4EK211 Základy ekonometrie

Umělé proměnné

Predikce

Multikolinearita

Cvičení 5

Umělé proměnné

Umělé proměnné

- **Dummy** / umělé / booleovské **proměnné**
- nabývají hodnotu: 0 nebo1

- Kvalitativní proměnné: vyjadřují přítomnost či nepřítomnost určité vlastnosti
  - přítomnost popisované vlastnosti… obvykle 1
  - absence vlastnosti … obvykle 0

  - Příklad: proměnná **žena**: $\check{z}ena_i = 1$ (*i*-tá osoba je žena)
    $= 0$ (*i*-tá osoba je muž)

  - Příklad: **vzdělání** zkoumaných osob

Dummy proměnná / pozorování	$D1_i$	$D2_i$	$D3_i$
i-tá osoba: základní vzdělání	1	0	0
i-tá osoba: ukončená maturita	0	1	0
i-tá osoba má VŠ titul	0	0	1

Umělé proměnné

- Příklad: **sezónnost**

  Pro jednotlivá čtvrtletí vytvoříme umělé proměnné D1 až D4 (analogicky jako vzdělání), tj.

  $D1_i = 1$ pokud *i*-té pozorování odpovídá 1. čtvrtletí

  $D1_i = 0$ pokud *i*-té pozorování odpovídá jinému než 1. čtvrtletí

PAST UMĚLÝCH PROMĚNNÝCH

Zahrneme-li do LRM umělé proměnné odpovídající všem variantám sledované kvalitativní proměnné, budou umělé proměnné perfektně lineárně závislé (porušení 4. G.M. předpokladu) a MNČ selže, protože matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ neexistuje.

Příklad perfektní lineární závislosti (sezónnost):

$$D1_i + D2_i + D3_i + D4_i = 1 \text{ pro } \forall\, i$$

Proč: při práci se čtvrtletními daty každému pozorování *i* vždy odpovídá právě jedno čtvrtletí (jedna umělá proměnná rovna 1, ostatní rovny nule).

Umělé proměnné

- **PAST UMĚLÝCH PROMĚNNÝCH - ŘEŠENÍ**

- Cíl: **vyvarovat se perfektní lineání závislosti mezi dummies (**též: **perfektní multikolinearita umělých proměnných)**

- do modelu zahrneme o jednu dummy proměnnou méně než je počet sledovaných vlastností

- Nezahrnutá dummy proměnná tvoří základ (bázi), ke kterému ostatní vlastnosti porovnáváme:

  – dvě pohlaví – jedna dummy
  – tři stupně vzdělání – dvě dummies
  – čtyři čtvrtletí – použijeme maximálně 3 dummies

- **Interpretace** parametrů β u dummies závisí na tom, kterou proměnnou jsem vynechal, bázi vůči které porovnávám.

Umělé proměnné – příklady

Soubor: CV5_PR1.xls

Data: ***y*** = plat učitelů (tis. USD)
x = roky praxe
m = pohlaví (1 = muž, 0 = žena)

Zadání: Odhadněte model závislosti ***y*** na ***x*** a ***m*** a interpretujte získané výsledky.

$y_i = \beta_0 + \beta_1 x_i + \beta_2 m_i + u_i, \quad i = 1, 2,...,15$

Umělé proměnné – příklady

Soubor: CV5_PR2.xls

Data: ***y*** = výdaje na cestování (tis. USD)
x = výše příjmu (tis. USD)
$\boldsymbol{D_1}$ = dosažené vzdělání (1 = základní, 0 = jiné)
$\boldsymbol{D_2}$ = dosažené vzdělání (1 = středoškolské, 0 = jiné)
$\boldsymbol{D_3}$ = dosažené vzdělání (1 = vysokoškolské, 0 = jiné)

Zadání: Odhadněte model závislosti ***y*** na ***x***, $\boldsymbol{D_2}$ a $\boldsymbol{D_3}$ a interpretujte získané výsledky.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i, \quad i = 1, 2,...,15$$

Umělé proměnné – příklady

Soubor: CV5_PR3.xls

Data: ***t*** = čas
R = příjmy státního rozpočtu (v mld. Kč)

Zadání: Odhadněte model závislosti ***R*** na ***t***.
Pokuste se zachytit v modelu vliv posledního čtvrtletí v daném roce (tj. explicitně zapojit čtvrtý kvartál do modelu).

