Téma 5 – Příklad 3

Zadání příkladu

Tabulka obsahuje údaje o výkonu za směnu a procentu vadných výrobků u 10 dělníků. Zjistěte, zda je pro popis průběhu závislosti procenta vadných výrobků na výkonu za směnu vhodná parabola, nebo zda stačí použít regresní přímku.

Výkon	73	55	53	60	70	69	68	63	65	58
Procento	4,8	3,8	3,9	4	4,6	4,4	4,4	3,8	4	3,9

Vypracování příkladu

Postup odhadů parametrů je obdobný jako u předešlého případu. Základem bude využití metody nejmenších čtverců, tedy minimalizace druhých mocnin reziduí

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right)^2$$

Poté provedeme testy o statistické významnosti modelu regresní přímky $\eta = \beta_0 + \beta_1 x$ nebo regresní paraboly $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$. Pokud budou oba modely na námi zvolené hladině významnosti 5 % vhodné, učiníme rozhodnutí na základě matematicko-statistických kritérií o vhodnosti regresního modelu.

Metodou nejmenších čtverců byly spočteny následující rovnice regresních modelů:

Procento = 1,216 + 0,046*Vykon

Procento = 16,988-0,461*Vykon + 0,004*Vykon^2

Nyní ověříme vhodnost modelů celkovým F-testem.

1. H_0: $\beta_0 = c$ – zvolený model není pro vystižení závislosti vhodný
   $\beta_1 = 0$
   H_1: non H_0
2. Testové kritérium

$$F_{PŘ} = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}} = 25{,}40 \qquad F_{PAR} = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}} = 46{,}65$$

3. $W \equiv (F, F \geq F_{1-\alpha}(p-1; n-p))$
4. Testové kritérium je v obou případech prvkem kritického oboru. Nulovou hypotézu zamítáme, přijímáme hypotézu alternativní.

Oba dva modely jsou statisticky významné, musíme se proto rozhodnout na základě jiného kritéria. Jelikož má každý model jiný počet parametrů a některé ukazatele zvýhodňují modely

s vyšším počtem parametrů (v tomto případě parabolu), zvolíme k posouzení upravený index determinace.

$$I^2_{adj}=1-\left(1-I^2\right)\cdot\frac{n-1}{n-p}=1-\frac{\left(n-1\right)S_R}{\left(n-p\right)S_y}$$

Pro přímku je $I^2_{adj}=73{,}05$, parabola má $I^2_{adj}=91{,}03$. Značný rozdíl svědčí ve prospěch sice složitějšího modelu, ale modelu, který vysvětluje více než 90 % variability.

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)

Při vstupním dialogu zadáme jako závislou proměnnou y *procento* a jako proměnnou nezávislou x *vykon*.

V druhém kroku – Analysis options – vybíráme regresní funkci. Regresní přímka je označena jako model *Linear.*

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. Pro daný příklad uvažujeme o R-squared adjusted, jelikož porovnáváme modely s rozdílným počtem parametrů. V poli StatAdvisor můžeme ještě jednou vidět kompletní rovnici konstruovaného modelu.

Simple Regression - Procento vs. Vykon

Dependent variable: Procento
Independent variable: Vykon
Linear model: Y = a + b*X
Number of observations: 10

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	1,21555	0,587216	2,07002	0,0722
Slope	0,0464425	0,00921515	5,0398	0,0010

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0,885194	1	0,885194	25,40	0,0010
Residual	0,278806	8	0,0348508		
Total (Corr.)	1,164	9			

Correlation Coefficient = 0,872053
R-squared = 76,0476 percent
R-squared (adjusted for d.f.) = 73,0535 percent
Standard Error of Est. = 0,186684
Mean absolute error = 0,121423
Durbin-Watson statistic = 1,27798 (P=0,0981)
Lag 1 residual autocorrelation = 0,293257

Procedura v SGP: Relate – One Factor - Polynomial Regression (Analysis Summary)

Postup pro regresní parabolu je obdobný, procedura však umožňuje tvorbu jakéhokoliv polynomu, tudíž je třeba v druhém kroku Analysis Options zadat hodnotu řádu polynomu (Order) jako 2. Výstup procedury je jinak obdobný. Parabolický model má tři parametry.

Polynomial Regression - Procento versus Vykon

Dependent variable: Procento
Independent variable: Vykon
Order of polynomial = 2
Number of observations: 10

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	16,9883	3,83793	4,42643	0,0031
Vykon	-0,460925	0,123088	-3,74467	0,0072
Vykon^2	0,0040374	0,000978568	4,12583	0,0044

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1,08276	2	0,541379	46,65	0,0001
Residual	0,0812424	7	0,0116061		
Total (Corr.)	1,164	9			

R-squared = 93,0204 percent
R-squared (adjusted for d.f.) = 91,0262 percent
Standard Error of Est. = 0,107731
Mean absolute error = 0,0749265
Durbin-Watson statistic = 1,69967 (P=0,1752)
Lag 1 residual autocorrelation = 0,105963

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.

Interpretace

Na 5% hladině významnosti bylo prokázáno, že modely regresní přímky i regresní paraboly jsou statisticky významné a vhodné k vystižení závislosti procenta zmetků na výkonu za směnu. Pro vystižení závislosti byl nakonec vybrán model regresní paraboly, jelikož vysvětluje více než 90 % celkové variability.

Pozn.: Z didaktických důvodů nebylo v tomto případě použito modelů bez konstantního členu.

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.