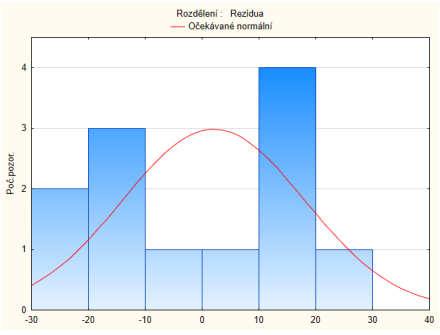
Obrázek 72: STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Celková tržba (Tabulka1)
R= ,85314118 R2= ,72784987 Upravené R2= ,70063486
F(1,10)=26,744 p<,00042 Směrod. chyba odhadu : 17,673

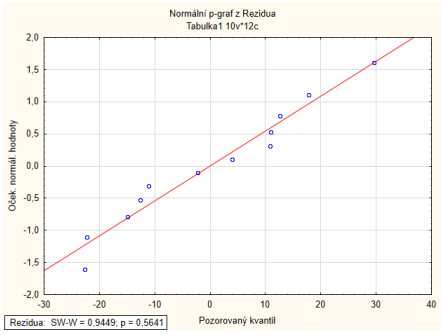
N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			-8,76424	10,34610	-0,847105	0,416749
Tržba stravovacího úseku	0,853141	0,164970	4,90405	0,94828	5,171501	0,000418

Předpoklady užití tohoto modelu jsou splněny. Dokládá to jak histogram reziduí, viz Graf 15, tak samotný Shapiro-Wilkův test normality ($p > 0{,}05$), viz Graf 16.

Graf 15: STATISTICA – Histogram reziduí 2



Graf 16: STATISTICA – Shapiro-Wilkův test 2



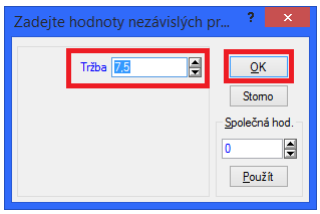
Stejně tak je splněn předpoklad nulové střední hodnoty reziduí, což potvrzuje provedený t-test, viz Obrázek 73.

Obrázek 73: STATISTICA – Výsledky t-testu 2

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka1)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	0,000001	16,85048	12	4,864314	0,00	0,000000	11	1,000000

Zkusme nyní určit bodovou předpověď v tomto zjednodušeném modelu. Vrátíme se do okna s výsledky regrese a zvolíme předpověď závislé proměnné (viz Obrázek 67). Do tržby dosadíme hodnotu 7,5, vyjadřující tržbu stravovacího úseku, viz Obrázek 74. Odhad celkové měsíční tržby je 28,016 mil. Kč, viz Obrázek 75.

Obrázek 74: STATISTICA – Zadání hodnoty



Obrázek 75: STATISTICA – Výsledek předpovědi

Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Celková tržba		
	b-váha	Hodnota	b-váha * Hodnot
Tržba stravovacího úseku	4,904046	7,500000	36,78035
Abs. člen			-8,76424
Předpověď			28,01611
-95,0%LS			15,89480
+95,0%LS			40,13742

3.b) Pokud bychom nyní porovnali výpočty na obou modelech, zjistíme, že se až tak výrazně neliší.

1. model: $Y = -9{,}126 + 3{,}729X_1 + 0{,}033X_2$,

2. model: $Y = -8{,}764 + 4{,}904X_1$.

Odhady koeficientů β_0 i β_1 jsou srovnatelné. Nízká důležitost proměnné X_2 napovídá i hodnota odhadu koeficientu β_2, která se téměř rovná nule (je ovšem samozřejmě třeba brát v potaz, v jakých řádech se hodnoty této proměnné pohybují).

Upravený koeficient determinace druhého modelu ($\bar{R}^2 = 0{,}7006$) je nižší jen o pouhé cca 1 % oproti původnímu modelu ($\bar{R}^2 = 0{,}7112$). Rozdíl není příliš velký ani v případě provedených bodových odhadů:

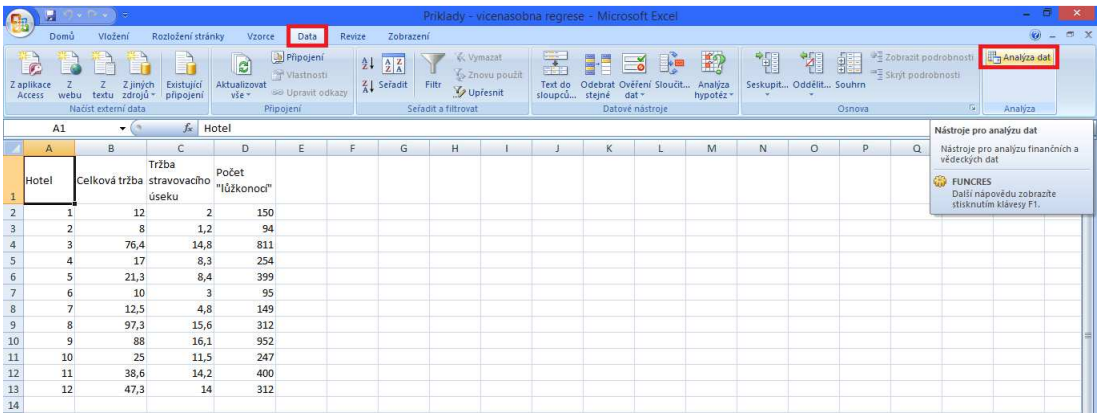
1. model: $\hat{y} = 25{,}46$,

2. model: $\hat{y} = 28{,}02$.

4.3.3 Řešení v softwaru Excel

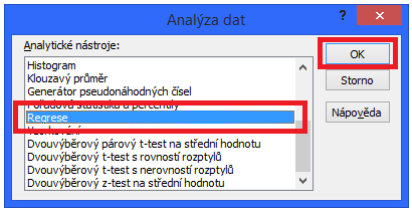
1.a) Nejprve si přepíšeme data (ideálně i s popisky v prvním řádku) a v menu „Data“ zvolíme v pravé části „Analýza dat“, viz Obrázek 76

Obrázek 76: Excel – Analýza dat

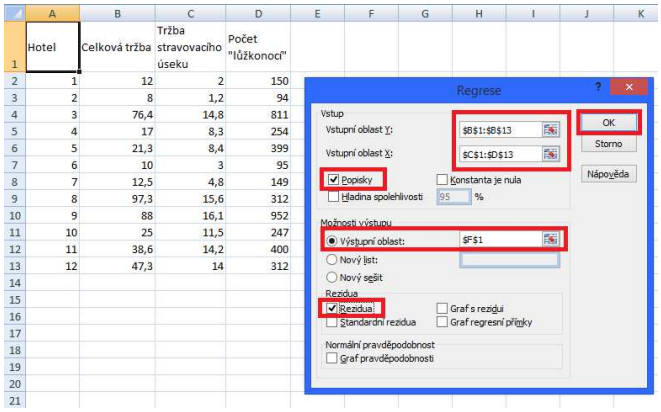


V seznamu Analýzy dat vybereme položku „Regrese“ a potvrdíme ji tlačítkem „OK“, viz Obrázek 77. Následně v novém dialogovém okně vybereme vstupní data, zatrhneme „Popisky“ (pokud jsme je zahrnuli ve vstupní oblasti), zvolíme buňku pro výstupní oblast a zatrhneme „Rezidua“, výpočet provedeme stisknutím tlačítka „OK“, viz Obrázek 78.

Obrázek 77: Excel – Regrese



Obrázek 78: Excel – Nastavení

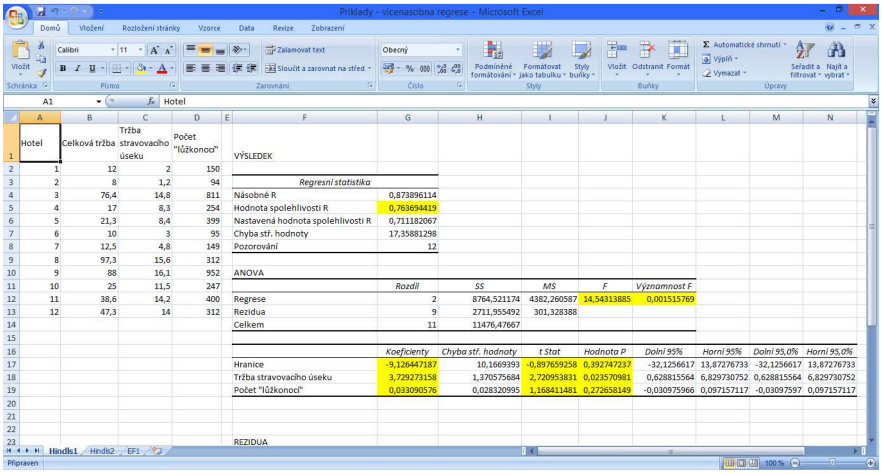


Výsledky se nám zobrazí do zvolené výstupní oblasti, viz Obrázek 79. Regresní koeficienty nalezneme ve třetí tabulce ve sloupci „Koeficienty“ z nichž sestavíme regresní rovnici $Y = -9{,}126 + 3{,}729X_1 + 0{,}033X_2$. Vhodnost modelu vyjadřuje koeficient determinace, který se nalézá v první tabulce pod názvem „Hodnota spolehlivosti R“, tedy $R^2 = 0{,}7637$.

1.b) Významnost modelu také prokazují výsledky analýzy rozptylu z druhé tabulky, kde $p = 0{,}00152$, což je méně než 5 %.

1.c) Důležitost jednotlivých regresních koeficientů určíme ze třetí tabulky ze sloupců „t Stat“ a „Hodnota P“.

