VYTVÁŘENÍ *DUMMY* PROMĚNNÝCH A JEJICH UŽITÍ V REGRESNÍ ANALÝZE

Do regresní analýzy nemohou vstupovat proměnné, které jsou nominální nebo ordinální (s krátkými stupnicemi). Pokud je závisle proměnná jiná než intervalová (kardinální), nelze regresi použít. Pokud je ale nezávisle proměnná jiné povahy, než kardinální, můžeme tento požadavek obejít. Víme totiž, že jako nezávisle proměnná může do regresní analýzy vstupovat proměnná dichotomická. A dichotomickou proměnnou lze vytvořit z každé proměnné – navíc poměrně jednoduše. Např. z proměnné sociální třída, která měla následující stupně měření: 1 nižší, 2. dělnická, 3. střední, 4. vyšší bychom mohli vytvořit dichotomii tak, že proměnnou rekódujeme na: 1. nižší třída, 0 všichni ostatní. Nebyl by to ale dobrý postup, neboť ve sběrné kategorii ‚ostatní' by byly velmi odlišné a heterogenní typy respondentů.

Lepším postupem je, že dichotomické proměnné vytvoříme z jednotlivých kategorií naší proměnné, takže se původní proměnná přemění na sérii dichotomií. Hovoříme o vytváření tzv. *dummy* (pomocných)[1] proměnných. S dummy proměnnými pak můžeme vstoupit do lineární regrese a vypočítat vliv každé kategorie na závisle proměnnou.[2]

Dummy proměnné lze vytvořit z jakékoliv nominální nebo ordinální proměnné. Postup, jak na to, si ukažme na příkladu. Provedeme na datovém souboru z výzkumu Akceptace populačních politik II (*Population Policy Acceptance* – PPA II), který byl v rámci mezinárodního srovnávacího výzkumu v ČR proveden v roce 2001.

Příklad: Vzdělání měřené podle dosažených stupňů je klasická ordinální proměnná: 1. základní, 2. vyučen/a, 3. SŠ, 4. VŠ. Ta není vhodná do regresní analýzy. Změňme ji tedy na sadu dichotomických proměnných, které vzniknou z jejích jednotlivých kategorií (viz tab. 10.9). Zapamatujme si, že počet dummy proměnných musí být o jednu menší, než je počet kategorií přetvářené proměnné. Je to proto, že z poslední kategorie vytváříme tzv. **referenční kategorii**, s níž výsledky porovnáváme. A dále si zapamatujme, že dichotomie, které vstupují do regrese je výhodné kódovat 0 a 1.

Tab. 10.9: Vytváření dummy proměnných z kategorií proměnné vzdělání

Kategorie	Jméno proměnné	Hodnoty proměnné
VŠ	Vzd_VS	1 = VŠ, 0 = ostatní
SŠ	Vzd_SS	1 = SŠ, 0 = ostatní
Vyučen/a	Vzd_vyuc	1 = vyuč., 0 = ostatní
Základní	Referenční kategorie	

Jako nové proměnné budou mít následující kódy:

Kategorie	Vzd_VS	Vzd_SS	Vzd_vyuc
VŠ	1	0	0
SŠ	0	1	0
Vyučen/a	0	0	1
Základní	0	0	0

Nové proměnné vytvoříme nám již známým postupem prostřednictvím procedury *TRANSFORM – RECODE – Into Different variables (syntax):*

[1] *Dummy* je anglicky atrapa nebo maketa nebo také předstíraný či fingovaný. V našem kontextu navrhuji používat adjektivum ‚pomocný'.

[2] Tímto krokem se de facto jednoduchá regrese mění na regresi vícenásobnou. Ale s tím si nedělejme starost.

```
RECODE
  vzd_resp
  (4=1)  (ELSE=0)  INTO  vzd_VS .
VARIABLE LABELS vzd_VS 'dummy'.
VALUE LABELS vzd_VS 1 'VS' 0 'ostatni'.
EXECUTE .
RECODE
  vzd_resp
  (3=1)  (ELSE=0)  INTO  vzd_SS.
VARIABLE LABELS vzd_SS 'dummy'.
VALUE LABELS vzd_SS 1 'SS' 0 'ostatni'.
EXECUTE .
RECODE
  vzd_resp
  (2=1)  (ELSE=0)  INTO  vzd_vyuc .
VARIABLE LABELS vzd_vyuc 'dummy'.
VALUE LABELS vzd_vyuc 1 'vyuc' 0 'ostatni'
EXECUTE .
```

Výstupy:

Toto je rozložení původní proměnné vzdělání respondenta.

