TECHNIKA UMĚLÝCH PROMĚNNÝCH V PRŮŘEZOVÉ ANALÝZE A V MODELECH ČASOVÝCH ŘAD

Umělé (dummy) proměnné se používají, pokud chceme do modelu zahrnout proměnné, které mají kvalitativní či diskrétní charakter, takže je nemůžeme přímo kvantifikovat. To nám umožní zkoumat působení kvalitativních faktorů jako pohlaví nebo vzdělání.

Kde používáme umělé proměnné?

1) v průřezové analýze:
    a. sociální, demografické, regionální charakteristiky jako vzdělání, pohlaví, úroveň ekonomického rozvoje země;
    b. pokud chceme rozdělit spojité kvantitativní veličiny (jako třeba věk) do kategorií (viz příklad 4)
2) v časových řadách:
    a. pro zahrnutí sezónnosti (sezónní očišťování) či cyklických vlivů (například vliv vánočních svátků na poptávku, výdaje na energie v zimě / v létě apod.)
    b. pro zahrnutí určitého zlomu do modelu (například doba před válkou a po válce nebo před zavedením určitého opatření a po něm – může jít o novou daň apod.)

Může jít o:

- diskrétní proměnné nabývající jen dvou hodnot (pohlaví)
- diskrétní proměnné nabývající několika hodnot (vzdělání)
- spojité proměnné, které lze rozdělit do několika kategorií (věk)
- interakci výše uvedených proměnných (pohlaví + vzdělání apod.)

Nejprve je potřeba určit klasifikační stupnici (škálu). Umělé proměnné jsou obvykle **binární** (dichotomické, tj. nabývají jen hodnot 0 a 1), ale lze použít i jinou škálu (0, 1, 2), v tom případě se však musíme zamyslet nad tím, zda je taková specifikace opodstatněná, protože hodnoty parametrů mohou být citlivé na použitou klasifikační stupnici (viz příklad 4).

V případě binárních proměnných označuje hodnota 1 přítomnost určitého znaku. Nula pak odpovídá základní kategorii (tzv. referenční skupině), s níž se druhá skupina srovnává.

Umělých proměnných musí být v modelu vždy o jednu méně, než kolik je kategorií, protože jinak by se v modelu vyskytla **perfektní multikolinearita**.

Pokud je vysvětlovaná proměnná **zlogaritmovaná**, pak koeficienty vysvětlujících umělých proměnných představují **relativní rozdíly** v proměnné Y. Jednotková změna nezávisle proměnné v tom případě vyvolá změnu ve výši $(e^{\beta} - 1)$ %. Například kdybychom měli model závislosti výše platu v nějaké firmě na pohlaví ve tvaru: $lnY_i = \beta_0 + \beta_1 D_i + u_i$, kde $D_i = 1$ pro muže a 0 pro ženy, a odhadli bychom parametry jako: $lnY_i = 10 + 0{,}2D_i$, pak exp(10) = 22 026 Kč by byl průměrný plat žen, exp(10 + 0,2) = 26 903 Kč by byl průměrný plat mužů, a to je o $(e^{0,2} - 1)$ % = 22 % více než plat žen.

Lze rozlišovat mezi umělými proměnnými, které mění **úrovňovou konstantu** (intercept dummy variables) a které mění **sklon křivky** (slope dummy variables).

➔ „intercept dummy variables“: mějme následující model, kde D_i = 1, má-li pozorování určitou sledovanou vlastnost (například je-li to muž):

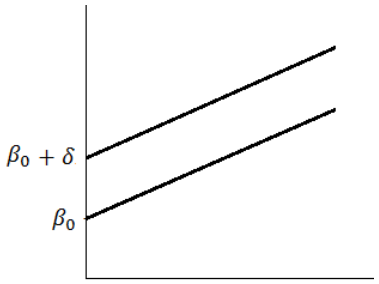
$$Y_i = \beta_0 + \delta D_i + \beta_1 X_{i1} \dots + \beta_k X_{ik} + u_i$$

Pak střední hodnota vysvětlované proměnné bude:

pro muže $E(Y_i) = (\beta_0 + \delta) + \beta_1 X_{i1} \dots + \beta_k X_{ik}$

pro ženy $E(Y_i) = \beta_0 + \beta_1 X_{i1} \dots + \beta_k X_{ik}$

To znamená, že se **mění úrovňová konstanta** pro jednotlivé skupiny, takže regresní přímka pro muže, resp. ženu, může vypadat například takto. Kdybychom například zkoumali závislost výše platu na pohlaví *D* a počtu let praxe *X*, znamenalo by to, **že nezávisle na počtu let praxe mají muži vyšší plat o δ Kč**.



