

Umělé (dummy) proměnné v ekonometrickém modelu

V ekonometrických modelech se často mezi vysvětlujícími proměnnými vyskytují veličiny, které nelze číselně vyjádřit měřením. Přítomnost těchto veličin je nicméně často velmi důležitá s ohledem na to, že tyto veličiny často vykazují významný vliv na závisle proměnnou. Nejčastěji se jedná o proměnné demografického, sociálního, urbanistického, etnického nebo obdobného charakteru. Společným znakem těchto kvalitativních či diskrétních proměnných je to, že mají omezený (a často uměle vytvořený) okruh přípustných hodnot, kterých nabývají.

Z hlediska možností obměn, kterých veličina nabývá, rozlišujeme:

- a) proměnné dichotomické (dvouznakové, binární) jako je pohlaví (muž/žena), místo bydliště jedince (městské/venkovské), příslušnost k etnické skupině (neRom/Rom), zvyk chování (kuřák/nekuřák) apod.
- b) proměnné kategoriální (víceznakové, leč s omezeným oborem přípustných hodnot). Příkladem může být stupeň vzdělání (základní/střední/vysokoškolské), věková skupina (řekněme v 5 nebo 10-letých agregacích), příjmové či majetkové rozvrstvení apod. Někdy lze hodnoty proměnné seřadit (věk, stupeň vzdělání, příjmové kategorie), jindy to postrádá smysl (pohlaví, profesní struktura apod.)

Poznamenejme, že umělou proměnnou není např. počet členů domácnosti, byť je vyjádřena vždy jen přirozeným číslem.

Bez zařazení těchto proměnných do regresních vztahů bychom byli ochuzeni o významný informační přínos, který právě zvláštnost příslušnosti k některé specifické skupině přináší.

V modelech založených na časových řadách se navíc často uplatní umělé proměnné k postižení sezónnosti: postihneme jimi právě vliv specifického měsíce nebo čtvrtletí v průběhu daného roku.

Všimněme si několika důležitých otázek hrajících úlohu při formulaci regresního vztahu s umělými proměnnými

1) stanovení hodnot umělé proměnné: při identifikaci pohlaví není podstatné, zda muž = 1, žena = 0, nebo opačně nebo hodnoty 1,2 či jiné.

2) Rozdělení stupnice pro věkovou strukturu by mělo vycházet z potřeb analýzy a z požadavku, aby homogenita sledované vlastnosti uvnitř skupin byla zřetelně vyšší než mezi jednotkami/příslušníky různých skupin. U věkových skupin sotvakdy požadujeme detailnější než 5-leté členění. Často se krajní intervaly (s početně méně zastoupenými jedinci) stanovují širší než vnitřní (např. společná věková třída : „ nad 80 let “).

3) Pokud je stupnice hodnot znaku příslušná dané proměnné více než dvouznaková, je užitečné nejprve vyšetřit, zda skutečný funkční tvar závislosti vysvětlované proměnné na dané (umělé) vysvětlující odpovídá předpokládanému, protože hodnoty odhadnutých parametrů mohou být citlivé na použitou klasifikační stupnici.

Je užitečné říci, že obvykle (byť na první pohled překvapivě) se upřednostňuje užití kombinovaných 0-1 vektorů než víceznakové vyjádření dané proměnné. Je tomu tak i přesto, že tato cesta vede často k podstatnému zvýšení počtu odhadovaných regresních koeficientů.

Volba nula-jedničkového schématu hodnot umělých proměnných (a obecná tendence preferovat spíše 0-1 schéma na úkor vícebodové ordinální stupnice) má svůj důvod mj. v možnosti, co nejpřirozeněji postihnout význam regresních koeficientů. Vždy se snažíme především o to, aby míra vlivu specifické umělé veličiny byla popsitelná co nejjednodušší kombinací modelových parametrů.

Dále, při zařazování umělých proměnných do regresního vztahu se musíme vystříhat toho, aby došlo (zařazením všech umělých proměnných) k nežádoucímu vzniku (přesné) multikolinearity. Vždy si můžeme dovolit zařadit do regresního vztahu (obsahuje-li tento vektor jedniček) umělé proměnné maximálně v takovém počtu, který je o 1 menší, než je jejich počet pro veličinu, kterou vystihují (tedy nanejvýš 3 čtvrtletní umělé proměnné, 11 měsíčních nebo 2 proměnné pro třístupňovou vzdělanostní klasifikaci). V případě kombinací více veličin vystižených umělými proměnnými se tento maximální přípustný počet dále snižuje.

Přibližme příkladem : Ve vzorku cca 200 osob sledujeme závislost mzdy pracovníka na nejvyšším dosaženém stupni jeho vzdělání. Za tímto účelem formulujeme regresní rovnici vztahem

$$(1) \quad Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t \quad , \quad \text{kde}$$

Y_t je (roční) mzda t-tého pracovníka

$X_{t1} = 1$ pro všechna t (jde o jedničkový vektor)

$X_{t2} = 1$, má-li pracovník (nejvýš) základní vzdělání

$X_{t2} = 0$ v ostatních případech

$X_{t3} = 1$, má-li pracovník (nejvýš) středoškolské vzdělání

$X_{t3} = 0$ v ostatních případech

ε_t je náhodná složka regresní rovnice s obvyklými stochastickými vlastnostmi (např. standardního lineárního regresního modelu).