Obrázek 79: Excel – Výsledky regrese



Hotel	Celková tržba	Tržba stravovacího úseku	Počet "lůžkonocí"
1	12	2	150
2	8	1,2	94
3	76,4	14,8	811
4	17	8,3	254
5	21,3	8,4	399
6	10	3	95
7	12,5	4,8	149
8	97,3	15,6	312
9	88	16,1	952
10	25	11,5	247
11	38,6	14,2	400
12	47,3	14	312

VÝSLEDEK

Regresní statistika

Násobné R	0,873896114
Hodnota spolehlivosti R	0,763694419
Nastavená hodnota spolehlivosti R	0,711182067
Chyba stř. hodnoty	17,35881298
Pozorování	12

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	8764,521174	4382,260587	14,54313885	0,001515769
Rezidua	9	2711,955492	301,328388		
Celkem	11	11476,47667			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
Hranice	-9,126447187	10,1669393	-0,897659258	0,392747237	-32,1256617	13,87276733	-32,1256617	13,87276733
Tržba stravovacího úseku	3,729273158	1,370575684	2,720953831	0,023570981	0,628815564	6,829730752	0,628815564	6,829730752
Počet "lůžkonocí"	0,033090576	0,028320995	1,168411481	0,272658149	-0,030975966	0,097157117	-0,03097597	0,097157117

REZIDUA

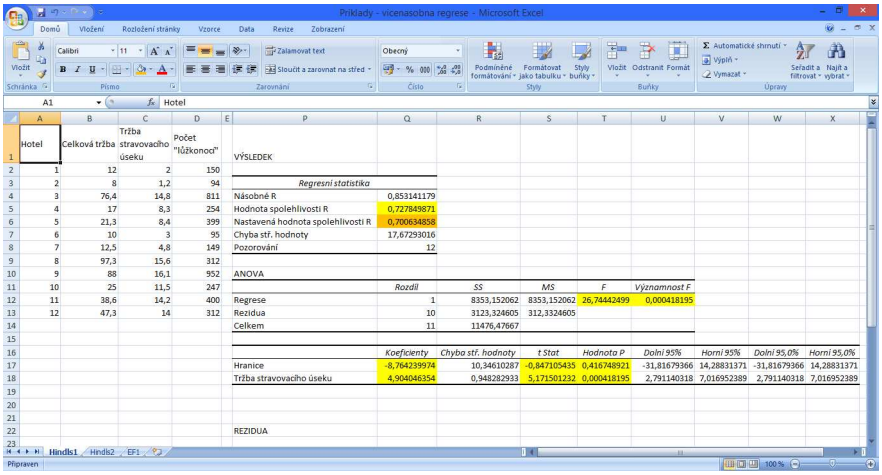
1.d) Předpoklady jsme již ověřili v programu STATISTICA, Excel pro tento účel není vhodný.

2) Bodovou předpověď opět získáme dosazením do regresní rovnice:

$$\hat{y} = -9{,}126 + 3{,}729 \cdot 7{,}5 + 0{,}033 \cdot 200 = 25{,}4415.$$

3.a) Vypuštěním proměnné X_2 získáme zjednodušený model, jehož výsledky můžeme vidět na obrázku Obrázek 80, které získáme znovu provedením výpočtu regrese (ve vstupní oblasti vynecháme hodnoty proměnné X_2).

Obrázek 80: Excel – Výsledky regrese 2



Hotel	Celková tržba	Tržba stravovacího úseku	Počet "lůžkonocí"
1	12	2	150
2	8	1,2	94
3	76,4	14,8	811
4	17	8,3	254
5	21,3	8,4	399
6	10	3	95
7	12,5	4,8	149
8	97,3	15,6	312
9	88	16,1	952
10	25	11,5	247
11	38,6	14,2	400
12	47,3	14	312

VÝSLEDEK

Regresní statistika

Násobné R	0,853141179
Hodnota spolehlivosti R	0,727849871
Nastavená hodnota spolehlivosti R	0,700634858
Chyba stř. hodnoty	17,67293016
Pozorování	12

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	8353,152062	8353,152062	26,74442499	0,000418195
Rezidua	10	3123,324605	312,3324605		
Celkem	11	11476,47667			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
Hranice	-8,764239974	10,34610287	-0,847105435	0,416748921	-31,81679366	14,28831371	-31,81679366	14,28831371
Tržba stravovacího úseku	4,904046354	0,948282933	5,171501232	0,000418195	2,791140318	7,016952389	2,791140318	7,016952389

REZIDUA

3.b) Porovnání modelů a výsledků jsme již provedli u řešení v programu STATISTICA v předchozí kapitole.

4.4 Příklady k procvičení

Tato kapitola zahrnuje příklady vhodné k procvičování. Kromě zadání všechny příklady obsahují také správné výsledky k případné kontrole.

4.4.1 Příklad č. 1

Výrobce mobilního telefonu testoval počet nepřerušených hovorů ve sto telefonátech v závislosti na teplotě pod bodem mrazu. Údaje jsou uvedeny v tabulce:

Tabulka 7: Data k příkladu č. 1

Teplota (°C)	0	-5	-10	-15	-20
Počet nepřerušených tel. hovorů/100 tel.	96	85	57	45	21

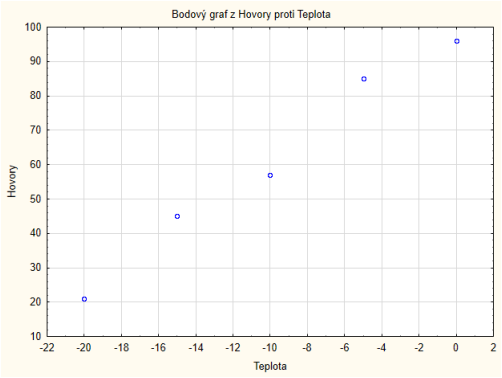
Zdroj: EF JU, přednášky prof. RNDr. Anny Čermákové, CSc.

Zadání

1) Nalezněte model jednoduché lineární regrese a ověřte vhodnost modelu.
2) Ověřte předpoklady užití nalezeného modelu.

Řešení v softwaru STATISTICA

Graf 17: Př. 1, STATISTICA – Korelační pole



Obrázek 81: Př. 1, STATISTICA – Korelační pole

Výsledky regrese se závislou proměnnou : Hovory (Tabulka14)
R= ,99141380 R2= ,98290133 Upravené R2= ,97720177
F(1,3)=172,45 p<,00095 Směrod. chyba odhadu : 4,5753

N=5	b*	Sm.chyba z b*	b	Sm.chyba z b	t(3)	p-hodn.
Abs.člen			98,80000	3,544009	27,87803	0,000101
Teplota	0,991414	0,075495	3,80000	0,289367	13,13211	0,000954

Řešení v softwaru Excel

Obrázek 82: Př. 1, Excel – Výsledky regrese

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Teplota	Hovory		VÝSLEDEK										
2	0	96												
3	-5	85		*Regresní statistika*										
4	-10	57		Násobné R	0,991413803									
5	-15	45		Hodnota spolehlivosti R	0,982901329									
6	-20	21		Nastavená hodnota spolehlivosti R	0,977201772									
7				Chyba stř. hodnoty	4,575295983									
8				Pozorování	5									
9														
10				ANOVA										
11					*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*					
12				Regrese	1	3610	3610	172,4522293	0,000953839					
13				Rezidua	3	62,8	20,93333333							
14				Celkem	4	3672,8								
15														
16					*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*		
17				Hranice	98,8	3,544009029	27,87803281	0,000101316	87,52138156	110,0786184	87,52138156	110,0786184		
18				Teplota	3,8	0,289367126	13,13210681	0,000953839	2,879104661	4,720895339	2,879104661	4,720895339		
19														
20														

Shrnutí výsledků

Regresní rovnice: $Y = 98{,}8 + 3{,}8X$.

Koeficient determinace: $R^2 = 0{,}9829$.

Test ANOVA: $p = 0{,}00095$ ..... model je významný.

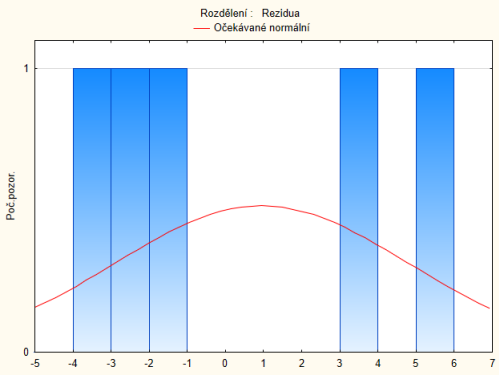
T-test pro β_0: $p = 0{,}00010$ ..... koeficient je významný.

T-test pro β_1: $p = 0{,}00095$ ..... koeficient je významný.

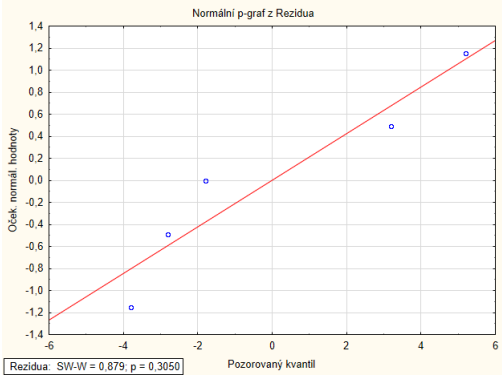
Model nelze zjednodušit.

Předpoklady modelu jsou splněny, viz Graf 18 a Graf 19.

Graf 18: Př. 1 – Histogram reziduí



Graf 19: Př. 1 – Shapiro-Wilkův test



4.4.2 Příklad č. 2

Máme k dispozici následující údaje týkající se ceny banánů za 1 kg a průměrně prodané množství v kg při dané ceně za den.

Tabulka 8: Data k příkladu č. 2

Cena	10	15	18	20	22	23	25	28	30	33
Prodáno	123	100	88	79	70	62	60	60	55	53

Zdroj: EF JU, přednášky prof. RNDr. Anny Čermákové, CSc.

Zadání

1) Nalezněte model jednoduché lineární regrese a určete jeho vhodnost.
2) Určete bodovou předpověď průměrně prodaného množství banánů v kg při ceně 12 Kč/kg.

