PROBLÉM NESPRÁVNÉ SPECIFIKACE MODELU

Chyby pramenící z nesprávné specifikace modelu (chápané v širokém smyslu slova) mohou mít několik příčin. Nejčastější z nich jsou:

A – nesprávný výběr proměnných zařazených do modelu

A1 – zařazení nepatřičné (*irelevantní, nedůležité* **) vysvětlující proměnné**
A2 – vynechání patřičné *(relevantní, důležité)* **vysvětlující proměnné**

B – nesprávná volba analytického funkčního tvaru:

B1 – v modelu uvažovaný lineární vztah je ve skutečnosti nelineární
B2 – nelinearita má ve skutečnosti jiný tvar než předpokládaný modelem

C – chybný předpoklad o vlastnostech náhodné složky regresní rovnice

C1 – aditivní vs. multiplikativní připojení s vysvětlujícím proměnným
C2– ***heteroskedasticita*****,** ***autokorelovanost náhodných složek*** v realitě**, zatímco model uvažuje splnění klasických předpokladů** (stejný rozptyl, nezávislost)

1. OBECNÁ FORMULACE

Uvažujeme jednorovnicový model v obvyklém maticovém zápisu

(1) $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$ s obvyklými vlastnostmi LRM.

Místo něho však formulujeme (naneštěstí) model v chybné specifikaci (ten bude mít mezi regresory obsaženými v matici $\tilde{\mathbf{X}}$ jiné proměnné než v matici $\mathbf{X}$.[1] Bude to model tvaru

(2) $$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^* \quad .$$

Odhadem parametrů (metodou OLS) bude vektor (chybně) odhadnutých parametrů roven

(3) $$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}$$ neboli po dosazení za $\mathbf{y}$ z **(1)**

(3A) $$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}).$$

Pro střední hodnotu tohoto vektoru parametrů platí

$$E(\hat{\boldsymbol{\beta}}^*) = E\left[(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right] = E\left[(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{X}\boldsymbol{\beta}\right] + E\left[(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\boldsymbol{\varepsilon}\right] = \mathbf{P}.\boldsymbol{\beta},$$

kde maticí $\mathbf{P} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{X}$ je násoben vektor správných koeficientů $\boldsymbol{\beta}$, leč $\mathbf{P} \neq \mathbf{I}_k$

Můžeme ji interpretovat jako matici v pomocné regresi správně specifikovaných regresorů $(\mathbf{X})$ na chybně specifikované regresory $(\mathbf{X}^*)$ v modelu (2)

V dalším zvlášť pojednáme o situaci, kdy je matice $\mathbf{X}^*$ částí matice $\mathbf{X}$ (tzn. dochází k vynechání jedné nebo více proměnných, které v modelu mají být jako vysvětlující přítomny) a zvlášť o situaci, je matice $\mathbf{X}$ částí matice $\mathbf{X}^*$ (tzn. jde o případ, kdy jsou do modelu zařazeny nadbytečné vysvětlující proměnné)

[1] Přirozeně, některé proměnné obsažené ve sloupcích obou matic mohou být (a zopravidla budou) společné.

2. VYNECHÁNÍ RELEVANTNÍCH PROMĚNNÝCH

konkrétně (pro 2 vysvětlující proměnné:

Předpokládejme, že místo správně specifikovaného modelu

(4) $$y_t = \beta_1 + \beta_1 x_{t2} + \beta_2 x_{t3} + \varepsilon_t$$

uvažujeme a následně odhadujeme nepřesný model (s vynecháním proměnné x_3) :

(5) $$y_t = \alpha_1 + \alpha_2 x_{t2} + u_t$$

Důsledky vynechání proměnné jsou tyto:

1. Pokud je vynechaná proměnná x_3 korelovaná se zařazenou proměnnou x_2, tj $r_{23} \neq 0$, pak budou odhady $\hat{\alpha}_1, \hat{\alpha}_2$ jak vychýlené[2], tak nekonzistentní, tzn. že platí jak $E\hat{\alpha}_1 \neq \beta_1, E\hat{\alpha}_2 \neq \beta_2$, tak také $\text{plim}\,\hat{\alpha}_1 \neq \beta_1, \text{plim}\,\hat{\alpha}_2 \neq \beta_2$. Míra nekonzistence nekonverguje k 0, i když rozsah vzorku $T \to \infty$.
2. I pokud jsou proměnné x_2 a x_3 nekorelované, tzn. Při $r_{23} = 0$, bude $\hat{\alpha}_1$ stále vychýlený, i když $\hat{\alpha}_2$ je nyní nestranný.
3. Reziduální rozptyl σ^2 je odhadnut nepřesně.
4. Obvykle užívané vyjádření pro rozptyl parametru $\hat{\alpha}_2$ $(= \sigma^2 / \sum x_{t2}^2)$ je vychýleným estimátorem rozptylu správného estimátoru $\hat{\beta}_2$.
5. Jako důsledek předchozího: procedury testování hypotéz a konstrukce intervalů spolehlivosti budou velmi pravděpodobně poskytovat scestné závěry, pokud jde o statistickou významnost odhadovaných parametrů: Lze ukázat, že

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 . b_{32} \quad , \quad \text{kde}$$

b_{32} je koeficient sklonu v regresi vyloučené proměnné x_3 na zařazenou proměnnou x_2:

$b_{32} = \sum x_{t3} x_{t2} / \sum x_{t2}^2$. Jestliže je $\beta_3 > 0$ a $b_{32} > 0$ (pozitivní korelovanost x_2 s x_3), pak odhad $\hat{\alpha}_2$ bude nadhodnocovat skutečnou hodnotu parametru β_2.