➔ „slope dummy variables:“ mějme následující model, kde D_i = 1, jde-li o muže:

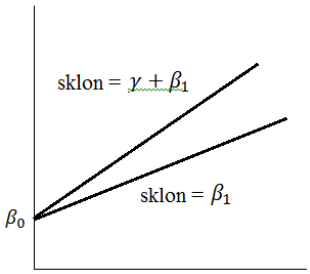
$$Y_i = \beta_0 + \gamma D_i X_{i1} + \beta_1 X_{i1} \dots + \beta_k X_{ik} + u_i$$

Pak střední hodnota vysvětlované proměnné bude:

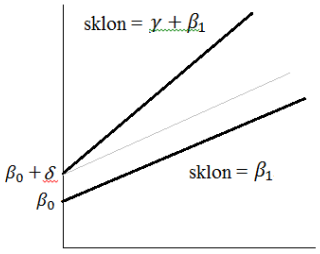
pro muže $E(Y_i) = \beta_0 + (\gamma + \beta_1) X_{i1} \dots + \beta_k X_{ik}$

pro ženy $E(Y_i) = \beta_0 + \beta_1 X_{i1} \dots + \beta_k X_{ik}$

To znamená, že se **mění sklon křivky** pro jednotlivé skupiny, takže regresní přímka pro muže, resp. ženu, může vypadat například takto. Pokud bychom opět zkoumali výši platu v závislosti na pohlaví a počtu let praxe, znamenalo by to, že platy žen a mužů bez praxe se shodují, ale **s rostoucím počtem let praxe roste plat mužů rychleji než plat žen**.



➔ nebo to lze obojí zkombinovat, jako třeba v modelu $Y_i = \beta_0 + \delta D_i + \beta_1 X_i + \gamma D_i X_{1i,} + \dots u_i$. kde X je počet let praxe a D_i = 1 pro muže, 0 pro ženy. Tím říkáme, že muži mají jednak vyšší nástupní plat než ženy, jednak že jejich plat s počtem let praxe roste rychleji. Regresní přímky by mohly vypadat nějak takto.



Příklad 1

Uvažujme model, v němž zkoumáme závislost platu na úrovni vzdělání (ZŠ, SŠ, VŠ). Specifikace modelu je:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i,$$

kde Y_i je plat i-tého pracovníka, D_{2i} = 1 pro pracovníka se ZŠ vzděláním, D_{3i} = 1 pro pracovníka se SŠ vzděláním. Pro podmíněné střední hodnoty platů tedy platí:

- *střední hodnota platu pracovníka se ZŠ vzděláním se rovná $\beta_1 + \beta_2$,*
- *střední hodnota platu pracovníka se SŠ vzděláním se rovná $\beta_1 + \beta_3$*
- *střední hodnota platu pracovníka s VŠ vzděláním se rovná β_1.*

To znamená, že úrovňová konstanta představuje průměrný plat vysokoškolsky vzdělaného člověka a parametry β_2 resp. β_3 pak rozdíly průměrného platu pracovníka se ZŠ resp. SŠ vzděláním proti pracovníkovi s VŠ vzděláním. Pravděpodobně by tedy byly záporné.

➔ *Mohli bychom testovat významnost parametrů β_2 a β_3, čímž bychom testovali, zda existuje statisticky významný rozdíl mezi platy osob s různým vzděláním. Všimněme si, že žádná umělá proměnná pro vysokoškolsky vzdělané osoby do modelu zahrnuto není, abychom se vyhnuli perfektní multikolinearitě (tato proměnná by byla lineární kombinací zbylých proměnných). Alternativně lze do modelu zahrnout všechny tři umělé proměnné, ale pak je potřeba **vynechat úrovňovou konstantu** a tím se také **změní interpretace parametrů**. Parametry β_2 a β_3 by pak už neříkaly, jaký je rozdíl v platech ZŠ a SŠ vzdělaného člověka proti vysokoškolákovi, ale šlo by zkrátka o průměrný plat v jednotlivých skupinách.*

Příklad 2

*Do modelu lze zahrnout i několik skupin umělých proměnných. Rozšíříme model z příkladu 1 o proměnnou MUŽ, kde $MUŽ_i$ = 1 pro muže a 0 pro ženy: $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \alpha MUŽ_i + u_i$. Jaká bude interpretace? Zkuste **sami spojit**, co patří k sobě:*

A) β_1	1. průměrný plat středoškoláka
B) $\beta_1 + \beta_2$	2. průměrný plat vysokoškolačky
C) $\beta_1 + \beta_3$	3. průměrný plat středoškolačky
D) $\beta_1 + \alpha$	4. průměrný plat ženy se ZŠ vzděláním
E) $\beta_1 + \beta_2 + \alpha$	5. průměrný plat muže se ZŠ vzděláním
F) $\beta_1 + \beta_3 + \alpha$	6. průměrná plat vysokoškoláka