Řešení v softwaru STATISTICA

Obrázek 83: Př. 2, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Prodáno (Tabulka18)
R= ,94972528 R2= ,90197811 Upravené R2= ,88972537
F(1,8)=73,614 p<,00003 Směrod. chyba odhadu : 7,5417

N=10	b*	Sm.chyba z b*	b	Sm.chyba z b	t(8)	p-hodn.
Abs.člen			143,9114	8,378338	17,17660	0,000000
Cena	-0,949725	0,110692	-3,0764	0,358560	-8,57988	0,000026

Řešení v softwaru Excel

Obrázek 84: Př. 2, Excel – Výsledky regrese

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Cena	Prodáno		VÝSLEDEK									
2	10	123											
3	15	100		*Regresní statistika*									
4	18	88		Násobné R	0,949725279								
5	20	79		Hodnota spolehlivosti R	0,901978106								
6	22	70		Nastavená hodnota spolehlivosti R	0,889725369								
7	23	62		Chyba stř. hodnoty	7,541697679								
8	25	60		Pozorování	10								
9	28	60											
10	30	55		ANOVA									
11	33	53			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*				
12				Regrese	1	4186,982369	4186,982369	73,61441989	2,62985E-05				
13				Rezidua	8	455,0176311	56,87720389						
14				Celkem	9	4642							
15													
16					*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*	
17				Hranice	143,9113924	8,378338087	17,17660363	1,3424E-07	124,5909101	163,2318747	124,5909101	163,2318747	
18				Cena	-3,076401447	0,358559769	-8,579884608	2,62985E-05	-3,903241757	-2,249561136	-3,903241757	-2,249561136	

Shrnutí výsledků

Regresní rovnice: $Y = 143{,}9 - 3{,}0764X$.

Koeficient determinace: $R^2 = 0{,}9020$.

Test ANOVA: $p = 0{,}00003$ ..... model je významný.

T-test pro β_0: $p \doteq 0$ ..... koeficient je významný.

T-test pro β_1: $p = 0{,}00003$ ..... koeficient je významný.

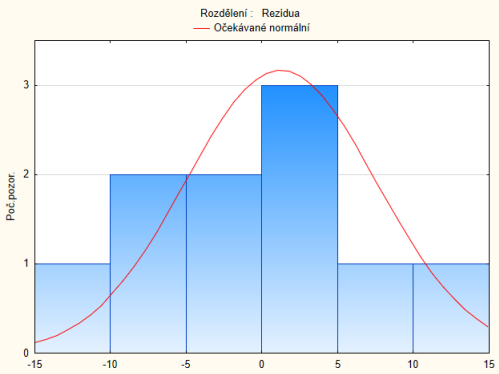
Model nelze zjednodušit.

Předpoklady modelu jsou splněny, viz Graf 20 a Graf 21.

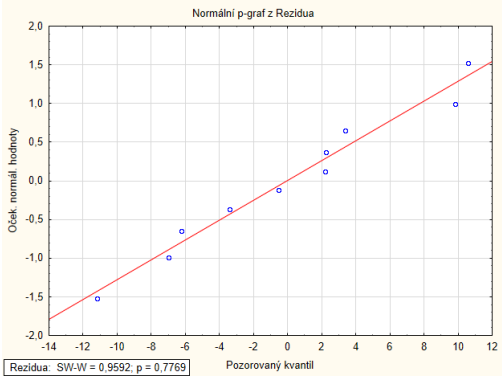
Bodová předpověď: $\hat{y}_{12} = 106{,}9946$.

Grafické znázornění regresní přímky můžeme vidět v grafu níže, viz Graf 22.

Graf 20: Př. 2 – Histogram reziduí



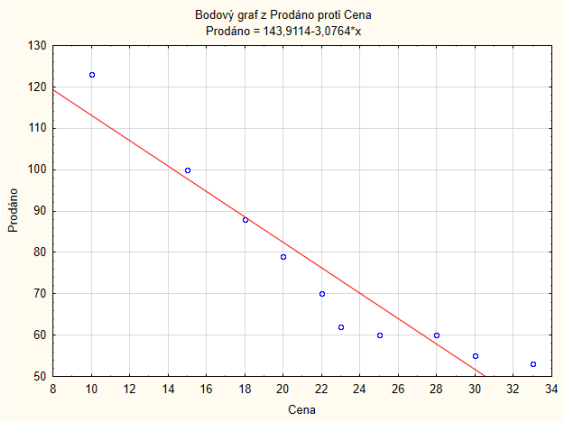
Graf 21: Př. 2 – Shapiro-Wilkův test



Obrázek 85: Př. 2 – Bodová předpověď

Proměnná	Předpovězené hodnoty (Tabulka18) proměnné: Prodáno		
	b-váha	Hodnota	b-váha * Hodnot
Cena	-3,07640	12,00000	-36,9168
Abs. člen			143,9114
Předpověď			106,9946
-95,0%LS			96,7872
+95,0%LS			117,2020

Graf 22: Př. 2 – Regresní přímka



4.4.3 Příklad č. 3

Nyní máme k dispozici údaje (časovou řadu) o roční výrobě elektrické energie (Y, v milionech kWh) v letech 1957 až 1979, přičemž zavedeme časový index t, který pro rok 1957 položíme rovný jedné. Data jsou shrnuty v tabulkách 9 až 11.

Tabulka 9: Data k příkladu č. 3 – 1. část

Rok	1957	1958	1959	1960	1961	1962	1963	1964
t	1	2	3	4	5	6	7	8
Y	17 720	19 620	21 884	24 450	26 962	28 795	29 861	31 983

Zdroj: Cipra [14]

Tabulka 10: Data k příkladu č. 3 – 2. část

Rok	1965	1966	1967	1968	1969	1970	1971	1972
t	9	10	11	12	13	14	15	16
Y	34 190	36 528	38 622	41 634	43 134	45 163	47 237	51 402

Zdroj: Cipra [14]

Tabulka 11: Data k příkladu č. 3 – 3. část

Rok	1973	1974	1975	1976	1977	1978	1979
t	17	18	19	20	21	22	23
Y	53 473	56 026	59 277	62 746	66 501	69 097	68 092

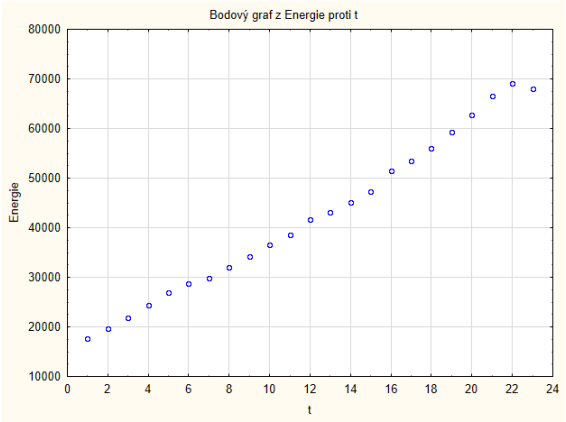
Zdroj: Cipra [14]

Zadání

1) Nalezněte vhodný model a ověřte jeho vhodnost.
2) Určete, o kolik se zvýší výroba elektrické energie po 5 letech.

Řešení v softwaru STATISTICA

Graf 23: Př. 3 – Korelační pole



Obrázek 86: Př. 3, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Energie (Tabulka1)
R= ,99614287 R2= ,99230062 Upravené R2= ,99193398
F(1,21)=2706,5 p<0,0000 Směrod. chyba odhadu : 1452,0

N=23	b*	Sm.chyba z b*	b	Sm.chyba z b	t(21)	p-hodn.
Abs.člen			13870,62	625,8277	22,16365	0,000000
t	0,996143	0,019148	2374,54	45,6432	52,02396	0,000000

Řešení v softwaru Excel

Obrázek 87: Př. 3, Excel – Výsledky regrese

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rok	t	Energie		VÝSLEDEK									
2	1957	1	17 720											
3	1958	2	19 620		*Regresní statistika*									
4	1959	3	21 884		Násobné R	0,996142871								
5	1960	4	24 450		Hodnota spolehlivosti R	0,992300619								
6	1961	5	26 962		Nastavená hodnota spolehlivosti R	0,991933982								
7	1962	6	28 795		Chyba stř. hodnoty	1451,9983								
8	1963	7	29 861		Pozorování	23								
9	1964	8	31 983											
10	1965	9	34 190		ANOVA									
11	1966	10	36 528			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*				
12	1967	11	38 622		Regrese	1	5706094466	5706094466	2706,491961	1,10912E-23				
13	1968	12	41 634		Rezidua	21	44274280,32	2108299,063						
14	1969	13	43 134		Celkem	22	5750368746							
15	1970	14	45 163											
16	1971	15	47 237			*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*	
17	1972	16	51 402		Hranice	13870,62451	625,8276921	22,16364773	4,76313E-16	12569,14458	15172,10443	12569,14458	15172,10443	
18	1973	17	53 473		t	2374,538538	45,6431755	52,02395565	1,10912E-23	2279,618358	2469,458717	2279,618358	2469,458717	
19	1974	18	56 026											
20	1975	19	59 277											

Shrnutí výsledků

Regresní rovnice:	$Y = 13\,870{,}62 + 2374{,}54X.$
Koeficient determinace:	$R^2 = 0{,}9923.$

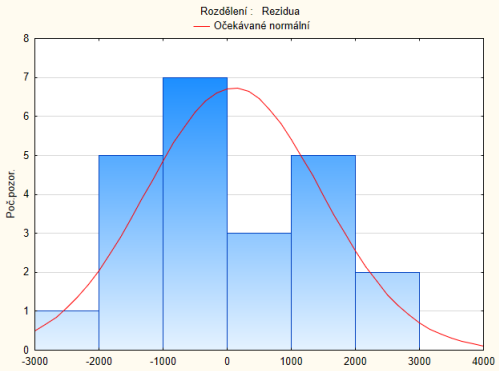
Test ANOVA:	$p \doteq 0$ ..... model je významný.
T-test pro β_0:	$p \doteq 0$ ..... koeficient je významný.
T-test pro β_1:	$p \doteq 0$ ..... koeficient je významný.

Model nelze zjednodušit.

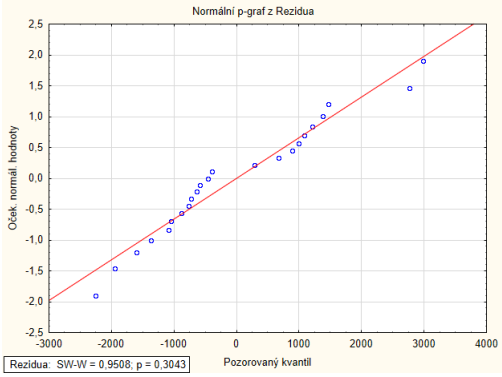
Předpoklady modelu jsou splněny, viz Graf 24 a Graf 25.

