POSUDEK OPONENTA NA BAKALÁŘSKOU PRÁCI

LENKA MALÍNSKÁ: STATISTICKÁ ANALÝZA DAT TÝKAJÍCÍCH SE NEZAMĚSTNANOSTI

Autorka na datech po roce 2000 zkoumá pomocí základních metod korelační a regresní analýzy vybrané souvislosti mezi veličinami týkajícími se odvětví výroby automobilů, HDP a zaměstnaností.

Splnění cílů práce

- ☐ nadstandardně
- ☐ velmi dobře
- ☐ splněny
- ☑ s výhradami
- ☐ nebyly splněny

Odborný přínos práce

- ☐ nové výsledky
- ☑ netradiční postupy
- ☐ zpracování výsledků z různých zdrojů
- ☐ shrnutí výsledků z různých zdrojů
- ☐ bez přínosu

Matematická (odborná) úroveň

- ☐ vynikající
- ☐ velmi dobrá
- ☐ průměrná
- ☑ podprůměrná
- ☐ nevyhovující

Věcné chyby

- ☐ téměř žádné
- ☐ vzhledem k rozsahu přiměřený počet
- ☐ méně podstatné, větší množství
- ☐ podstatnější, větší množství
- ☑ závažné

Grafická, jazyková a formální úroveň

- ☐ vynikající
- ☐ velmi dobrá
- ☑ průměrná
- ☐ podprůměrná
- ☐ nevyhovující

Text začíná popisem použitých dat a přehledem charakteristik (jen) některých z nich. V dalších částech autorka aplikuje model přímkové regrese a testuje významnost regrese resp. korelace. Vyšetřuje tři souvislosti:

(a) v částech 4.1–4.3 pro ČR souvislost mezi HDP na obyvatele a „produkcí aut“ na obyvatele, případně s vyloučením vlivu času a dále také zvlášť pro období před krizí a během krize,

(b) v částech 4.4 a 4.5 pro EU a Německo souvislost mezi ročními změnami zaměstnaností v automobilovém průmyslu a mimo něj, zde počítá i intervalové odhady pro regresní hodnotu a předpověď,

(c) v části 4.6 pro EU souvislost mezi produkcí osobních a nákladních aut, doplněnou o „ověřování“ některých z předpokladů modelu.

Práce a zkoumané souvislosti při zběžném prohlédnutí textu působí pěkným dojmem. Při pozornějším čtení ale narazíme na drobné nepřesnosti a při detailním zkoumání výpočtů zjistíme i skryté a zásadní chyby:

- chyby při základním zpracováním dat: při počítání průměrů a mediánů v tabulkách v kapitole 3 jsou nedostupná data započtena jako nuly, v posledním sloupečku není uveden průměr nebo medián ukazatele z jednotlivých let, nýbrž podíl mediánů či průměrů veličin z čitatele a jmenovatele, u dat o zaměstnanosti se ve skutečnosti nejedná o průměr dat z roků uvedených pod

tabulkou, ale jen do roku 2012, ve výpočtu pro obr. 3.1 se u „produkce aut“ do čitatele dosazuje hodnota ne z daného roku, ale hodnota o 2 roky starší,
- po prozkoumání přiložených souborů konstatuji, že *veškeré* výpočty v (a) *ve skutečnosti bohužel nejsou provedeny s indexem výroby automobilů*, ale zřejmě s indexem výroby ze všech odvětví,
- ověřování předpokladů modelu v (c) je špatně: dělat t-test 4.6.1 nedává žádný smysl (součet reziduí bude v daném modelu vždy nulový), způsob provedení výpočtu Goldfeldova-Quandtova testu je zcela nesmyslný, také jeden z předpokladů regresního modelu není vůbec zmíněn.

Práce obsahuje řadu dalších nepřesností. V obr. 3.1 chybí údaje vždy pro jeden z avizovaných roků, navíc růst v čase nelze z grafu vyčíst (nevíme, kde je začátek). Odlehlá pozorování podle rovnosti průměru a mediánu (str. 7)? Nevysvětlená označení, neúplný popis modelu, není řečeno, proti jaké alternativě byl test prováděn (str. 9), nepřesná vyjádření (korelační koeficient vs. nezávislost, resp. neuvedení předpokladů — str. 10 dole). Ve vzorci pro t na str. 11 se pro daný model má odečítat 3, na str. 14 jsou x_i a y_i popsána opačně. Na str. 16 není řečeno, co jsou b_i, a formulace o 95 % je nepřesná. Ztotožnění intervalů spolehlivosti pro parametry s SE na str. 17 jsem nepochopil, podobně tabulky tamtéž. Co mají být μ, x, μ_0, s na str. 19? A co s^2_{R1}, H_0, H_1 a q v 4.6.1? Nejasná věta „Hodnoty porovnáváme...“ na str. 20, v literatuře je nedostatečně specifikována položka Novák.

Hodnotu textu dále snižuje, že autorka nepopisuje dostatečně, s jakými daty pracuje a jaké s nimi provádí operace. Úvodní popis dat není úplný, v práci ve skutečnosti používá i další zdroj dat, v textu nezmíněný. Navíc organizace přiložených souborů není optimální. V textu není popsáno, co se myslí změnou zaměstnanosti v (b), svůj popis by si zasloužila i produkce automobilů. Matoucí je 2003 = 100 v obr. 3.2 (ve skutečnosti zde 2003 Q1 = 100), nalezneme také rozpory mezi popisem, jaká data jsou k dispozici, a obdobími, pro které je výpočet proveden. Tabulky 3.1 a 3.2 přesahují daleko zrcadlo sazby. Obrázky mají v různých částech práce různý formát. Je škoda, že na osách grafů v (b) a (c) nejsou označeny žádné hodnoty.

Nabízejí se i obecnější otázky. Proč se zpracování z (a) rozdělilo na před/v krizi a po krizi se nezkoumalo? Proč se o klamné korelaci uvažovalo pouze v (a), intervalové odhady počítaly zrovna v (b) a předpoklady ověřovaly v (c)? Proč se v (a) zkoumá ČR a jinde EU?

Myšlenka práce se mi líbí, předložené podání mnohem méně. Za předpokladu, že studentka při obhajobě prokáže pochopení metod použitých v práci, hodnotím práci ještě známkou *dobře*.

Otázka do rozpravy:
- Neměla by hodnota indexu produkce aut celkem v obr. 3.2 vždy ležet mezi dvěma ostatními? Nebo je v pořádku, že neleží?

Michal Friesl

Plzeň, 13. srpna 2015.