Obecně (pro k vysvětlujících proměnných):

Rozdělíme model na dvě skupiny vysvětlujících proměnných $X = (X_1, X_2)$, s celkovým jejich počtem $k = k_1 + k_2$ kde v k_1 sloupcích submatice X_1 jsou patřičné proměnné, zatímco matice X_2 obsahuje nepatřičných k_2 proměnných. V souladu s tím rozdělíme vektor parametrů $\beta = (\beta_1, \beta_2)$ na první subvektor o délce k_1 a druhý subvektor o délce k_2. Máme tedy přesně specifikovaný model

(6) $$y_{[T,1]} = X_{1[T,k_1]}\beta_{1[k_1,1]} + X_{2[T,k_2]}\beta_{2[k_2,1]} + \varepsilon_{[T,1]}$$

a oproti němu model s nesprávnou specifikací

(7) $$y_{[T,1]} = X_{1[T,k_1]}\beta_1^{*}{}_{[k_1,1]} + \eta_{[T,1]}$$

---

[2] Tedy ne nestranné.

Odhadovou funkcí OLS pro chybně specifikovaný model **(7)** lze psát jako:

$$\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'y = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$$
$$= \beta_1 + (X_1'X_1)^{-1}X_1'(X_2\beta_2 + \varepsilon) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$

takže $E(\hat{\beta}_1^*) = \beta_1 + P_2\beta_2$, kde $P_2 = (X_1'X_1)^{-1}X_1'X_2$ je matice regresních koeficientů z pomocné regrese proměnných X_2 na okruh proměnných v X_1. Velikost vychýlení je zde

(8) $$E(\hat{\beta}_1^*) - \beta_1 = \beta_1 + P_2\beta_2 - \beta_1 = P_2\beta_2 = (X_1'X_1)^{-1}X_1'X_2\beta_2$$

Odtud plyne, že vychýlení způsobené nezahrnutím některých důležitých vysvětlujících proměnných, je úměrné velikosti vektoru parametrů β_2 u vynechaných proměnných X_2 a stupni korelace mezi zahrnutými (v X_1) a nezahrnutými (v X_2) vysvětlujícími proměnnými. Vychýlení bude konvergovat k $\mathbf{0}$ jen tehdy, pokud bude platit $\text{plim}\, X_1'X_2 = 0$ pro $T \to \infty$.

3. ZAŘAZENÍ IRELEVANTNÍCH PROMĚNNÝCH

konkrétně (pro 2 vysvětlující proměnné):

Zde budeme naopak předpokládat, že správná podoba modelu je

(9) $$y_t = \beta_1 + \beta_1 x_{t2} + \varepsilon_t$$

zatímco my se pokoušíme kvantifikovat nepřesně specifikovaný model

(10) $$y_t = \alpha_1 + \alpha_2 x_{t2} + \alpha_3 x_{t3} + u_t$$

Odhady parametrů nepřesně specifikovaného modelu označme jako obvykle $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$.

Důsledky zařazení nadbytečné proměnné $x_{.3}$ jsou tyto:

1. Odhady parametrů (pořízené metodou OLS) chybně specifikovaného modelu jsou všechny nestranné a konzistentní. Pokud je vynechaná proměnná x_3 korelovaná se zařazenou proměnnou x_2, bude platit $E\hat{\alpha}_1 = \beta_1, E\hat{\alpha}_2 = \beta_2$, $\text{plim}\,\hat{\alpha}_1 = \beta_1, \text{plim}\,\hat{\alpha}_2 = \beta_2$ resp. též $E\hat{\alpha}_3 = 0\ (= \beta_3)$ $\text{plim}\,\hat{\alpha}_1 = \beta_1, \text{plim}\,\hat{\alpha}_2 = \beta_2$.
2. Reziduální rozptyl σ^2 je odhadnut přesně.
3. Konvenční postupy testování hypotéz a konstrukce intervalů spolehlivosti si zachovávají platnost.
4. Odhady parametrů $\hat{\alpha}_1, \hat{\alpha}_2$ budou zpravidla méně vydatné, tzn. jejich rozptyly budou obecně větší než u srovnatelných odhadů $\hat{\beta}_1, \hat{\beta}_2$ správně specifikovaného modelu. Srovnejme např. $\text{var}(\hat{\beta}_2) = \dfrac{\sigma^2}{\sum x_{t2}^2}$ $\text{var}(\hat{\alpha}_2) = \dfrac{\sigma^2}{\sum x_{t2}^2(1 - r_{23}^2)}$ a tedy
$$\frac{\text{var}(\hat{\alpha}_2)}{\text{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2} \geq 1$$

Zařazení nadbytečné proměnné tedy vykazuje znatelně méně slabin než vynechání důležité proměnné. Při větším počtu nadbytečných proměnných však mohou vzniknout problémy s *multikolinearitou* a ztrátou stupňů volnosti.

