Historická sociologie, Řízení a supervize FHS UK: diplomky, AKD3 (i AKD1 a 2)

**Praktické rady studentům k presentaci dat (nejen) z regresních modelů v textu např. diplomové práce**

**Jiří Šafr**, FHS UK & SOÚ AV ČR, v.v.i.

**Průběžně doplňováno a aktualizováno!** poslední aktualizace: **5/5/2017**; 21/6/16

### Co a jak presentovat v textu

Důležité, nikdy nedělejte tu chybu, že byste pouze vložili tabulku či graf a nenapsali k tomu byť kratinkou interpretaci, nebo alespoň nějakou poznámku. Takže **žádné „osamocené“ tabulky**.

Také začínat odstavec „Tady vidíte tabulku … „ bez toho, aniž byste nejprve uvedli, o co vám v dalším oddílu půjde na žádné tabulky neodkazujte, a už vůbec nepište nic neříkající balast (to dá rozum, že to co vidím je tabulka…).

Úplně nejhorší pak je popisovat, že jste něco (zcela samozřejmého, jako třeba tabulku) udělali. Nikdy **nepište takovéto věcně nic neříkající věty**: „Pak jsem udělal tabulku pro věkové kategorie a pohlaví, k čemuž jsem použil program MS Excel…“. To ale neznamená, že při použití složitějších analýz postup či metodu nepopíšete (ale stejně z toho nikdy nedělejte středobod textu, tedy pokud nepíšete metodologický článek.).

Vždy je správné čtenáře provázet věcným obsahem, a ostatní upozaďovat (odkazy do textu, opakující se nic neříkající slova atd.).

Příklad:

*„Dále jsme* *vytvořili* tabulku 1, kde jsou výsledky pro spokojenost se životem *podle pohlaví*. Dále se můžeme podívat na tabulku 2, kde je spokojenost se životem podle vzdělání.“ → **chybí interpretace!**, text je jen o konstrukci tabulek, čtenář se nic nedozvídá. → takže lépe třeba takto:

Dále nás bude zajímat, zda jsou se životem spokojenější ženy nebo muži. Porovnání průměrných hodnot ukazuje, že ženy jsou výrazně spokojenější, jejich průměrný pocit odpovídá hodnotě 7, zatímco u mužů pouze 4, což na škále 1-10 již naznačuje pro muže určitou míru nespokojenosti se životem (viz tabulka 1). Nicméně, rozrůzněnost odpovědí je u mužů výrazně vyšší než u žen (směrodatná odchylka pro muže dosahuje hodnoty 3,5 pro ženy pouze 2,1), což znamená, že v populaci najdeme jak výrazně spokojené tak výrazně nespokojené muže, zatímco ženy hodnotí spokojenost se životem mnohem podobněji.

Jiný příklad:

„Proměnné jsem pak pro lepší přehled uspořádal do kontingenční tabulky, kde je klesající "třídní" kongruence zdůrazněna silnějším písmem.“ → lépe: Vztah „třídní“ kongruence dokumentuje tabulka XY, kde je patrný trend ….

### Formulace v textu – čeho se vyvarovat

#### Často se opakující chyby:

- znak procent je v češtině oddělen od čísla mezerou, pokud se nejedná o přídavné jméno (např. „pětiprocentní“ rozdíl)

- nezaměňujete různé pojmy, např. : „fenomén“ a „proces“

Vyvarujte se následujících **chybných chápání významů**, **nic neříkajících banalit** a vyslovených nesmyslů (sbírám opakující se chybné formulace průběžně z diplomek a seminárek):

* *analýza dat* = proces, proto analýza sama nic „nevybádala“, ta přináší jen výsledky (takové či jinaké). A pokud v textu „analyzujete“, tj. používáte stále ta samá data (v našem případě z nějakého kvantitativního výzkumu), pak slovo „data“ ve spojení s analýzou není třeba stále opakovat.