Výroba elektrické energie se každých 5 let zvýší o $5 \cdot 2374{,}54 = 11872{,}7$ GWh.

Graf 24: Př. 3 – Histogram reziduí



Graf 25: Př. 3 – Shapiro-Wilkův test



4.4.4 Příklad č. 4

Firma má záznamy za prvních sedm měsíců roku o počtu hodin provozu výrobní linky X a o nákladech na její údržbu Y v tisících Kč [5], viz Tabulka 12.

Tabulka 12: Data k příkladu č. 4

X	275	350	250	325	375	400	300
Y	149	170	140	164	192	200	165

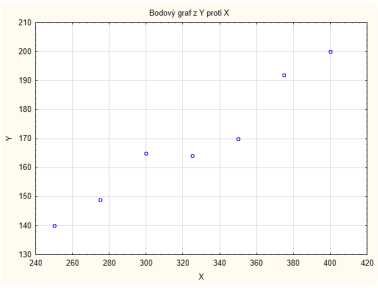
Zdroj: Příklad 8.1 [5], str. 84

Zadání

1) Nalezněte model vyjadřující závislost nákladů na údržbu firmy na provozu výrobní linky a ověřte vhodnost modelu.
2) Spočtěte bodovou předpověď nákladů firmy při 350 a 450 hodinovém provozu.

Řešení v softwaru STATISTICA

Graf 26: Př. 4, STATISTICA – Korelační pole



Obrázek 88: Př. 4, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Y (Tabulka1)
R= ,97278316 R2= ,94630708 Upravené R2= ,93556850
F(1,5)=88,122 p<,00023 Směrod. chyba odhadu : 5,4557

N=7	b*	Sm.chyba z b*	b	Sm.chyba z b	t(5)	p-hodn.
Abs.člen			42,75000	13,56100	3,152422	0,025310
X	0,972783	0,103627	0,38714	0,04124	9,387340	0,000231

Řešení v softwaru Excel

Obrázek 89: Př. 4, Excel – Výsledky regrese

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	X	Y		VÝSLEDEK											
2	275	149													
3	350	170		*Regresní statistika*											
4	250	140		Násobné R	0,972783163										
5	325	164		Hodnota spolehlivosti R	0,946307082										
6	375	192		Nastavená hodnota spolehlivosti R	0,935568498										
7	400	200		Chyba stř. hodnoty	5,455665469										
8	300	165		Pozorování	7										
9															
10				ANOVA											
11					*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*						
12				Regrese	1	2622,892857	2622,892857	88,12215023	0,000231306						
13				Rezidua	5	148,8214286	29,76428571								
14				Celkem	6	2771,714286									
15															
16					*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*			
17				Hranice	42,75	13,56100163	3,152422008	0,025309882	7,89033554	77,60966446	7,89033554	77,60966446			
18				X	0,387142857	0,041240954	9,387339891	0,000231306	0,281129609	0,493156106	0,281129609	0,493156106			

Shrnutí výsledků

Dle korelačního pole se zdál model jednoduché lineární regrese jako nejvhodnější kandidát, což potvrdily i výsledky.

Regresní rovnice: $Y = 42{,}75 + 0{,}387X$.

Koeficient determinace: $R^2 = 0{,}9463$.

Test ANOVA: $p = 0{,}00023$ ..... model je významný.

T-test pro β_0: $p = 0{,}02531$ ..... koeficient je významný.

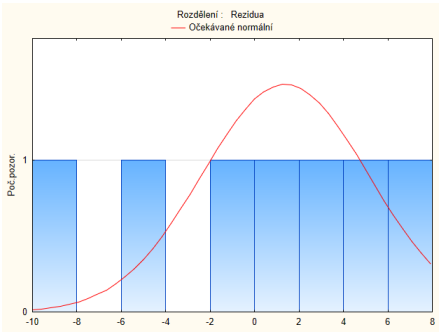
T-test pro β_1: $p = 0{,}00023$ ..... koeficient je významný.

Model nelze zjednodušit.

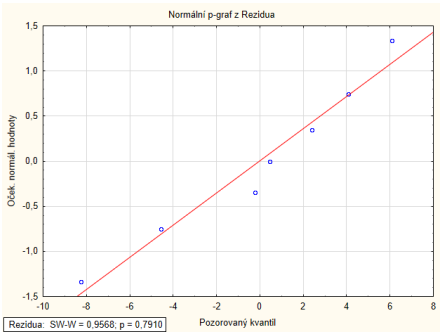
Předpoklady modelu jsou splněny, viz Graf 27, Graf 28 a Obrázek 90.

Bodové předpovědi: $\hat{y}_{350} = 178{,}25,\ \hat{y}_{450} = 216{,}96$.

Graf 27: Př. 4 – Histogram reziduí



Graf 28: Př. 4 – Shapiro-Wilkův test



Obrázek 90: Př. 4, STATISTICA – Výsledky t-testu

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka1)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	0,00	4,980318	7	1,882383	0,00	0,00	6	1,000000

Obrázek 91: Př. 4, STAT. – Bodová předpověď (350)

Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Y		
	b-váha	Hodnota	b-váha * Hodnot
X	0,387143	350,0000	135,5000
Abs. člen			42,7500
Předpověď			178,2500
-95,0%LS			172,3237
+95,0%LS			184,1763

Obrázek 92: Př. 4, STAT. – Bodová předpověď (450)

Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Y		
	b-váha	Hodnota	b-váha * Hodnot
X	0,387143	450,0000	174,2143
Abs. člen			42,7500
Předpověď			216,9643
-95,0%LS			202,6918
+95,0%LS			231,2368

4.4.5 Příklad č. 5

V tabulce (Tabulka 13) jsou uvedeny údaje o hodnotě produkce (Y) ve statisících Kč a o výši investic (X) v desetitisících Kč z roku 1998, kde bylo vybráno 12 soukromých firem s počtem zaměstnanců větším než 24 [6].

Tabulka 13: Data k příkladu č. 5

Firma	1	2	3	4	5	6	7	8	9	11	12	13
Y	52,8	48,4	54,2	50,0	54,9	53,9	53,1	52,4	53,0	52,9	53,1	60,1
X	16,3	16,8	18,5	16,3	17,9	17,4	16,1	16,2	17,0	16,7	17,5	19,1

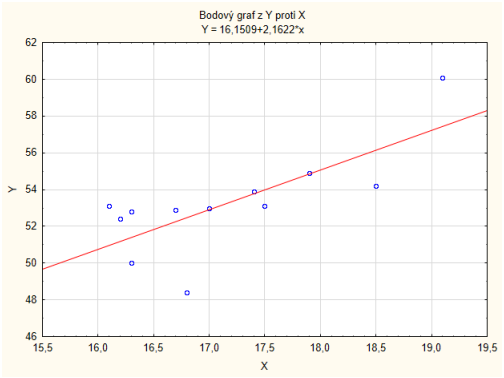
Zdroj: Příklad 4.5 [6], str. 188

Zadání

1) Nalezněte rovnici regresní přímky, modelující závislost hodnoty produkce na výši investice a určete spolehlivost nalezené rovnice.
2) Nalezněte bodovou předpověď cílové proměnné (produkce) při investici ve výši 250 000 Kč.
3) Zjistěte, o kolik se zvýší produkce, když do firmy investujeme 120 000 Kč.

Řešení v softwaru STATISTICA

Graf 29: Př. 5, STATISTICA – Korelační pole



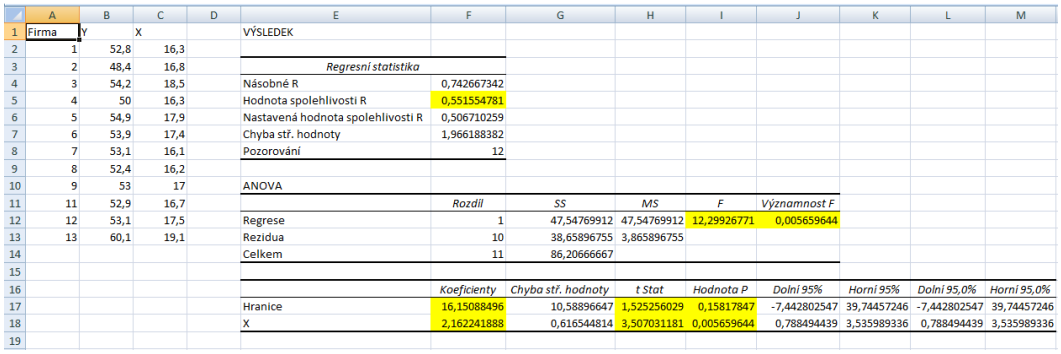
Obrázek 93: Př. 5, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Y (Tabulka8)
R= ,74266734 R2= ,55155478 Upravené R2= ,50671026
F(1,10)=12,299 p<,00566 Směrod. chyba odhadu : 1,9662

N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			16,15088	10,58897	1,525256	0,158178
X	0,742667	0,211765	2,16224	0,61654	3,507031	0,005660

Řešení v softwaru Excel

Obrázek 94: Př. 5, Excel – Výsledky regrese



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Firma	Y	X		VÝSLEDEK								
2	1	52,8	16,3										
3	2	48,4	16,8		*Regresní statistika*								
4	3	54,2	18,5		Násobné R	0,742667342							
5	4	50	16,3		Hodnota spolehlivosti R	0,551554781							
6	5	54,9	17,9		Nastavená hodnota spolehlivosti R	0,506710259							
7	6	53,9	17,4		Chyba stř. hodnoty	1,966188382							
8	7	53,1	16,1		Pozorování	12							
9	8	52,4	16,2										
10	9	53	17		ANOVA								
11	11	52,9	16,7			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*			
12	12	53,1	17,5		Regrese	1	47,54769912	47,54769912	12,29926771	0,005659644			
13	13	60,1	19,1		Rezidua	10	38,65896755	3,865896755					
14					Celkem	11	86,20666667						
15													
16						*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
17					Hranice	16,15088496	10,58896647	1,525256029	0,15817847	-7,442802547	39,74457246	-7,442802547	39,74457246
18					X	2,162241888	0,616544814	3,507031181	0,005659644	0,788494439	3,535989336	0,788494439	3,535989336
19													

Shrnutí výsledků

Regresní rovnice: $Y = 16{,}151 + 2{,}162X.$

Koeficient determinace: $R^2 = 0{,}9463.$

Test ANOVA: $p = 0{,}00566$ ..... model je významný.