K přiblížení interpretace regresních parametrů nám zde poslouží nejlépe vyjádření v podmíněných středních hodnotách. Tak lze zapsat

$$E(Y_t | X_{t2} = 0 ; X_{t3} = 0) = \beta_1$$

$$E(Y_t | X_{t2} = 1 ; X_{t3} = 0) = \beta_1 + \beta_2$$

$$E(Y_t | X_{t2} = 0 ; X_{t3} = 1) = \beta_1 + \beta_3$$

Odtud je patrné, že úrovnňová konstanta β_1 vyjadřuje průměrný plat vysokoškolačka. Regresní parametr β_2 představuje rozdíl v průměrných platech vysokoškolačka a osobou se základním vzděláním, a obdobně β_3 měří rozdíl mezi průměrným platem vysokoškolačka a středoškolačka. V případě testu hypotézy o neexistenci významného rozdílu mezi platy vysokoškolačka a středoškolačka bychom formulovali a testovali nulovou hypotézu tvaru $\beta_3 = 0$.

Povšimněme si, že do regresní rovnice nelze zařadit třetí umělou proměnnou ($X_{t4} = 1$ pro případ, že se jedná o vysokoškolačka), neboť by vznikla perfektní multikolinearita (součet vektorů všech tří umělých proměnných by poskytl vektor identický s jedničkovým vektorem).

V modelu (1) bychom mohli vynechat jedničkový vektor (s parametrem β_1) a uplatnit tak modifikovaný tvar rovnice

$$(1A) \quad Y_t = \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \varepsilon_t, \quad \text{kde}$$

$X_{t4} = 1$, má-li pracovník vysokoškolské vzdělán

$X_{t4} = 0$ v ostatních případech

avšak interpretace parametrů bude nyní jiná. Tak rozdíl mezi výší mezd středoškolačka a pracovníka se základním vzděláním bude nyní dán rozdílem parametrů $\beta_3 - \beta_2$, rozdíl v průměrných mzdách vysokoškolačka a osoby se základním vzděláním udává rozdíl $\beta_4 - \beta_2$, zatímco rozdíl mezi průměrnou mzdou vysoko- a středoškolačka udává rozdíl parametrů $\beta_4 - \beta_3$.

Formulujme nyní původní regresní rovnici (1) s modifikací představovanou přidáním proměnné pohlaví. Označme ji S_t

$$(2) \quad Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \gamma S_t + \varepsilon_t, \quad \text{kde}$$

$S_t = 0$, jedná-li se o muže

$S_t = 1$, jde-li se o ženu . Význam ostatních veličin modelu zůstává stejný.

Pak opět vyjádření v podmíněných středních hodnotách vede k výsledkům

$$E(Y_t | X_{t2} = 0 ; X_{t3} = 0 ; S_t = 0) = \beta_1$$

$$E(Y_t | X_{t2} = 1 ; X_{t3} = 0 ; S_t = 0) = \beta_1 + \beta_2$$

$$E(Y_t | X_{t2} = 0 ; X_{t3} = 1 ; S_t = 0) = \beta_1 + \beta_3$$

$$E(Y_t | X_{t2} = 0 ; X_{t3} = 0 ; S_t = 1) = \beta_1 + \gamma$$

$$E(Y_t | X_{t2} = 1 ; X_{t3} = 0 ; S_t = 1) = \beta_1 + \beta_2 + \gamma$$

$$E(Y_t | X_{t2} = 0 ; X_{t3} = 1 ; S_t = 1) = \beta_1 + \beta_3 + \gamma$$

Ve všech případech představuje parametr γ rozdíl mezi průměrnými mzdami žen a mužů majících jinak stejný nejvyšší dosažený stupeň vzdělán

Zde opět parametr β_1 vyjadřuje průměrnou mzdu pracovníka s VŠ vzděláním, zatímco parametr β_2 představuje rozdíl v průměrných platech vysokoškoláka a muže se základním vzděláním, a obdobně β_3 měří rozdíl mezi průměrným platem vysokoškoláka a středoškoláka.

Poznámka : z povahy zadání modelu lze vyvodit, že parametry β_2 , β_3 budou pravděpodobně záporné. Totéž očekávání lze vyslovit ve vztahu k parametru γ , pokud jsme zvolili $S_t = 0$ pro muže, resp. $S_t = 1$ u ženy.

Někdy se situace může dále komplikovat, pokud připustíme vzájemné interakce mezi určitými kvalitativními proměnnými (zde např. závislost dosaženého stupně vzdělání na pohlaví). To navíc vede k dalšímu nárůstu počtu umělých proměnných a k zvýšení pravděpodobnosti vzniku problémů spojených s nízkým počtem stupňů volnosti při statistickém testování.