Příklad: *Z analýzy dat se dá vysledovat*, že vzdělání otce hraje během let 1992 - 2009 postupně čím dál důležitější roli … → Z výsledků analýzy se dá vysledovat, že vzdělání otce hraje….

* *Respondent* = nenadužívejte tohoto docela nelibého a nic neříkajícího slova. Je to bohužel častý nešvar českých sociologů. Přeci smyslem dělání kvantitativního výzkumu je činit závěry z malého reprezentativního vzorku na celou populaci (vzpomeňte na principy inferenční statistiky). A **právě ta (celá) populace je váš referenční rámec (nikoliv výběr)**. Takže máte-li data z reprezentativního výběru (to zajistí správně provedený náhodný výběr podle určitých pravidel) z jasně definované (!) populace, pak se můžete vztahovat k této populaci či jejich zástupcích. Když například máte výzkum mezi respondenty seniorského věku to znamená, že píšete o seniorech, tak používejte pojem „senior“ (nebo jiné synonymum), když o žácích ZŠ, tak „žák“, atd. Univerzálně můžete mluvit třeba o jedincích (z dané populace), ale s „jedincem/i“ to raději nepřehánějte, někdy to v textu je samý „jedinec“ a pak to vypadá jak s těmi „respondenty“. Nezapomínejte ale, že i zdánlivě jasně ohraničená a definovaná populace nemusí být na první pohled jasná a čtenáře můžete mást, například označení „česká populace“, může být v datech ve skutečnosti jednou např. „ekonomicky aktivní ve věku starším 18-ti let v roce 1998“ ale také „obyvatelstvo starší 16-let v roce 2003“. Minimálně v metodologickém úvodu (sekce o datech) to musí být jasně a maximálně přesně definováno – za jakou populaci výběrový soubor vypovídá, i když v textu třeba budete průběžně používat různé obecnější pojmy. Nejde jen o zdánlivá slovíčka, **v názvu tabulek a grafů proto populaci vždy jasně definujte**.  
  Existuje ale jedna situace v textu, kde se naopak čtenáře použitím slova „respondent“ dá dobře upozornit na určitý problém – bývá to tam, kde chceme zdůraznit, že to co právě popisujeme, nejsou žádná „objektivní fakta“, **ale pouze(!) výpovědi lidí**. Tím také zdůrazňujeme, že v tom co presentujeme existuje určité omezení vyplývající z principů a zejména limitů dotazníkového šetření. Často se také v této souvislosti používá výraz „respondenty deklarované …“. Příkladem budiž vnímaný či uváděný „pocit chudoby“.

#### K definování cílů analýzy / práce

**Cílem celé práce (textu) by mělo být přispět řešení nějakého obecnějšího problému**, tj. odpovědět na nějakou obecnou otázku. Dodávám, otázku netriviální. Ostatně už proto říkáme, že definice výzkumného problému je v podobě formulace otázky. Například: Zvyšuje ekonomický kapitál rodiny šanci na úspěch žáka ve škole? *Cílem textu*je přispět k poznání tohoto problému z hlediska něčeho v kontextu toho a toho. A pokud by otázka byla snad pro někoho na první pohled triviální, tak jí musíte nastínit v novém světle, tak abyste ukázali, že se na problém dá dívat nějak jinak (a že tam tedy nějaký problém hodný řešení je). To ale musíte v textu pečlivě vyargumentovat.

Čeho se určitě vyvarujete je aby cílem práce byla „analýza“. Například: „Analýza vybraných dat a interpretace výsledků je tedy vlastně druhý cíl této práce.“ **Analýza dat sama o sobě přeci nemůže být cílem práce**. Ta je prostředkem k dosažení nějakého cíle (sociologického či jiného).

### Jak mít v textu tah na branku?

V práci (textu) musíte mít nejen (obecné) téma, to mít nějak konceptualizované, ale také musíte mít problém, kterých chcete řešit – otázku, na níž hledáte odpověď a na níž také tu odpověď v závěru čtenáři předložíte a to na základě výsledků analýzy a to pouze vlastní produkce (včetně) použitých dat.