$R_t = \beta_0 + \beta_1 t_t + u_t, \quad t = 1, 2,...,15$

Umělé proměnné – příklady

Soubor: CV5_PR4.xls

Data: ***pocet_domu*** **=** počet nově započatých staveb domů v USA (v tis.)

urok_mira **=** úroková míra (v %)

Zadání: Odhadněte model závislosti ***pocet_domu*** na ***urok_mira*** + zohledněte sezónní vliv v modelu.
Predikujte ***pocet_domu*** v roce 1999.

$$pocet\_domu_t = \beta_0 + \beta_1 urok\_mira_t + \beta_2 Q_{2t} + \beta_3 Q_{3t} + \beta_3 Q_{4t} + u_t, \quad t = 1, 2,...,40$$

Predikce

Aplikace EM – predikce obecně

- ekonomické prognózování, předpověď, předvídání
- hlavním cílem je odhad hodnot vysvětlované proměnné mimo interval pozorování s užitím minulé i současné informace
  - extrapolace modelu do budoucna
  - extrapolace modelu do minulosti – tj. před interval pozorování (tzv. retrospektiva)
- predikcí získáváme vyrovnané hodnoty (tj. hodnoty „fitted“)

Predikce ex-ante (resp. dopředu)

- tzv. podmíněná
- podmíněná volbou vysvětlujících proměnných
- na napozorované hodnoty musíme „čekat“

Predikce ex-post (resp. dozadu)

- tzv. pseudopředpověď
- slouží k testování kvality modelu
- napozorované hodnoty jsou již k dispozici

Aplikace EM – predikce ex-ante

- volba podmíněných exogenních proměnných, možné způsoby:
  - zadáno z jiné analýzy
  - zadáno pomocí procentuální změny oproti minulému období (např. o 10 %)
  - zadáno pomocí diferencí

- predikce **bodová**
- predikce **intervalová**
  - se směrodatnou odchylkou **sigma**: $\hat{y}_p \pm \mathbf{sigma}$
  - se směrodatnou odchylkou $\tilde{s}_p$: $\hat{y}_p \pm t^*_{1-\alpha/2(n-(k+1))} \tilde{s}_p$

$$\tilde{s}_p = s\sqrt{1 + x_p (X^T X)^{-1} x_p^T}$$

  - *„s“* ve vzorci je sigma, výpočet viz předchozí cvičení nebo to je sigma ve výstupu PcGive
  - vždy platí: $\tilde{s}_p > \mathbf{sigma}$
  - intervalový odhad se sigma bývá podhodnocený

Aplikace EM – predikce ex-ante + příklad

Soubor: CV5_PR5.xls

Data: ***Mira_nezam_obec*** = obecná míra nezaměstnanosti (%) → *x*
Inflace = míra inflace (%) → *y*

Zadání: Odhadněte závislost míry inflace (*y*) na obecné míře nezaměstnanosti (*x*).

Proveďte **bodovou předpověď ex-ante** ručně i pomocí PcGivu, víte-li, že hodnota ***Mira_nezam_obec*** v roce 2008 je 4,4.
Proveďte **intervalovou předpověď ex-ante** ručně i pomocí PcGivu, víte-li, že hodnota ***Mira_nezam_obec*** v roce 2008 je 4,4 a α = 1 %.

$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, 2,...,15$

Aplikace EM – predikce ex-post + příklad

- testuje se kvalita modelu
- vyřadíme určitý počet pozorování z modelu
- odhadneme model
- provedeme predikci vynechaných hodnot
- porovnáme získané předpovědi se skutečnými hodnotami
- obecně platí, že predikce je dobrá, pokud je absolutní hodnota chyby predikce menší než 5 % ze skutečné hodnoty pro dané období

Příklad: Proveďte predikci/předpověď ex-post pro roky 2006 a 2007 na datech CV5_PR5.xls.

Multikolinearita

Gauss-Markovy předpoklady

Opakování: Gaussovy-Markovovy předpoklady

1. $E(\boldsymbol{u}) = 0$
   - průměrná hodnota náhodné chyby je nula

2. $E(\boldsymbol{u}\,\boldsymbol{u}^T) = \boldsymbol{\sigma}^2\,\boldsymbol{I}_n$
   - diagonála: konečný a konstantní rozptyl = homoskedasticita
   - mimo diagonálu: náhodné složky jsou sériově nezávislé

3. ***X*** je nestochastická matice – $E(\boldsymbol{X}^T\boldsymbol{u}) = 0$
   - veškerá náhodnost ***y*** je obsažena v náhodné složce

4. ***X*** má plnou hodnost k+1 (hodnost matice ***X*** = počet sloupců ***X***)
   - matice ***X*** neobsahuje žádné perfektně lineárně závislé sloupce pozorování vysvětlujících proměnných

     → porušení: multikolinearita

Multikolinearita - definice

- Podmínky aplikace MNČ:
  - lineární nezávislost sloupců matice $\boldsymbol{X}$
  - matice E($\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$) není singulární,
  - existuje její nenulový determinant,
  - lze spočítat odhadovou funkci $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

- Porušení vede ke kolinearitě, resp. **multikolinearitě:**
  - **Multikolinearita**: existence více vztahů lineární závislosti mezi pozorováními vysvětlujících proměnných
    multikolinearita silná (statisticky významná)
    multikolinearita perfektní (past umělých proměnných)
  - **Kolinearita**: existence pouze jednoho lineárního vztahu (obvykle hovoříme o multikolinearitě i když jde pouze o kolinearitu).