Obecně (pro k vysvětlujících proměnných):

Opět rozdělíme model na dvě skupiny vysvětlujících proměnných $X=(X_1,X_2)$, kde celkový počet vysvětlujících proměnných $k=k_1+k_2$ kde v k_1 sloupcích submatice X_1 jsou řádně zařazené (patřičné) proměnné, zatímco matice X_2 obsahuje (omylem doplněných) nepatřičných k_2 proměnných. V souladu s tím rozdělíme opět parametrický vektor $\beta=(\beta_1,\beta_2)$, kde subvektor β_1 má délku k_1, subvektor β_2 délku k_2.

Máme tedy přesně specifikovaný model

(11)
$$y_{[T,1]}=X_{1[T,k_1]}\beta_{1[k_1,1]}+\eta_{[T,1]}$$

a oproti němu model s nesprávnou specifikací (rozšířenou o k_2 .nepatřičných proměnných)

(12)
$$y_{[T,1]}=X_{1[T,k_1]}\beta^*{}_{1[k_1,1]}+X_{2[T,k_2]}\beta_2{}^*{}_{[k_2,1]}+\varepsilon_{[T,1]}$$
, jinak souhrnně

(12A)
$$y_{[T,1]}=X_{[T,k]}\beta^*{}_{[k,1]}+\varepsilon_{[T,1]}$$

Odhadovou funkcí OLS aplikovanou na chybně specifikovaný model **(12)** lze nyní psát jako

(13)
$$\hat{\beta}^*=\begin{pmatrix}\hat{\beta}^*{}_1\\ \hat{\beta}^*{}_2\end{pmatrix}=(X'X)^{-1}X'y=\begin{pmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{pmatrix}^{-1}\cdot\begin{pmatrix}X_1'\\ X_2'\end{pmatrix}(X_1\beta_1+\eta)$$

Porovnejme nyní tu část vektoru $\hat{\beta}^*$, která je společná s prvním modelem **(11)**:

K vektoru parametrů $\hat{\beta}_1{}^*$ se váže jen „horní“ část **(13)** $\hat{\beta}_1{}^*=(X'X)^{-1}X'X_1\beta_1$, kde můžeme psát $X_{1[T,k_1]}=X_{[T,k]}\cdot\begin{pmatrix}I_{k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}$ a tedy . $X_1'X_1=\left(I_{k_1},0_{k-k_1,k_1}\right)X.X.\begin{pmatrix}I_{k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}$

Podobně máme pro „dolní úsek“:

$$X_{2[T,k-k_1]}=X_{[T,k]}\cdot\begin{pmatrix}0_{k_1,k-k_1}\\ I_{k-k_1}\end{pmatrix}\text{ a následně }X_2'X_1=\left(0_{k-k_1,k_1},I_{k-k_1}\right)X.X.\begin{pmatrix}I_{k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}$$

$$E(\hat{\beta}_1{}^*)=(X'X)^{-1}X_1'X_1\beta_1=(X'X)^{-1}\left(I_{k_1},0_{k-k_1,k_1}\right)X.X.\begin{pmatrix}I_{k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}\beta_1=I_k\cdot\begin{pmatrix}I_{k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}\beta_1=\beta_1$$

. Dále máme

$$E(\hat{\beta}_2{}^*)=(X'X)^{-1}X_2'X_1\beta_1=(X'X)^{-1}\left(0_{k-k_1,k_1},I_{k-k_1}\right)X.X.\begin{pmatrix}I_{k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}\beta_1=I_k\cdot\begin{pmatrix}0_{k_1,k_1}\\ 0_{k-k_1,k_1}\end{pmatrix}\beta_1=0$$

Odtud vyplývá, že

A) Odhad vektoru patřičných parametrů $\hat{\beta}^*{}_1$ je nestranný.

B) Střední hodnota odhadu vektoru nepatřičných parametrů $\hat{\beta}_2{}^*$ je nulový vektor (to je příznivý výsledek, protože k parametrům příslušné proměnné nemají v modelu co dělat).

Dále platí, že:

- odhadová funkce rozptylu náhodných složek je nestranná
- zvětšují se výběrové rozptyly odhadnutých parametrů patřičných nezávisle proměnných.(může to ovlivnit výsledky testování), zhoršuje se tím vydatnost odhadů
- přítomnost nepatřičných proměnných zvyšuje riziko multikolinearity (a snižuje se počet stupňů volnosti)