T-test pro β_0: $p = 0{,}15818$ ..... koeficient není významný.

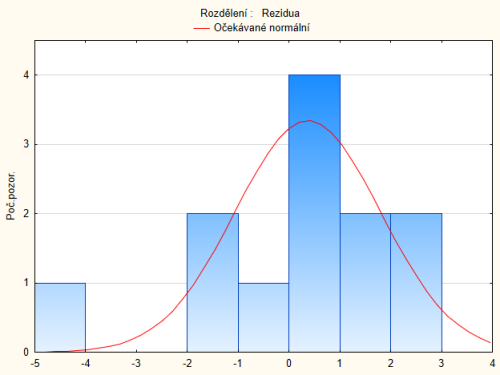
T-test pro β_1: $p = 0{,}00566$ ..... koeficient je významný.

Předpoklady modelu jsou splněny, viz Graf 30 a Graf 31.

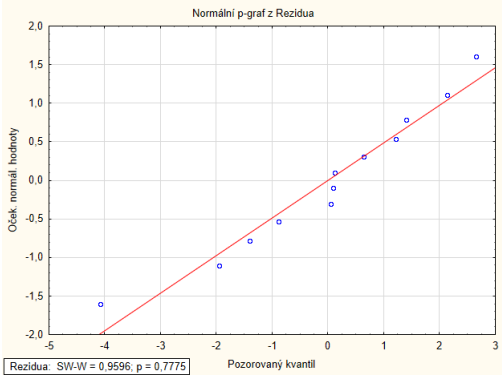
Bodová předpovědi: $\hat{y}_{25} = 70{,}20693 = 7\ 020\ 693$ Kč.

Pokud do firmy investujeme 12 000 Kč, produkce se zvýší o $1{,}2 \cdot 2{,}16 = 2{,}592 =$ 259 200 tisíc Kč.

Graf 30: Př. 5 – Histogram reziduí



Graf 31: Př. 5 – Shapiro-Wilkův test



Obrázek 95: Př. 5, STATISTICA – Bodová předpověď

Proměnná	Předpovězené hodnoty (Tabulka8) proměnné: Y		
	b-váha	Hodnota	b-váha * Hodnot
X	2,162242	25,00000	54,05605
Abs. člen			16,15088
Předpověď			70,20693
-95,0%LS			59,34911
+95,0%LS			81,06475

4.4.6 Příklad č. 6

V dalším příkladě máme k dispozici údaje o produkci (mil. FRF, z roku 1990, proměnná Y), o hrubé tvorbě fixního kapitálu (mil. FRF, z roku 1990, proměnná X_1) a o zaměstnanosti (tis. osob, z roku 1998, proměnná X_2) v jednotlivých odvětvích hospodářství ve Francii [6].

Tabulka 14: Data k příkladu č. 6

Odvětví	y_i	x_{1i}	x_{2i}
Zemědělství	288 443	18 781	1 055
Potravinářství	393 828	13 990	551
Energetika	330 300	33 813	223
Výroba polotovarů	602 182	32 022	1 101
Výroba výrobních zařízení	426 720	19 520	965
Výroba zařízení pro domácnosti	34 008	1 258	49
Výroba dopravních prostředků	185 887	10 462	358
Výroba spotřebních předmětů	427 766	16 392	1 030
Stavebnictví a veřejné práce	436 926	19 828	1 472
Obchod	495 319	36 354	2 691
Doprava a spoje	417 147	58 196	1 268
Tržní služby	1 002 132	116 083	4 617
Pojišťovací služby	61 827	2 053	158
Finanční služby	709 297	6 908	441
Netržní služby	840 622	136 923	6 148
Celkem	6 652 404	522 583	22 127

Zdroj: Příklady k procvičení [6], str. 241

Zadání

1) Nalezněte model vícenásobné regrese zkoumající závislost produkce na zadaných faktorech.
2) Pokud to bude vhodné, model zjednodušte.

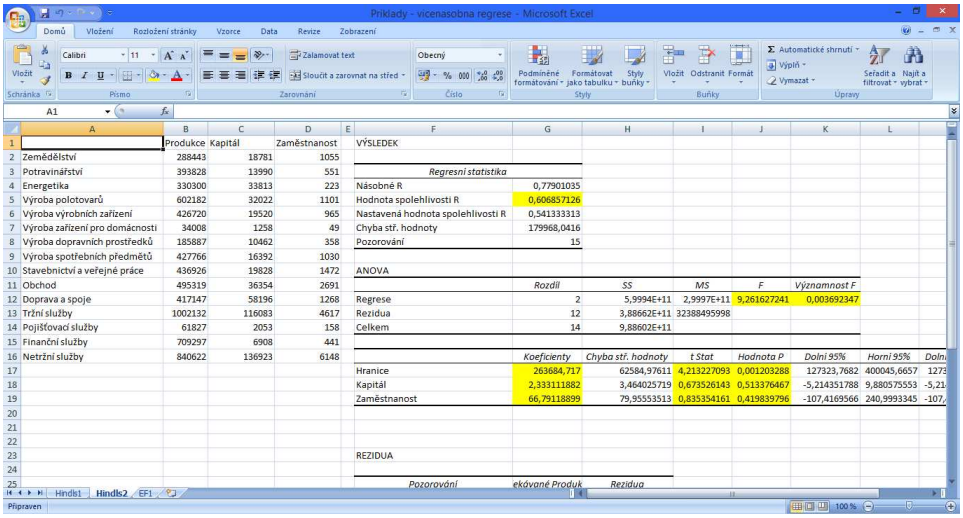
Řešení v softwaru STATISTICA

Obrázek 96: Př. 6, STATISTICA – Výsledky

N=15	Výsledky regrese se závislou proměnnou : Produkce (Tabulka1) R= ,77901035 R2= ,60685713 Upravené R2= ,54133331 F(2,12)=9,2616 p<,00369 Směrod. chyba odhadu : 1800E2					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(12)	p-hodn.
Abs.člen			263684,7	62584,98	4,213227	0,001203
Kapitál	0,353137	0,524311	2,3	3,46	0,673526	0,513376
Zaměstnanost	0,437985	0,524311	66,8	79,96	0,835354	0,419840

Řešení v softwaru Excel

Obrázek 97: Př. 6, Excel – Výsledky



Shrnutí výsledků

Regresní rovnice: $Y = 263\,684{,}7 + 2{,}333X_1 + 66{,}791X_2$.

Koeficient determinace: $R^2 = 0{,}6069$.

Test ANOVA: $p = 0{,}00369$ ..... model je významný.

T-test pro β_0: $p = 0{,}00120$ ..... koeficient je významný.

T-test pro β_1: $p = 0{,}51338$ ..... koeficient není významný.

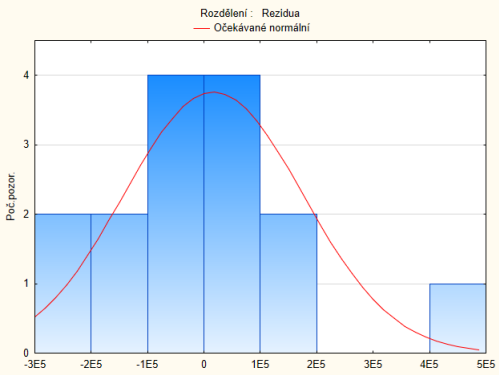
T-test pro β_2: $p = 0{,}41984$ ..... koeficient není významný.

Model lze zjednodušit.

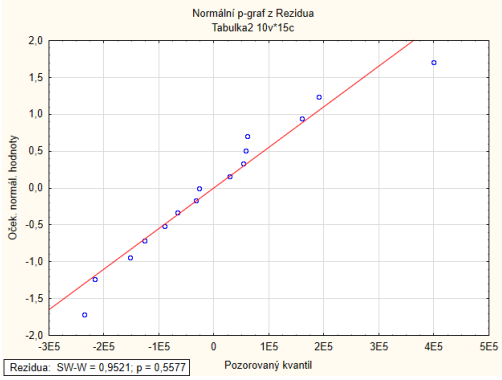
Předpoklady modelu jsou splněny, viz Graf 32 a Graf 33.

Ověření předpokladů v softwaru STATISTICA

Graf 32: Př. 6 – Histogram reziduí



Graf 33: Př. 6 – Shapiro-Wilkův test



Zjednodušený model

Obrázek 98: Př. 6, STATISTICA – Výsledky regrese – zjednodušený model

Výsledky regrese se závislou proměnnou : Produkce (Tabulka2)
R= ,76941218 R2= ,59199510 Upravené R2= ,56061011
F(1,13)=18,862 p<,00080 Směrod. chyba odhadu : 1761E2

N=15	b*	Sm.chyba z b*	b	Sm.chyba z b	t(13)	p-hodn.
Abs.člen			270412,4	60470,49	4,471807	0,000629
Zaměstnanost	0,769412	0,177158	117,3	27,02	4,343082	0,000797

Nejméně významnou proměnnou byl kapitál, po jeho vypuštění vzniknul jednodušší model $Y = 270\,412 + 117{,}3X_2$. Vhodnost modelu dle upraveného R^2 vzrostla z 0,5413 na 0,5606.

4.4.7 Příklad č. 7

Prodejce automobilů by chtěl zjistit, které ukazatele ovlivňují zájem zákazníků o koupi nového automobilu. Model byl vystaven v hypermarketu a za jeden den vyplnilo anketu 25 respondentů. Anketa se týkala těchto šesti ukazatelů:

Y ..... intenzita zájmu o koupi modelu (body od 0 do 10),

X_1 ..... počet let od koupě posledního automobilu,

X_2 ..... věk respondenta,

X_3 ..... subjektivní hodnocení poměru užitná hodnota/cena (body od 0 do 100),

X_4 ..... vzdálenost autorizované prodejny od bydliště (odhad v km),

X_5 ..... hrubý měsíční příjem respondenta v tisících Kč na 1 člena domácnosti.