Nepište „dělal jsem analýzu něčeho …“, ale „Výsledky analýzy ukazují, že …“ tj. nepopisujte, co jste dělali/děláte, ale k čemu jste došli.

„Ověřovala jsem platnost obecné hypotézy, zda došlo ke změně ve vlastnostech digitální generace Y …“

Zajímavá hypotéza není ZDA ale JAK. Takže správněji:

„Ověřovala jsem platnost obecné hypotézy, podle níž došlo ke změně *ve vlastnostech* digitální generace Y…“ (určitě by to šlo lépe, konkrétně v jakých vlastnostech)

### Praktické (formátová apod.)

- do wordu (formáty doc, docx) vkládat grafy jako **vektorovou** grafiku *EMF / WMS*, když je nejhůř tak alespoň jako *gif* a když už je úplně nejhůř tak *png*, ale **nikdy ne *jpg*** (tento formát je určen na fotky ne na grafy! (a právě proto je rozostřuje, aby to oko nepoznalo)).

- ideálně **číslujte tabulky a grafy** pomocí **dvojúrovňového** kódu, kde *první číslice odpovídá číslu kapitoly, druhá pořadí* tabulky / grafu v této kapitole. Velkou výhodou je, že pokud přidáte či odeberete tabulku vprostřed textu nebudete muset přečíslovávat celou diplomovou práci.

### Jak presentovat regresní modely

Co uvádět v textu z výstupů regresních modelů? Výstup z programu (SPSS) je třeba vždy upravit, především **zjednodušit**, některé statistiky se dublují, některé jsou zbytečné, čtenář potřebuje přehledně a stručně presentované výsledky, **zároveň ale vědecké dílo vyžaduje i doprovodné informace o přesnosti měření, důvěryhodnosti modelu** atp.

Takže, pokud jde o **lineárně regresní model či analýzu rozptylu**, ideálně uvádíme **nestandardizované regresní koeficienty (B)**, které přináší metrikou informaci (tam kde to dává smysl, především pokud jde o závislou proměnnou), na základně níž lze provést nejen predikci, ale i zhodnotit velikost účinku a výsledky věcně interpretovat. Dále nesmíme zapomenout na **velikost jejich chyby**, tj. přesnost jednotlivých regresních koeficientů v našem modelu. Minimální a základní standard je uvádět **standardní chyby koeficientů (SE)**, což má krom testu resp. posouzení jejich nenulovosti (v celé populaci) ještě jeden význam – pomocí nich lze totiž dodatečně spočítat testy rozdílů mezi jednotlivými regresními koeficienty a to jak v rámci jednoho modelu, tak i mezi postupně na sebe navazujícími tzv. zahnízděnými modely (a také intervaly spolehlivosti). Na což by měl čtenář mít právo, i když jste sami tento postup nepoužili. Někteří autoři (především ti, co jsou statistici) zde končí, více prostě netřeba, vše ostatní si lze spočítat. Přesto to není pro méně poučeného čtenáře komfortní, proto uvádíme buď *t-hodnotu* a nebo dosaženou statistickou signifikanci.

