Problemy z danymi (cz. I)

Natalia Nehrebecka
Stanisław Cichocki

Wykład 11

Plan zajęć

1. Zmienne pominięte

2. Zmienne nieistotne

3. Obserwacje nietypowe i błędne

4. Współliniowość

Zmienne pominięte

- Mamy 2 modele:

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

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

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

Zmienne pominięte

- Załóżmy, ze 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** (*ommitted variables*)

Zmienne pominięte

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

- Załóżmy , że prawdziwy jest model (2)

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

$$= \beta_1 + (X_1^{'}X_1)^{-1}X_1^{'}X_2\beta_2 + (X_1^{'}X_1)^{-1}X_1^{'}\varepsilon$$

Zmienne pominięte

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

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

- 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 skorelowane 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

ρ_{x_1, x_2} – wsp. korelacji miedzy 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:

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

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

- Załóżmy, ze 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 β_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 istotne zmienne?
3. Dlaczego z modelu powinno się usuwać zmienne nieistotne?
4. Parametry przy zmiennych x1 i x2 są dodatnie. Zmienne są ujemnie skorelowane. Jaki będzie wpływ pominięcia zmiennej x1 na oszacowanie parametru przy zmiennej x2?

Dziękuję za uwagę