Tabulka 15: Data k příkladu č. 7

i	Y	X_1	X_2	X_3	X_4	X_5
1	5	5	47	50	11	19
2	5	7	51	50	6	21
3	9	5	31	96	15	29
4	3	2	46	67	15	21
5	9	6	29	88	19	28
6	3	2	28	35	16	12
7	8	7	34	59	12	33
8	0	1	43	51	11	27
9	1	2	43	50	17	33
10	9	6	32	87	16	40
11	10	10	38	82	15	24
12	7	2	26	56	12	30
13	10	14	42	91	4	31
14	8	13	22	69	6	40
15	5	10	59	40	9	30
16	10	12	43	73	12	28
17	9	11	39	83	13	32
18	4	4	41	68	11	33
19	8	6	58	72	16	24
20	9	10	34	80	13	44
21	2	15	43	53	5	11
22	8	10	46	82	21	31
23	5	10	37	66	8	26
24	8	9	29	68	1	35
25	7	11	23	53	4	36

Zdroj: EF JU, přednášky prof. RNDr. Anny Čermákové, CSc.

Zadání

1) Nalezněte model vícenásobné lineární regrese a ověřte jeho spolehlivost.
2) Rozhodněte, které ukazatele ovlivňují zákazníky vzhledem k zájmu o koupi nového vozu nejvíce.

Řešení v softwaru STATISTICA

Obrázek 99: Př. 7, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Y (Tabulka20)
R= ,81852320 R2= ,66998023 Upravené R2= ,58313292
F(5,19)=7,7145 p<,00042 Směrod. chyba odhadu : 1,9200

N=25	b*	Sm.chyba z b*	b	Sm.chyba z b	t(19)	p-hodn.
Abs.člen			-2,52035	2,930359	-0,860083	0,400466
X1	0,372101	0,177406	0,27037	0,128905	2,097451	0,049564
X2	-0,135571	0,145071	-0,04081	0,043670	-0,934513	0,361765
X3	0,507794	0,177314	0,09070	0,031671	2,863806	0,009937
X4	0,129863	0,180109	0,07572	0,105020	0,721023	0,479674
X5	0,148466	0,155108	0,05542	0,057902	0,957181	0,350498

Řešení v softwaru Excel

Obrázek 100: Př. 7, Excel – Výsledky regrese

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		Y	X1	X2	X3	X4	X5		VÝSLEDEK										
2	1	5	5	47	50	11	19												
3	2	5	7	51	50	6	21		*Regresní statistika*										
4	3	9	5	31	96	15	29		Násobné R	0,818523202									
5	4	3	2	46	67	15	21		Hodnota spolehlivosti R	0,669980232									
6	5	9	6	29	88	19	28		Nastavená hodnota spolehlivosti R	0,583132925									
7	6	3	2	28	35	16	12		Chyba stř. hodnoty	1,920024609									
8	7	8	7	34	59	12	33		Pozorování	25									
9	8	0	1	43	51	11	27												
10	9	1	2	43	50	17	33		ANOVA										
11	10	9	6	32	87	16	40			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*					
12	11	10	10	38	82	15	24		Regrese	5	142,1966045	28,4393209	7,714461774	0,000415469					
13	12	7	2	26	56	12	30		Rezidua	19	70,0433955	3,6864945							
14	13	10	14	42	91	4	31		Celkem	24	212,24								
15	14	8	13	22	69	6	40												
16	15	5	10	59	40	9	30			*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*		
17	16	10	12	43	73	12	28		Hranice	-2,520352721	2,930359428	-0,860083134	0,400466029	-8,653665479	3,612960037	-8,653665479	3,612960037		
18	17	9	11	39	83	13	32		X1	0,270371896	0,128905003	2,097450754	0,049563946	0,000570624	0,540173168	0,000570624	0,540173168		
19	18	4	4	41	68	11	33		X2	-0,040810465	0,043670321	-0,934512607	0,361765348	-0,132213496	0,050592566	-0,132213496	0,050592566		
20	19	8	6	58	72	16	24		X3	0,090699946	0,031671121	2,863806011	0,009937301	0,024411529	0,156988363	0,024411529	0,156988363		
21	20	9	10	34	80	13	44		X4	0,075722104	0,105020379	0,721022953	0,479674073	-0,144088075	0,295532283	-0,144088075	0,295532283		
22	21	2	15	43	53	5	11		X5	0,055422382	0,057901678	0,957180931	0,350497894	-0,065767223	0,176611988	-0,065767223	0,176611988		
23	22	8	10	46	82	21	31												

Shrnutí výsledků

Regresní rovnice:

$$Y = -2{,}52 + 0{,}27X_1 - 0{,}041X_2 + 0{,}091X_3 + 0{,}076X_4 + 0{,}055X_5.$$

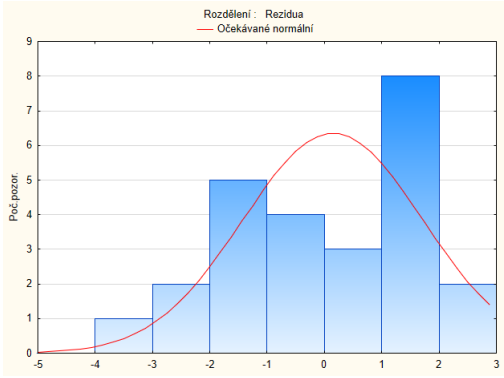
Koeficient determinace: $R^2 = 0{,}66998$.

Test ANOVA:	$p = 0{,}00042$	..... model je významný.
T-test pro β_0:	$p = 0{,}40047$	..... koeficient není významný.
T-test pro β_1:	$p = 0{,}04956$	..... koeficient je významný.
T-test pro β_2:	$p = 0{,}36177$	..... koeficient není významný.
T-test pro β_3:	$p = 0{,}00994$	..... koeficient je významný.
T-test pro β_4:	$p = 0{,}47967$	..... koeficient není významný.
T-test pro β_5:	$p = 0{,}35050$	..... koeficient není významný.

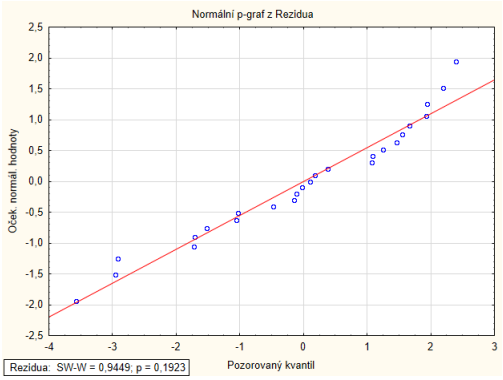
Předpoklady užití modelu jsou splněny, viz Graf 34 a Graf 35.

Zákazníky, kterých se anketa týkala, nejvíce ovlivňuje jejich subjektivní hodnocení poměru ceny a užitné hodnoty automobilu a potom, kdy si naposledy koupili nové auto. U ostatních ukazatelů jsme neprokázali výrazný vliv vzhledem k jejich zájmu o koupi nového automobilu.

Graf 34: Př. 7 – Histogram reziduí



Graf 35: Př. 7 – Shapiro-Wilkův test



4.4.8 Příklad č. 8

Máme k dispozici následující údaje z průzkumu zkoumající spokojenost studentů. Mezi ukazatele byly zařazeny finanční prostředky (v tisících Kč), vzdálenost školy od místa bydliště, shoda zájmů se zvoleným studijním oborem a velikost rodiny (počet členů v domácnosti).

Tabulka 16: Data k příkladu č. 8

Student	Spokojenost	Prostředky	Vzdálenost	Shoda zájmů	Velikost rodiny
1	7	5	10	8	3
2	9	8	3	10	4
3	4	9	26	4	3
4	2	3	30	2	3
5	3	4	20	4	5
6	1	8	30	1	6
7	5	5	15	6	6
8	6	2	5	7	5
9	3	3	25	4	3
10	7	2	10	9	4

Zdroj: EF JU, přednášky prof. RNDr. Anny Čermákové, CSc.

Zadání

1) Sestrojte model vícenásobné lineární regrese a určete jeho vhodnost.
2) Určete, které ukazatele jsou významné.

Řešení v softwaru STATISTICA

Obrázek 101: Př. 8, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Spokojenost (Tabulka10)
R= ,99727102 R2= ,99454948 Upravené R2= ,99018907
F(4,5)=228,09 p<,00001 Směrod. chyba odhadu : ,25166

N=10	b*	Sm.chyba z b*	b	Sm.chyba z b	t(5)	p-hodn.
Abs.člen			2,960851	1,417410	2,08892	0,091025
Prostředky	0,140007	0,034906	0,136751	0,034094	4,01096	0,010212
Vzdálenost	-0,299061	0,131871	-0,074653	0,032918	-2,26784	0,072635
Shoda zájmů	0,711074	0,130649	0,604095	0,110993	5,44263	0,002843
Velikost rodiny	-0,109952	0,043973	-0,227259	0,090887	-2,50047	0,054459

Řešení v softwaru Excel

Obrázek 102: Př. 8, Excel – Výsledky regrese

Student	Spokojenost	Prostředky	Vzdálenost	Shoda zájmů	Velikost rodiny
1	7	5	10	8	3
2	9	8	3	10	4
3	4	9	26	4	3
4	2	3	30	2	3
5	3	4	20	4	5
6	1	8	30	1	6
7	5	5	15	6	6
8	6	2	5	7	5
9	3	3	25	4	3
10	7	2	10	9	4

VÝSLEDEK

Regresní statistika

Násobné R	0,997271017
Hodnota spolehlivosti R	0,994549481
Nastavená hodnota spolehlivosti R	0,990189066
Chyba stř. hodnoty	0,251664516
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	4	57,78332486	14,44583121	228,0859667	7,6466E-06
Rezidua	5	0,316675143	0,063335029		
Celkem	9	58,1			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*
Hranice	2,960851263	1,417410097	2,088916446	0,091025025	-0,682717386
Prostředky	0,136750665	0,034094242	4,010960667	0,010211797	0,049108625
Vzdálenost	-0,074652984	0,032918166	-2,267835451	0,072635373	-0,159271823
Shoda zájmů	0,604094718	0,110993116	5,442632272	0,00284279	0,318777829
Velikost rodiny	-0,227259178	0,090886726	-2,500466113	0,054459041	-0,460890944

Shrnutí výsledků

Regresní rovnice:

$$Y = 2{,}96 + 0{,}137X_1 - 0{,}075X_2 + 0{,}604X_3 + 0{,}076X_4 - 0{,}227X_5.$$

Koeficient determinace: $R^2 = 0{,}9945$.