Nejvíce informací přináší v jednom čísle (proto stačí i místo chyby SE) **hodnota *t***, což je podíl koeficientu k jeho st. chybě (b/SE) a to již stačí k rychlému zhodnocení statistické významnosti koeficientů. Protože (skoro) každý ví, že jejich hodnota odpovídá *t* hodnotě (resp. ve velkých výběrových souborech hodnotě *z*), takže tam kde je *t* > 2 víme, že příslušný koeficient je statisticky významně odlišný od nuly na hladině spolehlivosti min. 0,05. Navíc z t-hodnoty si lze s trochou nepřesnosti dopočítat zpátky i SE. Uvádět lze i samotnou *p-hodnotu* testu, ale to bych pak doporučoval *doplnit o SE*. Statisticky významné výsledky pro dané koeficienty pak můžete nějak zvýraznit, například je uvést tučným písmem, zvýraznit pozadí buňky v tabulce, a nebo doplnit obligátními hvězdičkami jejichž počet odpovídá dosažené hladině významnosti v dílčích textech (pro *sig*.: 0,05 → \* 0,01 → \*\* 0,001 \*\*\*). Rozhodně ale **nikdy neuvádějte jen samotné hvězdičky odpovídající dosažené hladině významnosti** v dílčích testech nenulovati regresních koeficientů. V nich se ztrácí mnoho informací, je rozdíl, zda je výsledek významný na *p*=0,011 nebo 0,046, přitom v obou případech přiřadíte jednu stejnou \*. Navíc hvězdička sama o sobě nic moc neříká, a nelze s ní dále pracovat (např. testovat rozdíly koeficientů).

Co dále musíte **u modelu jako celku** uvést? Krom parametrů a jejich chyb, je třeba uvádět vydatnost – přesnost celého modelu. Celkový test rozptylu F-test se většinou neuvádí, většinou ani model nepresentuje, pokud nám nevyjde (tj. nezamítneme nulovou hypotézu). Určitě uvádějte **koeficient determinance R2** a pokud máte složitější (zahnízděné) modely, které mezi sebou porovnáváte tak i jeho adjustovanou verzi **Adj. R2**. A také ho řádně interpretujte. Strategie budování modelu může vycházet právě z porovnání vysvětlené variance mezi různými modely (doplněného také o příslušný F-test jejich rozdílu), kdy například přidáváme k základnímu modelu klíčovou proměnnou reprezentující určité teoretické východisko. K dobrým mravům také patří uvést **počet platných případů použitých v modelu**. To je důležité, protože většinou používáme metodu totálního průniku platných případů (listwise) a pak se velikost našeho výběrového (pod)souboru při větším počtu nezávislých proměnných nezřídka oproti výsledkům bivariátních analýz podstatně zmenší (je-li v dílčích bivariátních analýzách na ignorovatelné úrovni pod cca 5%, tak ve vícerozměrné analýze může být počet vynechaných případů i mnohem vyšší). Otázkou je, zda pak ještě lze hovořit o stejném výběru resp. prezentujeme výsledky z bivariátních a vícerozměrných analýz na stejné části populace. (Vždyť za neochotou odpovídat může být nějaká neměřená vlastnost, která ale může ovlivňovat také závislou proměnnou). Zde je pak na zvážení, zda nebudeme chybějící hodnoty imputovat. Nikdy ale nepoužívejte samotný průměr/medián (těm co neodpověděli)! Donedávna běžnou metodou bylo u té proměnné/ých, kde chybí nejvíce případů, nebo je to u ní/nich důležité sledovat, imputovat nejčastěji průměrnou hodnotu a zároveň do modelu vstoupit s dummy indikátorem ukazujícím na to, že hodnota chyběla. Pokud u něj zjistíme statisticky významný vliv, pak proces vzniku chybějících hodnot ovlivňuje závislou proměnnou. Ale zas tak jednoduché to není a proto je lepší používat nějaké moderní metody vícenásobné imputace chybějících hodnot (multiple imputation).

Co bychom dále měli? Pokud porovnáváme jednotlivé regresní koeficienty mezi různými zahnízděnými modely, měli bychom **testovat jejich rozdíl**. Pokud přidáváte do modelu jen jednu proměnnou, tak to můžete prostřednictvím příslušného t-testu u koeficientu resp. F-testem modelu jako celku (model s a bez). V případě složitějších modelů, musíte použít další testy, například Waldův test a nebo lépe likelihood ratio test nebo F-test a zároveň je také vhodné využít informační kritéria (AIC, BIC), která zohledňují úspornost (parsimonii) modelu.

<<pokračování někdy příště>>