Odpovědi: 1F, 2A, 3C, 4B, 5E, 6D

Příklad 3 – interakce mezi umělými proměnnými

Zkoumáme závislost poptávky domácností po službách v závislosti na vzdělání ženy a na tom, jestli je či není zaměstnaná. Model je specifikován jako: $\mathbf{Y_i = \alpha_0 + \alpha_2 V_{2i} + \alpha_3 V_{3i} + \beta_2 NZ_i + \gamma_1 D_{2i} + \gamma_2 D_{3i} + u_i}$, *kde*

- Y_i *jsou výdaje i-té domácnosti na služby*
- V_2 *= 1 pro středoškolsky vzdělanou ženu, jinak 0*
- V_3 *= 1 pro vysokoškolsky vzdělanou ženu, jinak 0*
- NZ_l *= 1 pro nezaměstnanou ženu, jinak 0*
- $D_2 = V_2NZ$ *= 1 pro středoškolsky vzdělanou nezaměstnanou ženu, jinak 0*
- $D_3 = V_3NZ$ *= 1 pro vysokoškolsky vzdělanou nezaměstnanou ženu, jinak 0.*
  Tím lze určit i ***vliv kombinace kvalitativních proměnných****. Interpretace je pak následující. Průměrné výdaje domácností na služby jsou …*

α_0	… v případě domácnosti se zaměstnanou ženou se ZŠ vzděláním
$\alpha_0 + \alpha_2$	… v případě domácnosti se zaměstnanou ženou se SŠ vzděláním
$\alpha_0 + \alpha_3$	… v případě domácnosti se zaměstnanou ženou s VŠ vzděláním
$\alpha_0 + \beta_2$	… v případě domácnosti s nezaměstnanou ženou se ZŠ vzděláním
$\alpha_0 + \alpha_2 + \beta_2 + \gamma_1$	… v případě domácnosti s nezaměstnanou ženou se SŠ vzděláním
$\alpha_0 + \alpha_3 + \beta_2 + \gamma_2$	… v případě domácnosti s nezaměstnanou ženou s VŠ vzděláním

Příklad 4

Mějme model, v němž zkoumáme závislost výše úspor (S_i) na příjmu (X_i) a věku. Rozdělíme si osoby do tří kategorií: 15-29 let, 30-44 let a 45-60 let. V modelu budou dvě umělé proměnné: proměnná $D_{2i} = 1$ *pro osoby ze střední věkové skupiny, jinak 0. Proměnná* D_{3i} *se rovná 1 pro osoby z nejstarší věkové skupiny, jinak 0. Myslíme si, že starší lidé více spoří. Model by mohl mít tvar* $S_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \alpha X_i + u_i$ *a interpretace by byla v tomto případě následující:*

- *průměrná výše úspor nejmladší skupiny (podmíněná střední hodnota) je* $\beta_1 + \alpha X_i$,
- *průměrná výše úspor střední skupiny (podmíněná střední hodnota) je* $\beta_1 + \beta_2 D_{2i} + \alpha X_i$
- *průměrná výše úspor nejstarší skupiny (podmíněná střední hodnota)* $\beta_1 + \beta_3 D_{3i} + \alpha X_i$.

Regresní parametr alfa (mezní sklon k úsporám) je stejný ve všech skupinách. Mohli bychom místo toho použít i proměnné D = 0, 1, 2 pro jednotlivé skupiny, ale v tom případě by rozdíly ve výši úspor mezi jednotlivými skupinami musely být ekvivalentní.

Příklad 5 – časové řady

Specifikace časové řady s umělými proměnnými by mohla vypadat například takto:

$$Y_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \alpha X_t + u_i,$$

kde závislou proměnnou jsou spotřební výdaje domácností, X_i *jsou příjmy i-té domácnosti, a umělé proměnné D odpovídají jednotlivým čtvrtletím, přičemž referenční je první čtvrtletí. Tím očistíme časovou řadu o sezónnost.*

JEŠTĚ PÁR PŘÍKLADŮ…
DOPLŇ CHYBĚJÍCÍ ÚDAJE

1. Chceme zkoumat, zda má absolvovaná fakulta VŠE vliv na nástupní plat. Uvažujeme jen pět pražských fakult, kde pro absolventa *i*-té fakulty platí, že $D_i = 1$. Kolik v modelu použijeme celkem umělých proměnných?