VZD_RESP vzdelání respondenta

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 základní	116	17,5	17,6	17,6
	2 vyuc bez mat.	250	37,6	37,8	55,5
	3 SŠ	229	34,5	34,8	90,2
	4 VŠ	65	9,7	9,8	100,0
	Total	660	99,4	100,0	
Missing	System	4	,6		
Total		664	100,0		

Dummy proměnná pro kategorii vysokoškolské vzdělání:

VZD_VS dummy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 ostatni	599	90,3	90,3	90,3
	1 VS	65	9,7	9,7	100,0
	Total	664	100,0	100,0	

Dummy proměnná pro kategorii středoškolského vzdělání:

VZD_SS dummy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 ostatni	435	65,5	65,5	65,5
	1 SS	229	34,5	34,5	100,0
	Total	664	100,0	100,0	

Dummy proměnná pro kategorii vzdělání vyučen/a:

VZD_VYUC dummy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 ostatni	415	62,4	62,4	62,4
	1 vyuc	250	37,6	37,6	100,0
	Total	664	100,0	100,0	

Počty jednotlivých kategorií odpovídají původnímu rozložení v proměnné vzdělání, *recode* tedy proběhl úspěšně. Tato sada proměnných nyní může vstoupit do konkrétní úlohy regresní analýzy. Zajímá nás nyní, jak vzdělání ovlivňuje záměr žen mít v budoucnu dítě. Analýza je provedena na podsouboru českých žen ve věku 18-49 let, data opět pocházejí z reprezentativní šetření PPA II. Podívejme se nejdříve, jak je rozložena závisle proměnná a jaká je mezi těmito dvěma proměnnými korelace:

Frekvence:

CF1_3 Pocet detí

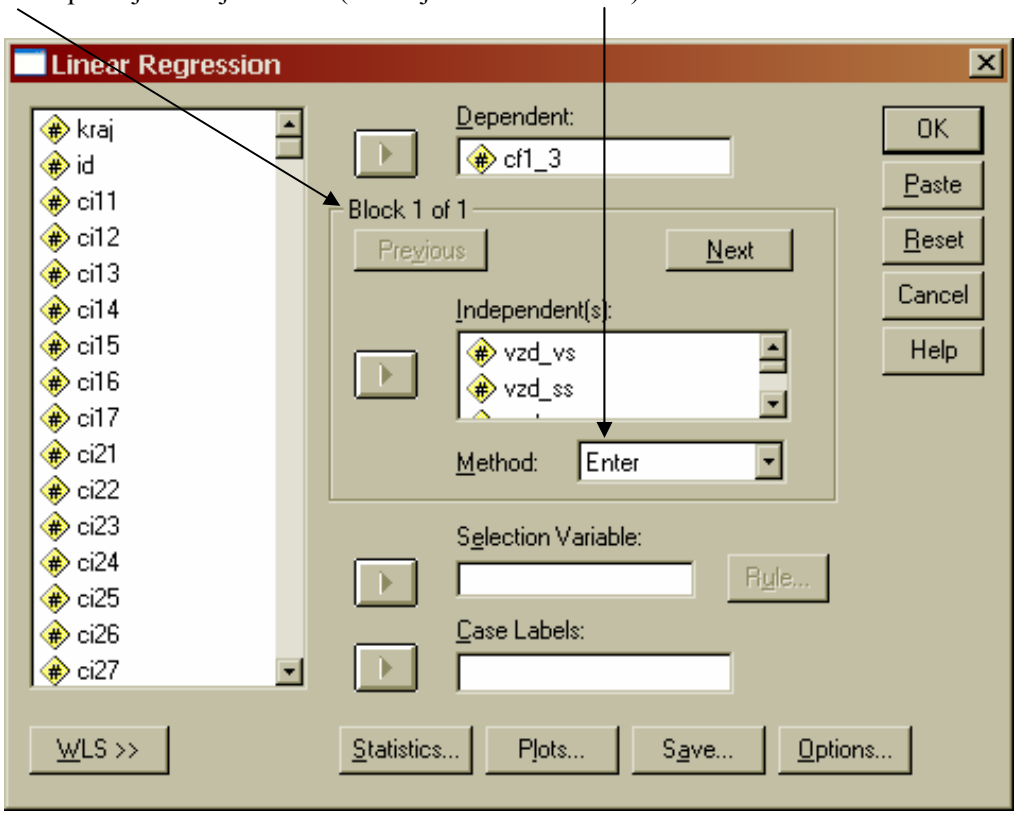
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 0	438	66,0	69,4	69,4
	1 1	60	9,1	9,5	78,9
	2 2	105	15,9	16,7	95,6
	3 3	23	3,5	3,7	99,3
	4 4	2	,3	,3	99,6
	5 5 a více	2	,4	,4	100,0
	Total	631	95,1	100,0	
Missing	-1 neodpovedel/a	32	4,8		
	Total	33	4,9		
Total		664	100,0		