Poznámka : V komplikovanějších úlohách se někdy ukazuje vhodnější než regresi s více umělými diskrétními proměnnými uplatnit analýzu rozptylu, která je ekvivalentní regresní analýze, pokud model obsahuje výlučně nula-jedničkové *vysvětlující* proměnné.

Příkladem modelu, který v sobě zahrnuje jako vysvětlující jak umělé proměnné, tak konvenční ekonomické (měřitelné) proměnné, může být model zobrazující funkci úspor v následující specifikaci :

$$(3) \quad S_t = \beta_1 X_{t1} + \beta_2 D_{t2} + \beta_3 D_{t3} + \gamma \cdot Y_t + \varepsilon_t \quad , \quad \text{kde}$$

S_t je objem úložek (alokovaných za daný rok ke stávajícím úsporám)

$D_{t2} = 1$ pro 2. věkovou skupinu
 $= 0$ jinak (pro jiné skupiny)

$D_{t3} = 1$ pro 3. věkovou skupinu
 $= 0$ jinak (pro jiné skupiny)

Y_t je disponibilní příjem t-tého spořitele

$X_{t1} = 1$ pro všechna t (jde opět o jedničkový vektor s interpretací jisté „minimální“ hladiny úložek)

ε_t je náhodná složka regresní rovnice s obvyklými vlastnostmi

Předpokládáme přitom, že 3 užití věkové skupiny jsou stanoveny takto :

1. skupina : věk 16 - 29 let
2. skupina : věk 30 - 44 let
3. skupina : věk 45 - 60 let

Usuzujeme tedy, že kromě disponibilního příjmu Y_t je roční objem úspor S_t (úložky na vklady) závislý na věkové struktuře spořitelů, přičemž v souladu s realitou lze očekávat, že s přibývajícím věkem roste tendence ke spořivosti (s ohledem na zabezpečení přibližujícího se stáří). Mezní sklon k úsporám (koeficient γ) je (jako průměrná hodnota) neutrální vůči věku (vztahuje se k průměrnému spořiteli).

Poznámka : Veličina S_t by neměla být zaměňována s hodnotou úspor vyjádřených ve stavové formě (např. jako zůstatek na účtech či jiných vkladových depozitech a hodnota likvidních cenných papírů), neboť ta je silně závislá na dříve (v minulých letech) naspořených částkách. Pro vystižení takové závisle proměnné bychom se neobešli (přinejmenším) bez její hodnoty v minulém roce S_{t-1} , a patrně též bez proměnné vyjadřující objemy výběrů z těchto účtů.

Pro model (3) tedy máme

$$E(S_t | D_{t2} = 0 ; D_{t3} = 0, Y_t) = \beta_1 + \gamma \cdot Y_t$$

$$E(S_t | D_{t2} = 1 ; D_{t3} = 0, Y_t) = (\beta_1 + \beta_2) + \gamma \cdot Y_t$$

$$E(S_t | D_{t2} = 0 ; D_{t3} = 1, Y_t) = (\beta_1 + \beta_3) + \gamma \cdot Y_t$$

Každý ze vztahů představuje závislost výše úložek na disponibilním příjmu v první, druhé a třetí věkové kategorii. Nejmladší věková skupina je zde přijata jako základní hladina, vůči které jsou porovnávány ostatní dvě. S ohledem na tendenci růstu spořivosti s věkem, lze očekávat, že $\beta_2 > 0$, $\beta_3 > 0$. S ohledem na svůj význam bude koeficient γ také kladný.

Poznámka : Věkové skupiny bychom mohli také ohodnotit pořadovými čísly 1, 2, 3 a pracovat s jedinou vysvětlující proměnnou D. Model by pak pozměnil tvar na

$$(3a) \quad S_t = \beta_1 X_{t1} + \beta_2 D_t + \gamma \cdot Y_t + \varepsilon_t, \quad \text{kde}$$

$D_t = 1$ pro osobu z 1. věkové skupiny

$D_t = 2$ pro osobu z 2. věkové skupiny

$D_t = 3$ pro osobu z 3. věkové skupiny

(význam ostatních veličin S_t, Y_t, ε_t zůstává nezměněn)

Tento postup však není plně ekvivalentní s předchozím (nehledě na jinou interpretaci parametru β_2), neboť se zde předpokládá „ekvidistantnost“ rozdílů ve spořivosti (tzn. rozdíl mezi 1. a 2. skupinou by v této specifikaci musel být stejný jako rozdíl mezi spořivostí 2. a 3. skupiny). Tento předpoklad nemusí být plně realistický.

Nahrazení původních hodnot umělými proměnnými (dichotomickými nebo i kategoriálními) však vede pouze k aproximativnímu odhadu vlivu původní vysvětlující proměnné na změny závisle proměnné. Přesnost takového odhadu přirozeně klesá s početností a nestejnorodostí vytvořených skupin/kategorií.