2. Zkoumáme nástupní plat absolventů VŠE a UK, kdy vysvětlovaná proměnná Y_i je ve tvaru logaritmů. Vysvětlujícími proměnnými počet let praxe při studiu X_i a absolvovaná škola D_i, přičemž platí, že $D_i = 1$ pro absolventa VŠE a 0 pro absolventa UK. Model vyšel následovně: $lnY = 10 + 0{,}05X_i + 0{,}1D_i$. Znamená to, že s každým rokem praxe vzroste nástupní plat absolventa v průměru o ____ % a že při stejném počtu let praxe budou mít absolventi VŠE plat v průměru o ____ % vyšší. Průměrný nástupní plat absolventa VŠE s dvěma roky praxe bude ____ Kč. Aby měl absolvent UK v průměru stejný nástupní plat, musel by mít o ____ roky/let praxe více.

ANALÝZA VÝSTUPU

Zdroj dat a specifikace modelu: University of Queensland, ECON2300, přednáška 6: Models with dummy variables, 2012.

Budeme zkoumat závislost výše hodinové mzdy v USD ($WAGE_i$) na pohlaví ($FEMALE_i$ = 1 pro ženy, jinak 0), barvě pleti ($BLACK_i$ = 1 pro osoby černé pleti, jinak 0i) a počtu let vzdělání ($EDUC_i$). Model specifikujeme takto:

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \delta_1 BLACK_i + \delta_2 FEMALE_i + \gamma(BLACK_i \cdot FEMALE_i) + u_i$$

Výstup z programu E-views je zde:

Dependent Variable: WAGE
Method: Least Squares
Date: 04/12/08 Time: 10:13
Sample: 1 1000
Included observations: 1000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.230327	0.967499	-3.338841	0.0009
EDUC	1.116823	0.069714	16.01998	0.0000
BLACK	-1.831240	0.895726	-2.044418	0.0412
FEMALE	-2.552070	0.359686	-7.095280	0.0000
BLACK*FEMALE	0.587905	1.216954	0.483096	0.6291

R-squared	0.248164	Mean dependent var	10.21302
Adjusted R-squared	0.245141	S.D. dependent var	6.246641
S.E. of regression	5.427245	Akaike info criterion	6.225728
Sum squared resid	29307.71	Schwarz criterion	6.250266
Log likelihood	-3107.864	F-statistic	82.10655
Durbin-Watson stat	0.480319	Prob(F-statistic)	0.000000

I. Úrovňová konstanta je dle modelu stejná pro ženy i pro muže téže pleti. ANO / NE
II. Sklon regresní přímky je dle modelu stejný pro ženy i pro muže téže pleti. ANO / NE
III. V modelu se vyskytuje umělá proměnná, která zachycuje interakci mezi pohlavím a barvou pleti, přesněji vliv této kombinace na mzdu. ANO / NE
IV. Model jako celek je na 5% hladině významnosti statisticky významný. ANO / NE
V. Všechny proměnné modelu jsou na 5% hladině významnosti statisticky významné. ANO / NE
VI. Referenční skupinou jsou: muži bílé pleti / ženy bílé pleti / muži černé pleti / ženy černé pleti.
VII. Muž bílé pleti s 10 lety vzdělání bude průměrně dostávat ____ USD na hodinu.
VIII. Žena černé pleti s 20 lety vzdělání bude průměrně dostávat ____ USD na hodinu.
IX. Žena bílé pleti s 15 lety vzdělání bude průměrně dostávat _____ USD na hodinu.
X. Kdybychom chtěli zároveň testovat významnost všech umělých proměnných v modelu, musíme použít t-test / F-test.
XI. Každý rok vzdělání zvýší průměrnou hodinovou mzdu o 1,11 USD / o 1,11 % / o 11 % / o 0,11 %.

Odpovědi:

1. Použijeme 4 umělé proměnné.
2. S každým rokem praxe vzroste nástupní plat absolventa v průměru o 5 %. Při stejném počtu let praxe budou mít absolventi VŠE plat v průměru o 10,5 % vyšší. Průměrný nástupní plat absolventa VŠE s dvěma roky praxe bude 26 903 Kč. Aby měl absolvent UK v průměru stejný nástupní plat, musel by mít o 2 roky praxe více.
3. I. NE
   II. ANO
   III. ANO
   IV. ANO
   V. NE
   VI. muži bílé pleti
   VII. 7,9 USD
   VIII. 15,3 USD
   IX. 10,9 USD
   X. F-test
   XI. o 1,11 USD

ZDROJE

Hušek, R: Ekonometrická analýza. Nakladatelství Oeconomica, Praha 2007.
University of Queensland, ECON2300, přednáška 6: Models with dummy variables, 2012.

Lenka Fiřtová (2014)