Test ANOVA: $p = 0{,}00001$ ..... model je významný.

T-test pro β_0: $p = 0{,}09103$ ..... koeficient není významný.

T-test pro β_1: $p = 0{,}01021$ ..... koeficient je významný.

T-test pro β_2: $p = 0{,}07264$ ..... koeficient není významný.

T-test pro β_3: $p = 0{,}00284$ ..... koeficient je významný.

T-test pro β_4: $p = 0{,}05446$ ..... koeficient není významný.

Nejvýznamnějšími se pro spokojenost studentů ukázaly finanční prostředky a shoda zájmů se zvoleným studijním oborem.

4.4.9 Příklad č. 9

V tabulce Tabulka 17 máme k dispozici údaje z podniku Canard, kde sledovali závislost vlastních nákladů připadajících na jednotku produkce (Y) na objemu produkce v 1 000 ks (X) [6].

Tabulka 17: Data k příkladu č. 9

X	60	71	92	144	192	306	437	481	747	989	1383
Y	5 157	2 620	1 986	1 582	1 100	954	729	456	200	196	110

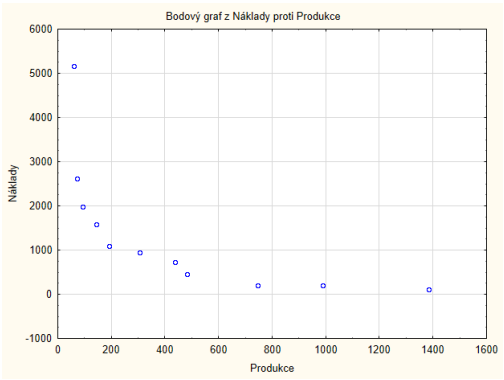
Zdroj: Příklad 4.5 [6], str. 196

Zadání

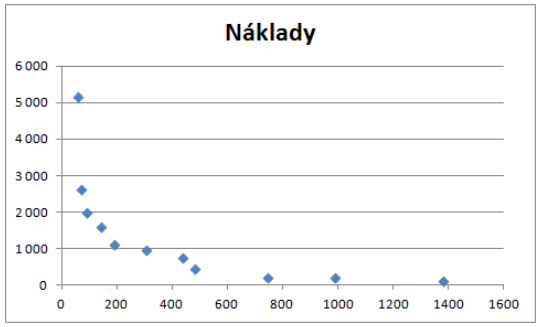
1) Nalezněte hyperbolický model nelineární regrese a ověřte jeho spolehlivost.

Korelační pole

Graf 36: Př. 9, STATISTICA – Korelační pole



Graf 37: Př. 9, Excel – Korelační pole



Řešení v softwaru STATISTICA

Obrázek 103: Př. 9, STATISTICA – Výsledky regrese

Výsledky regrese se závislou proměnnou : Náklady (Tabulka6)
R= ,94501273 R2= ,89304905 Upravené R2= ,88116561
F(1,9)=75,151 p<,00001 Směrod. chyba odhadu : 513,11

N=11	b*	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			-96,8	229,42	-0,421825	0,683050
1/V1	0,945013	0,109011	250528,9	28899,56	8,668952	0,000012

Řešení v softwaru Excel

Obrázek 104: Př. 9, Excel – Výsledky regrese

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Produkce	Náklady	1/Produkce		VÝSLEDEK									
2	60	5 157	0,01667											
3	71	2 620	0,01408		*Regresní statistika*									
4	92	1 986	0,01087		Násobné R	0,945012726								
5	144	1 582	0,00694		Hodnota spolehlivosti R	0,893049053								
6	192	1 100	0,00521		Nastavená hodnota spolehlivosti R	0,881165614								
7	306	954	0,00327		Chyba stř. hodnoty	513,1086168								
8	437	729	0,00229		Pozorování	11								
9	481	456	0,00208											
10	747	200	0,00134		ANOVA									
11	989	196	0,00101			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*				
12	1383	110	0,00072		Regrese	1	19785717,56	19785717,56	75,15072754	1,1589E-05				
13					Rezidua	9	2369524,074	263280,4526						
14					Celkem	10	22155241,64							
15														
16						*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*	
17					Hranice	-96,77523991	229,4205708	-0,421824597	0,683050103	-615,7606263	422,2101465	-615,7606263	422,2101465	
18					1/Produkce	250528,8794	28899,55804	8,668951929	1,1589E-05	185153,5373	315904,2215	185153,5373	315904,2215	
19														

Shrnutí výsledků

Korelační pole potvrzuje, že hyperbolický model se zdá být vhodným.

Regresní rovnice: $Y = -96{,}78 + \frac{250\,528{,}9}{X}$.

Koeficient determinace: $R^2 = 0{,}8930$.

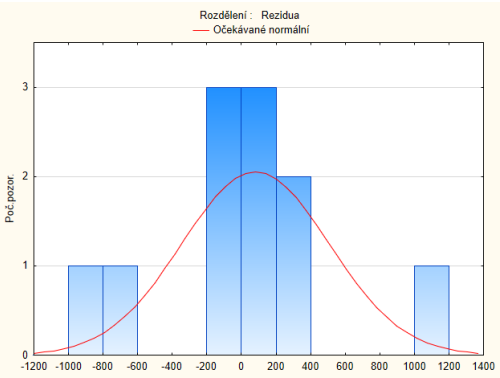
Test ANOVA: $p = 0{,}00001$ ..... model je významný.

T-test pro β_0: $p = 0{,}68305$ ..... koeficient není významný.

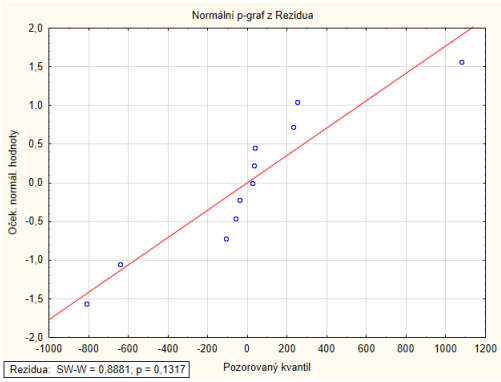
T-test pro β_1: $p = 0{,}00001$ ..... koeficient je významný.

Předpoklady užití modelu jsou splněny, viz Graf 38 a Graf 39.

Graf 38: Př. 9 – Histogram reziduí



Graf 39: Př. 9 – Shapiro-Wilkův test



4.4.10 Příklad č. 10

Prodejna softwarových produktů a počítačových her nechala své prodavače absolvovat prodejní kurz. Poté náhodně provedla 20 měření, ve kterých se sledovalo, kolik osob navštíví během prodejní doby prodejnu (X) a jaká je v onen den tržba (Y), zaokrouhleno v tisících Kč. Údaje jsou uvedeny v tabulkách (viz Tabulka 18 a Tabulka 19). [6]

Tabulka 18: Data k příkladu č. 10 – 1. část

i	1	2	3	4	5	6	7	8	9	10
X	20	21	26	27	28	29	30	31	32	34
Y	5	6	7	7	8	9	10	11	12	13

Zdroj: Příklad 4.4 [6], str. 192

Tabulka 19: Data k příkladu č. 10 – 2. část

i	11	12	13	14	15	16	17	18	19	20
X	35	37	38	39	42	44	48	49	51	54
Y	13	14	14	15	16	15	15	14	13	13

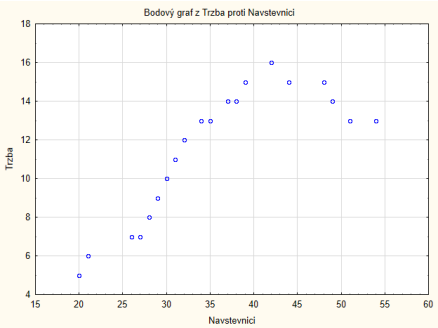
Zdroj: Příklad 4.4 [6], str. 192

Zadání

1) Nalezněte model, který by nejlépe vyjadřoval závislost denní tržby prodejny na počtu návštěvníků během prodejní doby, a ověřte jeho vhodnost.

Řešení v softwaru STATISTICA

Graf 40: Př. 10, STATISTICA – Korelační pole



Obrázek 105: Př. 10, STATISTICA – Lineární model

Výsledky regrese se závislou proměnnou : Trzba (Tabulka13)
R= ,82022007 R2= ,67276097 Upravené R2= ,65458102
F(1,18)=37,006 p<,00001 Směrod. chyba odhadu : 1,9953

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(18)	p-hodn.
Abs.člen			1,360994	1,725401	0,788798	0,440496
Navstevnici	0,820220	0,134833	0,283609	0,046621	6,083228	0,000010

Obrázek 106: Př. 10, STATISTICA – Kvadratický model

Výsledky regrese se závislou proměnnou : Trzba (Tabulka13)
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
Navstevnici	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
V1**2	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

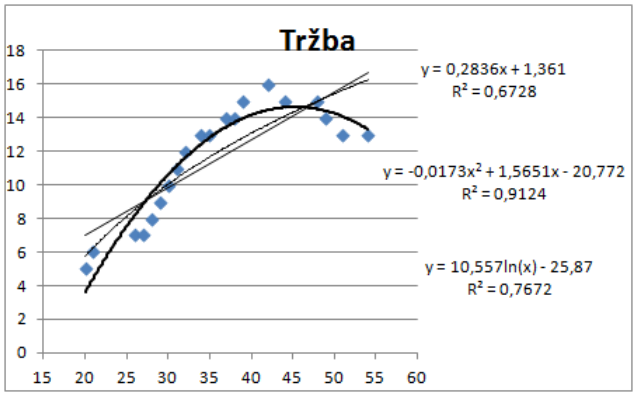
Obrázek 107: Př. 10, STATISTICA – Logaritmický model

Výsledky regrese se závislou proměnnou : Trzba (Tabulka13)
R= ,87587225 R2= ,76715219 Upravené R2= ,75421620
F(1,18)=59,304 p<,00000 Směrod. chyba odhadu : 1,6831

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(18)	p-hodn.
Abs.člen			-25,8698	4,867232	-5,31509	0,000047
LN-V1	0,875872	0,113736	10,5573	1,370922	7,70089	0,000000