Correlations

			CF1_3 Pocet detí	VZD_RESP vzdelání resp.
Spearman's rho	CF1_3 Pocet detí	Correlation Coefficient	1,000	-,067
		Sig. (2-tailed)	.	,093
		N	629	625

Regresní analýza

Závisle proměnná je počet dětí, které si žena v budoucnu přeje (cf1_3) bez ohledu na to, kolik dětí již má, nezávisle proměnné pak všechny tři dummy vzdělanostní proměnné. Pozor, do výpočtu regrese musí vstoupit najednou jako blok (což zajistí metoda Enter)



Výstup:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,158[a]	,025	,020	,946

a. Predictors: (Constant), VZD_VYUC dummy, VZD_VS dummy, VZD_SS dummy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,813	,088		9,203	,000
	VZD_VS dummy	-,435	,151	-,133	-2,880	,004
	VZD_SS dummy	-,169	,109	-,085	-1,555	,120
	VZD_VYUC dummy	-,377	,108	-,191	-3,510	,000

a. Dependent Variable: CF1_3 Poèet dìtí

Regresní rovnice má tvar:

*Počet dětí = 0,81 – 0,44*vzd_VS – 0,17*vzd_SS – 0,38*vzd_vyuc.*

Interpretace: Dummy proměnné vysvětlují pouze 2,5 % variance v počtu dětí, které si ženy přejí. Konstanta je zde průměrem pro referenční kategorii, z níž jsme dummy proměnnou nevytvářeli, to je pro základní vzdělání. Jak to? Vzpomeňme si, že pro kategorii základního vzdělání jsme žádnou dummy nevytvářeli a že měla ve všech ostatních dummy proměnných hodnotu 0. Dosadíme-li tedy do vypočtené regresní rovnice namísto jednotlivých hodnot dummy proměnných 0, výsledek výpočtu bude:

Počet dětí = 0,81 – 0,44*0 – 0,17*0 – 0,38*0
Počet dětí = 0,81 – 0 – 0 – 0
Počet dětí = 0,81

Tedy ženy se základním vzděláním si v průměru přály 0,81 dítěte. Hodnoty ostatních regresních koeficientů musí být čteny ve vztahu k této referenční kategorii. Takže ženy se vzděláním VŠ si přály v průměru o 0,44 dětí méně než ženy se vzděláním základním, neboť hodnota regresního koeficientu pro VŠ vzdělání, jak říká rovnice, je -0,44. Celkem si tak VŠ ženy přály 0,81-0,44 = 0,37 dětí.
Ukažme si tento výsledek s použitím regresní rovnice, kterou tedy řešíme pro kategorii žen s VŠ vzděláním. Dosaďme příslušné hodnoty do vypočtené regresní rovnice:

*Počet dětí = 0,81 – 0,44*vzd_VS – 0,17*vzd_SS – 0,38*vzd_vyuc.*

U proměnných SŠ a vyuč. musíme dosadit 0, pro proměnnou VŠ dosadíme hodnotu 1, takže:

Počet dětí = 0,81 – 0,44*1 – 0,17*0 – 0,38*0
Počet dětí = 0,81 – 0,44– 0 – 0 = 0,37.

Ženy se vzděláním SŠ si přály méně o 0,17 dítěte než ženy se základním vzděláním a vyučené o 0,38 dětí méně. Pro konkrétní počet dětí, které si přály, postupujeme ve výpočtu podobně jako v případě vysokoškoláček.

Poznamenejme, že u hledání vztahu mezi jednou nezávisle proměnnou, byť proměněnou na sérii *dummy variables*, tento způsob práce nemá příliš velký smysl[3], ve vícenásobné regresi, kdy nezávisle proměnných je více než jedna, to ovšem už význam má.

