

Cvičenie 1: Opakovanie 2

Naimportujte údaje do Gretlu a kvantifikujte vplyv jednotlivých vysvetľujúcich premenných na spotrebu piva. Výsledky Interpretujte!

V tabuľke boli uvedené vstupné údaje za nasledujúce premenné:

1. Q = litre spotrebovaného piva
2. PB = cena piva (\$)
3. PL = cena iných alkoholických nápojov (\$)
4. PR = cena ostatných tovarov a služieb (index)
5. I = príjem(\$)

Pre každú premennú bolo k dispozícii 30 pozorovaní. Údaje boli ročné a pochádzali z jednej domácnosti. Údaje sme najprv naimportovali do Gretlu a priradili sme im charakteristiku časového radu. Následne sme špecifikovali jednoduchý lineárny regresný model (Ordinary least square-metóda najmenších štvorcov) v ktorom Q bola závislá premenná a ostatné premenné boli nezávislé. Získali sme tým nasledujúci výstup:

Model 1: OLS, using observations 1983-2012 (T = 30)
Dependent variable: Q

	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	82,1587	17,9618	4,5741	0,00011	***
PB	-23,7426	5,42941	-4,3730	0,00019	***
PL	-4,07741	3,89049	-1,0480	0,30465	
PR	12,9243	4,1639	3,1039	0,00470	***
I	0,00199456	0,000775908	2,5706	0,01649	**
Mean dependent var	56,11333	S.D. dependent var	7,857381		
Sum squared resid	318,4831	S.E. of regression	3,569219		
R-squared	0,822118	Adjusted R-squared	0,793657		
F(4, 25)	28,88559	P-value(F)	4,77e-09		
Log-likelihood	-78,00374	Akaike criterion	166,0075		
Schwarz criterion	173,0135	Hannan-Quinn	168,2487		
rho	-0,313035	Durbin-Watson	2,508801		

Výstup je označený ako model 1 a je pri ňom uvedené že sme použili údaje ktoré tvorili časové rady za obdobie rokov 1983-2012, v každom časovom rade bolo 30 pozorovaní (ročné údaje). Závislá premenná bola Q.

Časť regresného výstupu v ktorej sú uvedené odhadnuté regresné koeficienty je veľmi podobná výstupu z excelu. V prvom stĺpci sú názvy jednotlivých vysvetľujúcich premenných (prvá premenná je konštanta). V druhom stĺpci (*coefficient*) sú hodnoty odhadnutých regresných koeficientov pre každú vysvetľujúcu premennú.

Lokujúca konštanta má hodnotu 82,1587 čo by sme mohli zinterpretovať ako priemernú hodnotu závisle premennej pri nulovej hodnote všetkých vysvetľujúcich premenných. Z logického hľadiska však takáto interpretácia neobstoí, pretože situácia s nulovou cenou ostatných nápojov, tovarov a služieb a nulovým príjmom je ťažko predstaviteľná.

Hodnota regresného koeficientu pri PB (cena piva) môže byť interpretovaná nasledovne. Nárast ceny piva o 1 \$ vyvolá v priemere pokles spotrebovaného množstva piva o 23,74 litrov za rok. Ostatné odhadnuté regresné koeficienty by sme mohli interpretovať obdobným spôsobom. (Treba dávať pozor na to, že ak by sme chceli porovnávať vplyv jednotlivých vysvetľujúcich premenných na závisle premennú, museli by sme model odhadnúť s použitím štandardizovaných premenných vzhľadom na to že sú jednotlivé premenné uvádzané v rôznych merných jednotkách).

Std. Error –štandardné chyby jednotlivých regresných koeficientov.

t-ratio –testovacia štatistika významnosti jednotlivých regresných koeficientov. Pre vyhodnotenie významnosti regresných koeficientov pomocou testovacej štatistiky by bolo potrebné najprv nájsť tabuľkové hodnoty t-rozdelenia s prislúchajúcimi stupňami voľnosti a následne ich porovnať s testovacími štatistikami.

p-value – p-hodnota –hodnota pravdepodobnosti pomocou ktorej vyhodnocujeme významnosť jednotlivých regresných koeficientov. Ak je p-hodnota nižšia ako 0,05 regresný koeficient je významný, ak je vyššia ako 0,05 koeficient je nevýznamný. Významnosť je nutné zinterpretovať pri každom odhadnutom regresnom koeficiente zvlášť.