Řešení v softwaru Excel

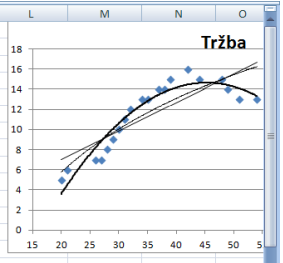
Graf 41: Př. 10, Excel – Korelační pole s rovnicemi



Obrázek 108: Př. 10, Excel – Výsledky regrese (kvadratický model)

	A	B	C	D	E	F	G	H	I	J	K
1	Prodejna	Počet návštěvníků		Tržba		VÝSLEDEK					
2		X	X^2	Y							
3	1	20	400	5		*Regresní statistika*					
4	2	21	441	6		Násobné R	0,955192764				
5	3	26	676	7		Hodnota spolehlivosti R	0,912393215				
6	4	27	729	7		Nastavená hodnota spolehlivosti R	0,902086535				
7	5	28	784	8		Chyba stř. hodnoty	1,062347174				
8	6	29	841	9		Pozorování	20				
9	7	30	900	10							
10	8	31	961	11		ANOVA					
11	9	32	1024	12			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*
12	10	34	1156	13		Regrese	2	199,8141142	99,90705709	88,52444904	1,027E-09
13	11	35	1225	13		Rezidua	17	19,18588581	1,128581518		
14	12	37	1369	14		Celkem	19	219			
15	13	38	1444	14							
16	14	39	1521	15			*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*
17	15	42	1764	16		Hranice	-20,77225468	3,373255904	-6,157924354	1,05223E-05	-27,88920247
18	16	44	1936	15		X	1,565102222	0,189558768	8,256554104	2,36857E-07	1,165168184
19	17	48	2304	15		X^2	-0,017289198	0,002535401	-6,819118197	2,98665E-06	-0,022638426
20	18	49	2401	14							
21	19	51	2601	13							
22	20	54	2916	13							

	L	M	N	O
	Horní 95%	*Dolní 95,0%*	*Horní 95,0%*	
17	-13,65530689	-27,88920247	-13,65530689	
18	1,965036259	1,165168184	1,965036259	
19	-0,01193997	-0,022638426	-0,01193997	



Shrnutí výsledků

Z korelačního pole jsem usoudil, že nejvhodnější bude zřejmě nalézt kvadratický model. Pro srovnání jsem vytvořil také lineární model a logaritmický model, který se také nabízel jako vhodnou alternativou.

Lineární model:	$Y = 1{,}361 + 0{,}2836X.$
Koeficient determinace:	$R^2 = 0{,}6728.$
Test ANOVA:	$p = 0{,}00001$ ..... model je významný.
T-test pro β_0:	$p = 0{,}44050$ ..... koeficient není významný.
T-test pro β_1:	$p = 0{,}00001$ ..... koeficient je významný.
Kvadratický model:	$Y = -20{,}77 + 1{,}565X - 0{,}0173X^2.$
Koeficient determinace:	$R^2 = 0{,}9124.$
Test ANOVA:	$p \doteq 0$ ..... model je významný.
T-test pro β_0:	$p = 0{,}00001$ ..... koeficient je významný.
T-test pro β_1:	$p \doteq 0$ ..... koeficient je významný.
T-test pro β_2:	$p = 0{,}000003$ ..... koeficient je významný.
Logaritmický model:	$Y = -25{,}87 + 10{,}557 \ln X.$
Koeficient determinace:	$R^2 = 0{,}7672.$
Test ANOVA:	$p \doteq 0.$
T-test pro β_0:	$p = 0{,}00005$ ..... koeficient je významný.
T-test pro β_1:	$p \doteq 0$ ..... koeficient je významný.

Nejlépe vyšel kvadratický model s koeficientem determinace $0{,}9124$ a s přesvědčivými výsledky ohledně významnosti modelu i všech regresních koeficientů. Lineární model nevyšel až tak bezvýznamně s indexem spolehlivosti $0{,}6728$ a logaritmický model k mému překvapení nedopadl příliš dobře, jen o něco málo lépe než lineární model s indexem spolehlivosti $0{,}7672$. I přesto můžeme na základě výsledků analýzy rozptylu říci, že všechny nalezené modely jsou jako celek významné.

5 Diskuze

V této kapitole se pokusím shrnout hlavní výhody a nevýhody statistického softwaru STATISTICA a tabulkového procesoru Excel. Osobně mám zkušenosti s oběma programy, nicméně pojďme se zaměřit na ty části, které se nás týkají především.

Nejprve si řekneme pár slov o jejich dostupnosti. Programy STATISTICA i Excel patří k velmi známým aplikacím a oba jsou studenty často využívány. STATISTICA je sice cenově méně dostupná než Excel, který je součástí balíčku Microsoft Office a je řádově výrazně levnější, nicméně studenti Jihočeské univerzity v Českých Budějovicích mají možnost ji během studia využívat bezplatně (ke studijním účelům), zatím (!).

Software STATISTICA je velice robustní program a na statistické výpočty je přímo stvořený, nabízí mnohem větší množství statistických nástrojů než program Excel, což je její velikou výhodou. Navíc se software STATISTICA začíná využívat při výuce na vysokých školách čím dál častěji a tak vznikají i návody na její ovládání. Tím bych ovšem zároveň přešel k nevýhodám softwaru STATISTICA. Proč vlastně vznikají návody na její ovládání? Protože software STATISTICA je silně nepřehledný. Někde něco zapomenete nastavit a máte problém. Některá nastavení se mezi záložkami v dialogových oknech ovlivňují a jiná zas jsou na sobě úplně nezávislá. V řešených příkladech této práce si můžete všimnout, že se proměnné vybírají během jednoho výpočtu vícekrát. Téměř při kterékoli práci se několikrát překlikáváte mezi dialogovými okny, než dojdete k výsledku. Ve finále se často ani nedozvíte, jakým způsobem se k výsledkům došlo, jaké výpočty tam proběhly, o kontrole ani nemluvě a vše funguje jako taková „černá skříňka“.

S tabulkovým procesorem Excel se dnes většina studentů setkává již na základní nebo střední škole a tak bývá i mezi studenty na vysokých školách preferovaný před ostatními „početními“ programy, protože se v něm orientují lépe. V Excelu oproti STATISTICE víte, co zrovna kde počítáte a tak se studenti mohou více zaměřit na postup řešeného příkladu a lépe se pak v daném tématu orientují. Silná stránka Excelu je jeho jednoduchost. Je pravda, že bez doplňku Analýzy dat bychom stěží počítali příklady z regresní analýzy tak jednoduše (jak které části příkladu, určité věci zvládá bodový graf), ale s tímto doplňkem provádíte nastavení výpočtu pouze v jednom dialogovém okně.

Ve zkratce, oba programy mají své výhody a své nevýhody. Osobně si troufám říci, že z hlediska řešení regresní analýzy jsou oba programy na srovnatelné úrovni. STATISTICA sice poskytuje více možností, ale je třeba se s ní naučit správně pracovat.

6 Závěr

Domnívám se, že by tato práce mohla posloužit čtenářům, kteří se chtějí blíže seznámit s regresní analýzou, a že by mohla být přínosem pro studenty na vysokých školách během studia, případně pro vyučující jako výukový materiál, či byla uváděna jako doporučená literatura. Osobně mi přijde, že přehledných komplexně zpracovaných materiálů vztahujících se k regresní analýze není mnoho a tak bych chtěl tuto oblast obohatit a byl bych rád, kdyby tato práce nalezla své využití v praxi.

Mně samotnému bylo největším přínosem zpracování teoretické části, při kterém jsem si uvědomil řadu nejasností a získal jsem nové poznatky. Nyní mohu tyto poznatky předávat dál a také na nich postavit základy pro své další studium statistických disciplín.

7 Seznam použité literatury

[1] ZVÁRA, Karel. *Regresní analýza*. Praha: Academia, 1989. ISBN 80-200-0125-5.

[2] HENDL, Jan. Přehled statistických metod zpracování dat: analýza a metaanalýza dat. Vyd. 2., opr. Praha: Portál, 2006. ISBN 80-7367-123-9.

[3] KOHN, Stanislav. *Základy teorie statistické metody*. Praha: Státní úřad statistický, 1929.

[4] Linear regression. *Wikipedia: the free encyclopedia* [online]. ©2018 [cit. 12.7.2018]. Dostupné z: https://en.wikipedia.org/wiki/Linear_regression

[5] MRKVIČKA, Tomáš a Vladimíra PETRÁŠKOVÁ. *Úvod do statistiky*. České Budějovice: Jihočeská univerzita, 2006. ISBN 80-7040-894-4.

[6] HINDLS, Richard. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[7] Regression analysis. *Wikipedia: the free encyclopedia* [online]. ©2018 [cit. 12.7.2018] Dostupné z: https://en.wikipedia.org/wiki/Regression_analysis

[8] Assumtions of Linear Regression. *r-statistics.co* [online] ©2017 [cit. 12.7.2018]. Dostupné z: http://r-statistics.co/Assumptions-of-Linear-Regression.html

[9] Partition of sums of squares. *Wikipedia: the free encyclopedia* [online]. ©2018 [cit. 12.7.2018] Dostupné z: https://en.wikipedia.org/wiki/Partition_of_sums_of_squares

[10] Coefficient of determination. *Wikipedia: the free encyclopedia* [online]. ©2018 [cit. 12.7.2018] Dostupné z: https://en.wikipedia.org/wiki/Coefficient_of_determination

[11] Nákup a porovnání produktů Microsoft Office | Office. *Microsoft* [online]. ©2018 [cit. 12.7.2018]. Dostupné z: https://products.office.com/cs-cz/compare-all-microsoft-office-products/

[12] General linear model. *Wikipedia: the free encyclopedia* [online]. ©2018 [cit. 12.7.2018] Dostupné z: https://en.wikipedia.org/wiki/General_linear_model

[13] ANDĚL, Jiří. *Matematická statistika*. Praha: SNTL - Nakladatelství technické literatury, 1978.

[14] CIPRA, Tomáš. *Analýza časových řad s aplikacemi v ekonomii*. Praha: Státní nakladatelství technické literatury, 1986.

8 Seznam tabulek, grafů a obrázků

Tabulky

Grafy

Obrázky