Téma 5 – Příklad 2

Zadání příkladu

Předpokládejme, že jsme od 6 domácností získali údaje o počtu členů a měsíčních výdajích za určitou komoditu.

1. Odhadněte parametry regresní přímky, která vystihuje závislost měsíčních výdajů (y) na počtu členů domácnosti (x) a interpretujte hodnotu regresního koeficientu.
2. Ověřte smysl volby počtu členů jako vysvětlující proměnné pomocí t-testu (nebo celkového F-testu).
3. Pomocí regresního modelu odhadněte průměrné výdaje čtyřčlenné domácnosti.
4. Posuďte kvalitu modelu pomocí koeficientu determinace.

Počet členů	Výdaje
1	550
2	750
3	1200
4	1450
5	2200
6	2250

Vypracování příkladu

Zvolený model regresní přímky má dva základní parametry – směrnici a konstantu (průsečík s osou y). Její obecný vzorec je $y_i = \eta_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Pro odhad těchto parametrů použijeme metodu nejmenších čtverců, tedy budeme minimalizovat čtverce odchylek reziduí. Ze soustavy normálních rovnic vznikají následující vzorce pro výpočet jednotlivých parametrů:

$$b_1 = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n} \cdot \frac{\sum_{i=1}^{n} y_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x^2} = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \frac{\sum y_i \cdot \sum x_i^2 - \sum x_i \cdot \sum x_i \cdot y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} = \bar{y} - b_1 \cdot \bar{x}$$

Parametr b_1 se rovná 90. Parametr b_0 vyšel přibližně 374,286. Výběrová regresní přímka má podobu: y = 90,000 + 374,286 x

Nyní musíme ověřit vhodnost jednotlivých parametrů pomocí t-testů a celkový model F-testem.

1. H_0: $\beta_0 = 0$ – parametr konstanty (průsečíku) je statisticky nevýznamný
   H_1: non H_0
2. Testové kritérium
   $$t = \frac{b_0}{s(b_0)} = 0{,}616$$
3. $$W \equiv \left\{ t;\ t \le t_{\frac{\alpha}{2}}(n-p) \cup t \ge t_{1-\frac{\alpha}{2}}(n-p) \right\}$$
4. Testové kritérium je prvkem oboru přijetí. Nulovou hypotézu nezamítáme, nepřijímáme hypotézu alternativní.

Stejným postupem bude testován parametr směrnice

1. H_0: $\beta_1 = 0$ – parametr směrnice je statisticky nevýznamný
   H_1: non H_0
2. Testové kritérium
   $$t = \frac{b_1}{s(b_1)} = 9{,}981$$
3. $$W \equiv \left\{ t;\ t \le t_{\frac{\alpha}{2}}(n-p) \cup t \ge t_{1-\frac{\alpha}{2}}(n-p) \right\}$$
4. Testové kritérium je prvkem kritického oboru. Nulovou hypotézu zamítáme, přijímáme hypotézu alternativní.

Na základě dílčích t-testů můžeme říci, že v modelu regresní přímky je statisticky významný pouze parametr směrnice. Jelikož je to ovšem parametr determinující podobu přímky, lze považovat model za statisticky významný.

Pozn.: Užití a interpretace konstanty v modelu je nutné zvážit dle logického smyslu analýzy závislostí. Někdy je zahrnutí konstanty v modelu dokonce nejednoznačné. Například v našem příkladu si můžeme říci, že bezečlenná domácnost nemůže existovat, proto je možné konstruovat model bez konstanty (učiníme tak v postupu s programem SGP). Na druhou stranu můžeme domácnost chápat jako bytovou jednotku, kde dočasně nemusí nikdo pobývat. Pak by interpretace takové konstanty mohla představovat třeba paušál za nějakou službu, pronájem elektroměru apod. Vždy je velice důležité znát dobře svá data!

Dalším krokem regresní analýzy bude provedení celkového F-testu o vhodnosti modelu. Jde o analogii analýzy rozptylu, kdy rozložení variability proběhne na modelovou a reziduální.

1. H_0: $\beta_0 = c$ – zvolený model regresní přímky není pro vystižení závislosti vhodný
   $\beta_1 = 0$ – jedná se o konstantní funkci
   H_1: non H_0
2. Testové kritérium
$$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}} = 99{,}630$$
3. $W \equiv (F, F \geq F_{1-\alpha}(p-1; n-p))$
4. $99{,}630 \geq 7{,}709$
   Testové kritérium je prvkem kritického oboru. Nulovou hypotézu zamítáme, přijímáme hypotézu alternativní.

Bodový odhad průměrných výdajů čtyřčlenné domácnosti učiníme prostým dosazením do rovnice odhadované regresní příjmy. Pokud x = 4, pak y = 1587,14 Kč.

Kvalitu regresního modelu posoudíme indexem determinace (u přímkového modelu shodný s koeficientem determinace). Jde o podíl modelové variability na variabilitě celkové:

$$I^2 = \frac{S_T}{S_y} = 96{,}14\,\%$$

Opět jde o poměr části z celku, tudíž je smysluplné procentní vyjádření.

Řešení v SGP

Zadání do programu není v tomto případě vůbec problematické. Stačí přepsat obě proměnné do samostatných sloupců.

Procedura v SGP: Relate – One Factor - Simple Regression (Analysis Summary, Forecasts)

Při vstupním dialogu zadáme jako závislou proměnnou y v*ydaje* a jako proměnnou nezávislou x p*ocet_clenu*.