Posledný stĺpec bez názvu, v ktorom sú uvedené len hviezdičky hodnotí samotnú významnosť koeficientov. Podľa tohto stĺpca ju vieme posúdiť dokonca aj bez hlbšej znalosti štatistiky.

*** -znamená vysoko významný regresný koeficient (na hladine významnosti $\alpha=0,01$, to znamená že prislúchajúca p-hodnota je nižšia ako 0,01)

** -významný koeficient (na hladine významnosti $\alpha=0,05$, teda prislúchajúca p-hodnota je nižšia ako 0,05 ale vyššia ako 0,01)

* - koeficient je významný na hladine 0,1, teda prislúchajúca p-hodnota je nižšia ako 0,1 ale vyššia ako 0,05 (teda na hladine významnosti $\alpha=0,05$ je koeficient už nevýznamný)

Z výstupu vyplýva, že okrem premennej PL sú všetky odhadnuté regresné koeficienty štatisticky vysoko významné na hladine významnosti $\alpha=0,01$ a premenná I je významná na hladine významnosti $\alpha=0,05$.

Viac o štandardných chybách a testovaní významnosti je uvedené pri vypracovaní prvého príkladu z cvičenia č. 1.

Všimnite si, že na rozdiel od regresného výstupu z MS Excel, výstup Gretlu nezobrazuje intervaly spoľahlivosti. Tieto je možné nájsť dodatočne v hornom menu v záložke *Analysis* pod voľbou *Confidence Intervals for Coefficients*. Získali by sme nasledujúci výstup:

$$t(25, 0,025) = 2,060$$

Variable	Coefficient	95 confidence interval
const	82,1587	(45,1658, 119,152)
PB	-23,7426	(-34,9247, -12,5605)
PL	-4,07741	(-12,0900, 3,93521)
PR	12,9243	(4,34864, 21,5000)
I	0,00199456	(0,000396546, 0,00359257)

V tomto výstupe je najprv uvedená tabuľková hodnota t pre prislúchajúce stupne voľnosti a 5% hladinu významnosti. Ďalej sú znovu zopakované názvy jednotlivých premenných, odhadnuté regresné koeficienty a 95% intervaly spoľahlivosti pre každú premennú. (V prípade že by sme chceli zmeniť hladinu významnosti, klikneme na ikonu s obrázkom alfy v ľavom hornom menu a vyberieme si vyberieme si inú hladinu významnosti.) Podobne ako v Exceli a ich interpretácia je tiež totožná s tým čo sme uviedli pri interpretácii výstupu z MS Excel.

Ďalšia časť regresného výstupu programu Gretl je v porovnaní s výstupom v programe Excel obširnejšia s tým že v nej chýba tabuľka ANOVY. Ak nás táto tabuľka zaujímala (rozklad variability na vysvetlenú a nevysvetlenú) nájdeme ju v hornom menu na lište okna s regresným modelom, v záložke *Analysis* pod voľbou *ANOVA*. Výstup ANOVY k regresii v Gretli je nasledovný:

Analysis of Variance:

	Sum of squares	df	Mean square
Regression	1471,93	4	367,983
Residual	318,483	25	12,7393
Total	1790,41	29	61,7384

$$R^2 = 1471,93 / 1790,41 = 0,822118$$

$$F(4, 25) = 367,983 / 12,7393 = 28,8856 \text{ [p-value } 4,77e-009]$$

Podobne ako vo výstupe regresie z Excelu je v tabuľke uvedená vysvetlená (Regression), nevysvetlená (Residual) a celková variabilita (Total). Pri každej sú uvedené stĺpce so Sumou štvorcov (Sum of squares), stupňami voľnosti (df) a priemerom štvorcov (Mean Square). Viac ku každému môžete nájsť vo vypracovaní k prvému príkladu cvičenia č. 1.

