

# Problemy z danymi (cz. II)

**Natalia Nehrebecka**  
**Stanisław Cichocki**

Wykład 13

# Plan zajęć

1. Zmienne pominięte
2. Zmienne nieistotne
3. Obserwacje nietypowe i błędne
4. Współliniowość

# Plan zajęć

1. Zmienne pominięte
2. Zmienne nieistotne
3. Obserwacje nietypowe i błędne
4. Współliniowość

# Zmienne pominięte

- Mamy 2 modele:

$$y = X_1\beta_1 + u \quad (1)$$

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (2)$$

- Potencjalnie każdy z tych modeli może prawidłowo opisywać zmienną  $y$   $\longrightarrow$  problemy gdy przy liczeniu estymatorów zastosujemy niewłaściwy model

# Zmienne pominięte

- Załóżmy, że estymujemy model (1) a prawdziwy jest model (2)
- Zakładamy, że  $\beta_2 = 0$  gdy w rzeczywistości  $\beta_2 \neq 0$
- Przypadek ten nazywamy problemem **zmiennych pominiętych** (*omitted variables*)

# Zmienne pominięte

-  $\hat{\beta}_1$  - estymator MNK wektora parametrów w modelu (1)

- Załóżmy , że prawdziwy jest model (2)

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

# Zmienne pominięte

$$\begin{aligned} - \quad E(\hat{\beta}_1) &= \beta_1 + (X_1'X_1)^{-1} X_1'X_2\beta_2 + (X_1'X_1)^{-1} X_1'E(\varepsilon) = \\ &= \beta_1 + (X_1'X_1)^{-1} X_1'X_2\beta_2 \end{aligned}$$

- Jeśli więc pominiemy istotne zmienne, estymator nie jest estymatorem nieobciążonym

$$\text{- Obciążenie: } E(\hat{\beta}_1) - \beta_1 = (X_1'X_1)^{-1} X_1'X_2\beta_2$$

# Zmienne pominięte

- Dwa przypadki, dla których pominięcie zmiennej nie powoduje obciążenia estymatora

a)  $\beta_2 = 0$

b)  $X_1'X_2 = 0$  - zmienna pominięta nie jest skorelowana ze zmiennymi objaśniającymi, które zostały uwzględnione w modelu



# Zmienne pominięte

- Obciążenie może prowadzić do:

a) Uznania za zmienną istotną zmiennej, która nie ma żadnego wpływu na zmienna zależną **—————→ najgorszy przypadek**

b) **Przeszacowania/niedoszacowania** wpływu zmiennej objaśniającej na zmienną objaśnianą

# Zmienne pominięte

- Kierunek obciążenia dla najprostszego przypadku (model ze stałą i jedną zmienną objaśniającą, pominięta jedna dodatkowa zmienna objaśniająca):

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \frac{S_{x_2}}{S_{x_1}} \rho_{x_1 x_2}$$

gdzie:

$S_{x_1}, S_{x_2}$  – odchylenie standardowe  $x_1, x_2$

$\rho_{x_1, x_2}$  – wsp. korelacji między  $x_1$  a  $x_2$

# Zmienne pominięte

- ▶ Kierunek obciążenia dla najprostszego przypadku (model ze stałą i jedną zmienną objaśniającą, pominięta jedna dodatkowa zmienna objaśniająca):

Przypadek	Wpływ zmiennej pominiętej na zmienną zależną	Korelacja między zmienną pominiętą a zmienną niezależną	Znak obciążenia
I	+	+	+ (przeszacowanie)
II	-	-	+ (przeszacowanie)
III	+	-	- (niedoszacowanie)
IV	-	+	- (niedoszacowanie)

# Plan zajęć

1. Zmienne pominięte
2. Zmienne nieistotne
3. Obserwacje nietypowe i błędne
4. Współliniowość

# Zmienne nieistotne

- Mamy 2 modele:

$$y = X_1\beta_1 + u \quad (1)$$

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (2)$$

- Załóżmy, że estymujemy model (2) a prawdziwy jest model (1)

- Zakładamy, że  $\beta_2 \neq 0$  gdy w rzeczywistości  $\beta_2 = 0$

- Przypadek ten nazywamy problemem **zmiennych nieistotnych**

# Zmienne nieistotne

- ▶ Estymator  $\beta_1$  - **nieobciążony**, ale będzie miał **większą wariancję** niż estymator uzyskany na podstawie modelu (1)
- ▶ Inaczej mówiąc, w modelu w którym występują zmienne nieistotne estymator MNK ma wyższą wariancję niż w modelu, z którego usunięto zmienne nieistotne

# Zmienne nieistotne

- Usuwamy z modelu zmienne nieistotne bo:

a) **Poprawia to precyzję** oszacowań parametrów przy zmiennych istotnych (estymator MNK ma mniejszą wariancję)

b) Uzyskujemy **uproszczenie modelu**

# Pytania teoretyczne

1. Jaki skutek może mieć pominięcie istotnej zmiennej w modelu?
2. W jakim szczególnym przypadku można uzyskać prawidłowe oszacowania parametrów mimo, że w modelu pominięto zmienne?
3. Dlaczego z modelu powinno się usuwać zmienne nieistotne?
4. Parametry przy zmiennych  $x_1$  i  $x_2$  są dodatnie. Zmienne są ujemnie skorelowane. Jaki będzie wpływ pominięcia zmiennej  $x_1$  na oszacowanie parametru przy zmiennej  $x_2$ ?



**Dziękuję za uwagę**