

## 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

$$1) \quad y = X\beta + \varepsilon \quad \text{s obvyklými vlastnostmi LRM.}$$

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

$$2) \quad y = \tilde{X}\beta^* + \varepsilon^*$$

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

$$3) \quad \hat{\beta}^* = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y \quad \text{neboli po dosazení za y z 1)}$$

$$3a) \quad \hat{\beta}^* = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'(X\beta + \varepsilon).$$

Pro střední hodnotu tohoto vektoru parametrů platí

$$E(\hat{\beta}^*) = E[(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(X\beta + \varepsilon)] = E[(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\beta] + E[(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\varepsilon] = P\cdot\beta,$$

<sup>1</sup> Přirozeně, některé proměnné obsažené ve sloupcích obou matic mohou být (a zpravidla budou) společné.

kde maticí  $P = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'X$  je násoben vektor správných koeficientů  $\beta$ , leč  $P \neq I_k$

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

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

## 2. Vynechání relevantních proměnných

**Konkrétně** (pro 2 vysvětlující proměnné):

Předpokládejme, že místo správně specifikovaného modelu

$$4) \quad y_t = \beta_0 + \beta_1 x_{t2} + \beta_2 x_{t3} + \varepsilon_t$$

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

$$5) \quad y_t = \alpha_0 + \alpha_1 x_{t2} + u_t$$

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<sup>2</sup>, 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  $T \rightarrow \infty$ .
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 už nestranný.
3. Reziduální rozptyl  $\sigma^2$  bude 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

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 \cdot b_{32}$$

---

<sup>2</sup> Tedy ne nestranné.

kde  $b_{32}$  je koeficient sklonu v regresi vyloučené proměnné  $x_3$  na zařazenou proměnnou  $x_2$ :  $b_{32} = \frac{\sum x_{t3}x_{t2}}{\sum x_{t2}^2}$ . 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  $\beta_2$ .

**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 patříčné proměnné, zatímco matice  $X_2$  obsahuje  $k_2$  nepatříčných 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

$$6) \quad \mathbf{y}_{[T,1]} = \mathbf{X}_{1[T,k_1]} \beta_{1[k_1,1]} + \mathbf{X}_{2[T,k_2]} \beta_{2[k_2,1]} + \varepsilon_{[T,1]}$$

a oproti němu model s nesprávnou specifikací

$$7) \quad \mathbf{y}_{[T,1]} = \mathbf{X}_{1[T,k_1]} \beta_{1[k_1,1]}^* + \eta_{[T,1]}$$

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

$$\begin{aligned} \hat{\beta}_1^* &= (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{y} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' (\mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \varepsilon) \\ &= \beta_1 + (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' (\mathbf{X}_2 \beta_2 + \varepsilon) = \beta_1 + (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2 \beta_2 + (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \varepsilon \end{aligned}$$

takže  $E(\hat{\beta}_1^*) = \beta_1 + \mathbf{P}_2 \beta_2$ , kde  $\mathbf{P}_2 = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2$  je matice regresních koeficientů z pomocné regrese proměnných  $X_2$  na okruh proměnných v  $X_1$ . Velikost vychýlení je zde

$$8) \quad E(\hat{\beta}_1^*) - \beta_1 = \beta_1 + \mathbf{P}_2 \beta_2 - \beta_1 = \mathbf{P}_2 \beta_2 = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2 \beta_2$$

Odtud plyne, že vychýlení způsobené nezahrnutím některých důležitých vysvětlujících proměnných, je úměrné velikosti vektoru parametrů  $\beta_2$  u vynechaných proměnných  $X_2$  a stupni korelace mezi zahrnutými (v  $X_1$ ) a nezahrnutými (v  $X_2$ ) vysvětlujícími proměnnými. Vychýlení bude konvergovat k 0 jen tehdy, pokud bude platit  $\text{plim} \mathbf{X}_1' \mathbf{X}_2 = 0$  pro  $T \rightarrow \infty$ .

### 3. Zařazení irelevantních proměnných

**Konkrétně** (pro 2 vysvětlující proměnné):

Zde budeme naopak předpokládat, že správná podoba modelu je

$$9) \quad y_t = \beta_1 + \beta_2 x_{t2} + \varepsilon_t$$

zatímco my se pokoušíme kvantifikovat nepřesně specifikovaný model

$$10) \quad \mathbf{y}_t = \alpha_1 + \alpha_2 \mathbf{x}_{t2} + \alpha_3 \mathbf{x}_{t3} + \mathbf{u}_t$$