Pod tabuľkou Anovy je uvedený výpočet a hodnota koeficientu determinácie, výpočet testovacej štatistiky F testu pre otestovanie celkovej významnosti modelu ako aj p -hodnotu pomocou ktorej je možné vyhodnotiť celkovú významnosť modelu. Podobne ako pri hodnotení významnosti jednotlivých regresných koeficientov aj tu platí, že ak je p -hodnota nižšia ako 0,05 model ako celok je významný, ak je to naopak model ako celok je nevýznamný.

Vrátime sa znovu k pôvodnému regresnému výstupu programu Gretl. Pod tabuľkou s odhadnutými regresnými koeficientmi sú uvedené štatistiky pričom Gretl pri regresii uvádza viac štatistík ako je tomu pri výstupe v Exceli. Sú uvedené nasledujúce štatistiky:

Mean dependent var – stredná hodnota závisle premennej

S.D. dependent var – štandardná odchýlka závisle premennej

Sum squared resid – suma štvorcov reziduí –nevysvetlená variabilita (rovnaké číslo ako v tabuľke anovy pri ss residual)

S.E. of regression – Štandardná chyba regresie –niekedy používaná aj ako miera presnosti modelu, štandardná chyba odchýlok skutočných hodnôt od hodnôt predikovaných modelom (viac ako výpočte a interpretácii štandardnej chyby modelu je možné nájsť vo vypracovanom prvom príklade k prvému cvičeniu)

R-squared – Koeficient determinácie –podiel variability závisle premennej vysvetlenej modelom, t.j. V tomto prípade sme regresným modelom vysvetlili 82,21% variability závisle premennej

Adjusted R-squared – korigovaný koeficient determinácie, používa sa pri porovnávaní modelov s rôznym počtom vysvetľujúcich premenných. Zohľadňuje počet premenných v modeli. Je vždy nižší ako pôvodný (nekorigovaný) nekorigovaný koeficient determinácie.

F(4, 25) – testovacia štatistika celkovej významnosti modelu, vo výstupe programu MS Excel bola uvádzaná v tabuľke ANOVY

P-value(F) – p-hodnota pre testovanie celkovej významnosti modelu (resp. hodnota pravdepodobnosti že náhodná premenná nadobudne hodnotu F a vyššiu). Pokiaľ je P-hodnota nižšia ako 0,05 model ako celok je významný, ak je p-hodnota vyššia ako 0,05 model ako celok je nevýznamný.

Log-likelihood – v niektorých prípadoch sa uvádza ako LnL, hodnota funkcie vierohodnosti, používa sa pri porovnávaní viacerých modelov. Pri výbere z viacerých modelov si vyberieme model pri ktorom je táto hodnota vyššia.

Akaike criterion- Akaikeho informačné kritérium, používa sa pri vyhodnocovaní kvality modelu, okrem podielu vysvetlenej variability zohľadňuje aj komplexnosť modelu. Pri aplikácii Akaikeho informačného kritéria pri výbere najlepšieho modelu z viacerých možností si vyberáme model kde tento indikátor dosahuje čo najnižšie hodnoty. Podobne je tomu aj pri Schwartzovom kritériu a Hannah-Quinn ktoré sú tiež uvedené v regresnom výstupe.

Posledné dve štatistiky sú uvádzané len vo výstupe z regresíí kde boli ako premenné použité časové rady.

Rho – autokorelačný koeficient prvého rádu

DW – štatistika Durbin-Watsonovho testu, používa sa pri posudzovaní prítomnosti autokorelácie v modeli. Keďže autokorelácia sa vyskytuje spravidla pri časových radoch, v prípade použitia prierezových údajov Gretl tieto posledné dve štatistiky vo výstupoch neuvádza.

Na záver regresného výstupu je uvedené že premenná pri ktorej p-hodnota dosahovala najvyššiu úroveň bola PL, táto premenná je teda spomedzi všetkých najmenej významná. V tomto prípade je jej p-hodnota vyššia ako 0,05, teda je nevýznamná. Spolu s vysokou hodnotou koeficientu determinácie to môže svedčiť o prítomnosti multikolinearity v modeli ale tomu sa budeme venovať až na ďalších cvičeniach.

Vypracoval Jozef Palkovič