- Zkoumání multikolinearity: intenzita závislosti mezi dvěma nebo více vysvětlujícími proměnnými (je či není multikolinearita únosná?)

Multikolinearita – příčiny a důsledky

Příčiny

- tendence časových řad ekonomických ukazatelů (makroúdajů) vyvíjet se stejným směrem (např. HDP, C, I, S, Ex, Im)
- použití průřezových dat
- zahrnutí zpožděné endogenní nebo exogenní proměnné
- chybné použití dummies (past umělých proměnných)

Důsledky

- snížená přesnost odhadů regresních koeficientů
- velké standardní chyby odhadové funkce MNČ
  - pochybnosti či nejistota pokud jde o správnost specifikace modelu
- odhady zůstávají nestranné, vydatné
- velká citlivost odhadové funkce MNČ na velmi malé změny v matici $\boldsymbol{X}$
- obtížné vyjádření odděleného působení silně kolineárních proměnných

Měření multikolinearity – metoda I

- použití párových korelačních koeficientů
- pro pouze 2 vysvětlující proměnné:

$$r_{x_1,x_2} = \frac{\mathrm{cov}(x_1 x_2)}{s_{x_1} s_{x_2}} \in \langle -1,1 \rangle$$

- multikolinearita je únosná, pokud:

  $r_{x_1 x_2} \leq 0{,}9$ a současně

  $r^2_{x_1 x_2} < R^2_{(y.x_1 x_2)} \rightarrow$ koeficient vícenásobné determinace modelu

- modul PcGive → *Package* → Descriptive Statistics
  → *Model* → Formulate → x_1, x_2,
- zvolit nabídku korelační matice

Měření multikolinearity – příklad na metodu I

Soubor: CV3_PR1.xls

Data: ***y*** = maloobchodní obrat potřeb pro domácnost v mld. CZK
***x*$_1$** = disponibilní příjem v mld. CZK
***x*$_2$** = cenový index

Zadání: Odhadněte závislost maloobchodního obratu (***y***) na disponibilním příjmu (***x*$_1$**) a cenovém indexu (***x*$_2$**). Vyhodnoťte multikolinearitu.

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, $i = 1, 2,...,8$

Měření multikolinearity – metoda II

- více vysvětlujících proměnných (tj. nestačí zkoumat párové korelační koeficienty)
- využívá se pomocné regrese a koeficientů R_i^2
- $y = f(x_1, x_2, x_3) \longrightarrow$ z modelu $\longrightarrow R^2$
  - $x_1 = f(x_2, x_3) \longrightarrow R_1^2$
  - $x_2 = f(x_1, x_3) \longrightarrow R_2^2$
  - $x_3 = f(x_1, x_2) \longrightarrow R_3^2$
- jsou-li všechna $R_i^2 < R^2$, pak je multikolinearita únosná

Měření multikolinearity – příklad na metodu II

Soubor: CV5_PR6.xls

Data: ***y*** = počet prodaných kuřat (v desítkách milionů kusů)
x_1 = výše dotace do zemědělství (v mld. Kč)
x_2 = cena za kuře (Kč/kg)
x_3 = cena vepřového (Kč/kg)

Zadání: Odhadněte závislost počtu prodaných kuřat (***y***) na proměnných x_1, x_2 a x_3.
Vyhodnoťte multikolinearitu.

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t,\ t = 1, 2,...,23$$

Měření multikolinearity – příklad na metodu I a II

Soubor: CV5_PR7.xls

Data: ***HDP*** = hrubý domácí produkt (v mld. Kč)
C = spotřeba (v mld. Kč)
I = investice (v mld. Kč)
G = vládní výdaje (v mld. Kč)

Zadání – metoda I:
Odhadněte závislost *HDP* na proměnných *C* a *I*. Vyhodnoťte multikolinearitu.

$$HDP_t = \beta_0 + \beta_1 C_t + \beta_2 I_t + u_t, \quad t = 1, 2,...,19$$

Zadání – metoda II:
Odhadněte závislost *HDP* na proměnných *C*, *I* a *G*. Vyhodnoťte multikolinearitu.

$$HDP_t = \beta_0 + \beta_1 C_t + \beta_2 I_t + \beta_3 G_t + u_t, \quad t = 1, 2,...,19$$

Multikolinearita – řešení

Možnosti řešení neúnosné multikolinearity v LRM:

- získání dalších pozorování
- snížení počtu exogenních proměnných
- použití jiného modelu
- použití jiné odhadové techniky
- transformace pozorování
  - první diference - pozor na autokorelaci
  - poměrové veličiny - pozor na heteroskedasticitu