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) takto 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  $E\hat{\alpha}_3 = 0 (= \beta_3)$ .
2. Reziduální rozptyl  $\sigma^2$  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. Srovnajme např.  $\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum \mathbf{x}_{t2}^2}$

$$\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum \mathbf{x}_{t2}^2 (1 - r_{23}^2)} \quad \text{a tedy}$$

$$\frac{\text{var}(\hat{\alpha}_2)}{\text{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2} \geq 1$$

**Poznámka:** Zařazení nadbytečné proměnné tedy vykazuje ztlačení méně slabín 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 (patříčné) proměnné, zatímco matice  $X_2$  obsahuje (omylem doplněných) nepatříčných  $k_2$  proměnných. V souladu s tím rozdělíme opět parametrický vektor  $\beta = (\beta_1, \beta_2)$ , kde subvektor  $\beta_1$  má délku  $k_1$ , subvektor  $\beta_2$  délku  $k_2$ .

Máme tedy přesně specifikovaný model

$$11) \quad \mathbf{y}_{[T,1]} = \mathbf{X}_{1[T,k_1]} \beta_{1[k_1,1]} + \eta_{[T,1]}$$

a oproti němu model s nesprávnou specifikací (rozšířenou o  $k_2$  nepatříčných proměnných)

$$12) \quad \mathbf{y}_{[T,1]} = \mathbf{X}_{1[T,k_1]} \beta_{1[k_1,1]}^* + \mathbf{X}_{2[T,k_2]} \beta_{2[k_2,1]}^* + \varepsilon_{[T,1]} \quad , \quad \text{jinak}$$

souhrnně

$$12a) \quad \mathbf{y}_{[T,1]} = \mathbf{X}_{[T,k]} \beta_{[k,1]}^* + \varepsilon_{[T,1]}$$

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

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

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)  $\hat{\beta}_1^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_1\beta_1$ , kde můžeme

psát  $\mathbf{X}_{1[T,k_1]} = \mathbf{X}_{[T,k]} \cdot \begin{pmatrix} \mathbf{I}_{k_1} \\ 0_{k-k_1, k_1} \end{pmatrix}$  a tedy  $\mathbf{X}_1'\mathbf{X}_1 = \begin{pmatrix} \mathbf{I}_{k_1} & 0_{k-k_1, k_1} \end{pmatrix} \mathbf{X} \cdot \mathbf{X} \cdot \begin{pmatrix} \mathbf{I}_{k_1} \\ 0_{k-k_1, k_1} \end{pmatrix}$

Podobně máme pro „dolní úsek“:

$$\mathbf{X}_{2[T,k-k_1]} = \mathbf{X}_{[T,k]} \cdot \begin{pmatrix} 0_{k_1, k-k_1} \\ \mathbf{I}_{k-k_1} \end{pmatrix} \text{ a následně } \mathbf{X}_2'\mathbf{X}_2 = \begin{pmatrix} 0_{k-k_1, k_1} & \mathbf{I}_{k-k_1} \end{pmatrix} \mathbf{X} \cdot \mathbf{X} \cdot \begin{pmatrix} \mathbf{I}_{k_1} \\ 0_{k-k_1, k_1} \end{pmatrix}$$

$$\mathbb{E}(\hat{\beta}_1^*) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{X}_1\beta_1 = (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} \mathbf{I}_{k_1} & 0_{k-k_1, k_1} \end{pmatrix} \mathbf{X} \cdot \mathbf{X} \cdot \begin{pmatrix} \mathbf{I}_{k_1} \\ 0_{k-k_1, k_1} \end{pmatrix} \beta_1 = \mathbf{I}_{k_1} \cdot \begin{pmatrix} \mathbf{I}_{k_1} \\ 0_{k-k_1, k_1} \end{pmatrix} \beta_1 = \beta_1.$$

Dále máme

$$\mathbb{E}(\hat{\beta}_2^*) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2'\mathbf{X}_2\beta_1 = (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 0_{k-k_1, k_1} & \mathbf{I}_{k-k_1} \end{pmatrix} \mathbf{X} \cdot \mathbf{X} \cdot \begin{pmatrix} \mathbf{I}_{k_1} \\ 0_{k-k_1, k_1} \end{pmatrix} \beta_1 = \mathbf{I}_{k-k_1} \cdot \begin{pmatrix} 0_{k_1, k_1} \\ \mathbf{I}_{k-k_1, k_1} \end{pmatrix} \beta_1 = 0$$

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:

C) odhadová funkce rozptylu náhodných složek je nestranná

D) zvětšují se výběrové rozptyly odhadnutých parametrů patřičných nezávisle proměnných (může to ovlivnit výsledky testování), zhoršuje se tím vydatnost odhadů

E) přítomnost nepatřičných proměnných zvyšuje riziko *multikolinearity* (a snižuje se počet stupňů volnosti).