Pro ilustraci přidejme do naší úlohy další proměnnou – podívejme se, jak do vztahu mezi vzděláním a počtem dětí, které si žena přeje, intervenuje věk respondentek. Lze totiž předpokládat, že starší respondenty již nějaké dítě mají, a proto by si měli přát menší počet dětí než ženy mladší. Věk budeme pro jednoduchost dichotomizovat na mladší skupinu 18-30 (kód 0) a starší skupinu 31-49 (kód 1) – při práci s dummy proměnnými je výhodné kódovat dichotomické proměnné nulou a jedničkou. Zde jsou výsledky:

Z tabulky sestavme regresní rovnici, která má po zaokrouhlení hodnot parametrů tvar:

*Počet dětí = 2,43 – 0,07*vzd_VS + 0,02*vzd_SS – 0,12* vzd_vyuc – 0,49* age_2*

[3] Stejného výsledku bychom, vzpomeňte si, dosáhli, když bychom tento vztah posoudili prostřednictvím jednofaktorové analýzy rozptylu. Ta by ukázala jednak rozdíly v počtu dětí jednotlivých kategorií ve srovnání s ženami se základním vzděláním, jednak by určila, který rozdíl je statisticky signifikantní (mimochodem je to u vyučených a u žen s VŠ vzděláním).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2,432	,222		10,961	,000
	vzd_VS dummy	-,068	,241	-,023	-,283	,778
	vzd_SS dummy	,016	,139	,010	,116	,907
	vzd_vyuc dummy	-,123	,147	-,075	-,834	,406
	age_2	-,486	,194	-,186	-2,502	,013

a. Dependent Variable: cf1_3 Počet dětí

Interpretace probíhá opět v relaci k referenční kategorii. Hodnota regresního koeficientu pro věk je -0,486 (= -0,49). Ten říká, že když budeme kontrolovat vliv působení vzdělanostních kategorií (budeme je držet konstantní), tak starší skupina (ve věku 31-49 let) žen si přeje v průměru o 0,49 dítěte méně než skupina mladší. Koeficienty u dummy proměnných se oproti předchozímu výpočtu trochu pozměnily právě působením proměnné věk ve výpočtu – např. ženy s VŠ vzděláním si přejí o 0,07 dítěte méně, než ženy se vzděláním základním (to je naše referenční kategorie), v předchozím výpočtu to bylo 0 0,44 dětí méně.

A jak se interpretuje konstanta? Konstanta je hodnota pro respondentky, které byly ve všech nezávisle proměnných kódovány 0. Takže konstanta, jejíž hodnota je 2,43, je vlastně předpovězenou hodnotu průměrného počtu očekávaných dětí pro ženy ve věku 18-29 let se základním vzděláním. Dokažme si to. Rovnice: *Počet dětí = 2,43 – 0,07*vzd_VS + 0,02*vzd_SS – 0,12 vzd_vyuc – 0,49 age_2*

bude po dosazení příslušných hodnot vypadat takto, neboť ženy se základním vzděláním mají ve všech dummy proměnných hodnotu 0 a věkově mladší ženy mají také hodnotu 0:

*Počet dětí = 2,43 – 0,07*0+ 0,02*0 – 0,12*0 – 0,49*0 =* <u>*2,43*</u>

Starší ženy ve věku 31-49 let (kódované 1) očekávají, že budou mít v průměru ještě:

Počet dětí = 2,43 – 0,07*0+ 0,02*0 – 0,12*0 – **0,49*1** = 2,43 – 0,49 = <u>1,94.</u>

A kolik dětí budou mít v průměru ženy se SŠ vzděláním ve věku 31-49 let? Dosaďme do rovnice:

Počet dětí = 2,43 – 0,07*0+ **0,02*1**– 0,12*0 – **0,49*1** = 2,43 + 0,02 – 0,49 = <u>1,96.</u>

Model celkově vysvětluje pouze 5 % variance závisle proměnné, a to je málo.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,217[a]	,047	,027	,743

a. Predictors: (Constant), age_2, vzd_SS dummy, vzd_VS dummy, vzd_vyuc dummy

Z hodnot standardizovaných beta koeficientů vidíme, že nejsilněji působí na počet očekávaných dětí právě věk respondentky (-0,19), dummy proměnné mají beta koeficienty velmi nízké, navíc jsou všechny nesignifikantní.

Tolik tedy k dummy proměnným, který svým způsobem posloužil jako jistý vstup do vícenásobné regrese. O ní více až v magisterském kursu.

Literatura:

Vaus D. A. de. 2002. *Analyzing Social Science Data*. London: Sage, str. 368-373. (v knihovně FSS).