V druhém kroku – Analysis options – vybíráme regresní funkci. K dispozici je široké spektrum funkcí. Doplňkově lze zvolit začlenění či nezačlenění konstanty do modelu (Include Constant), případně jiný výpočet parametru regresní funkce než je metoda nejmenších čtverců (Alternative Fit).

Okno Analysis Options shrnuje výsledky regresní analýzy. V tabulce Coefficients jsou odhady jednotlivých parametru a jejich dílčí t-testy. V tomto případě parametr průsečíku (Intercept) a směrnice (Slope). T-Statistic a P-Value ukazuje příslušnou hodnotu kritéria a jeho významnosti. Druhá tabulka s názvem Analysis of Variance zobrazuje provedený F-test. Rozklad variability na teoretickou (Model) a residuální (Residual). Hodnota kritéria F (F-Ratio) a jeho významnost (P-Value) nám vypoví, zda je model na zvolené hladině významnosti vhodný k vystižení závislosti.

Pod tabulkami jsou hodnoty ukazatelů, které slouží k charakteristice a posouzení vhodnosti modelu. Index determinace odpovídá hodnotě R-squared. V poli StatAdvisor můžeme ještě jednou vidět kompletní rovnici konstruovaného modelu.

Simple Regression - vydaje vs. pocet_clenu
Dependent variable: vydaje
Independent variable: pocet_clenu
Linear model: Y = a + b*X
Number of observations: 6

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	90,0	146,035	0,616291	0,5710
Slope	374,286	37,4983	9,98141	0,0006

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	2,45157E6	1	2,45157E6	99,63	0,0006
Residual	98428,6	4	24607,1		
Total (Corr.)	2,55E6	5			

Correlation Coefficient = 0,98051
R-squared = 96,1401 percent
R-squared (adjusted for d.f.) = 95,1751 percent
Standard Error of Est. = 156,867
Mean absolute error = 108,095
Durbin-Watson statistic = 3,02633 (P=0,8307)
Lag 1 residual autocorrelation = -0,587808

The StatAdvisor
The output shows the results of fitting a linear model to describe the relationship between vydaje and pocet_clenu. The equation of the fitted model is

vydaje = 90 + 374,286*pocet_clenu

Stejně jako v ostatních případech používá Statgraphics k vyhodnocení testu hypotézy ukazatel P-Value (v jiných programech např. Significance Level apod.), což je maximální možná hodnota hladiny významnosti, na které ještě nezamítáme nulovou hypotézu. Není tudíž nutné určovat kritický obor pro námi zvolenou hladinu významnosti.

Pro odhady slouží okno Forecasts, které je v analýze také označeno jako Predicted Values. Za pomoci doplňkových možností Pane Options lze zadat jakoukoliv hodnotu pro námi požadovanou predikci. Procedura tvoří jak bodový odhad, tak i intervalové odhady pro průměrnou (Confidence) a individuální (Prediction) hodnotu.

Predicted Values

		95,00%		*95,00%*	
	Predicted	*Prediction*	*Limits*	*Confidence*	*Limits*
X	*Y*	*Lower*	*Upper*	*Lower*	*Upper*
4,0	1587,14	1113,84	2060,44	1401,87	1772,41

Odškrtnutím položky Include Constant lze vytvořit model přímky procházející počátkem. Jak bylo zmíněno v poznámce výše, občas je třeba tvořit modely bez použití konstantního členu, pokud je logicky takový model smysluplnější. Výsledný model zachycuje následující obrázek. K vystižení závislosti se jeví jako velice vhodný.

Simple Regression - vydaje vs. pocet_clenu
Dependent variable: vydaje
Independent variable: pocet_clenu
Linear model: Y = b*X
Number of observations: 6

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Slope	395,055	15,3905	25,6688	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1,42022E7	1	1,42022E7	658,88	0,0000
Residual	107775,	5	21554,9		
Total	1,431E7	6			

Correlation Coefficient = 0,996227
R-squared = 99,2469 percent
R-squared (adjusted for d.f.) = 99,2469 percent
Standard Error of Est. = 146,816
Mean absolute error = 114,194
Durbin-Watson statistic = 2,84997
Lag 1 residual autocorrelation = -0,603541

The StatAdvisor
The output shows the results of fitting a linear model to describe the relationship between vydaje and pocet_clenu. The equation of the fitted model is

vydaje = 395,055*pocet_clenu

Interpretace

Na 5% hladině významnosti bylo prokázáno, že model regresní přímky je statisticky vhodný k vystižení závislosti výdajů domácností na určitou komoditu a počtu jejich členů. Výsledná přímka má podobu výdaje = 90,000 + 374,286 počet členů. Při zvýšení počtu členů domácnosti o jednotku se zvýší výdaje průměrně o 374,286 Kč. Odhadované průměrné výdaje čtyřčlenné domácnosti činí 1587,14 Kč. Model vysvětluje asi z 96,14 % celkové variability (96,14 % celkové variability proměnné y je vysvětleno pomocí faktoru x).

Řešení v MS Excel

Dle verze programu se liší i možnosti zpracování regresní analýzy. Nejblíže výstupu ze Statgraphicsu je použití nástroje *Analýza dat*. Po kliknutí na toto tlačítko lze vybrat *Regrese*. V následujícím dialogu určíte pole se vstupními daty pro závislou a nezávislou proměnnou. Lze také definovat podobu reziduí. Nezapomeňte také určit pole pro výstup procedury (případně nový list).

Výstup je obdobný ANOVA Table a Coefficients Table ze Statgraphicsu. Variabilita je rozložena na modelovou (Regression) a reziduální (Residual). Kromě P-Value jsou zobrazeny i intervalové odhady parametrů regresní funkce.