Błąd zmiennej pominiętej

Ekonometria WNE

Sebastian Zalas

University of Warsaw

s.zalas@uw.edu.pl

Błąd zmiennej pominiętej

- Jesteśmy zainteresowani oszacowaniem jak $x \longrightarrow y$

$$y = \beta_0 + \beta_1 x + \varepsilon = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

- Prawdziwy model:

$$y = \beta_0 + \beta_1 x + \gamma z + \varepsilon = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\gamma + \varepsilon \tag{1}$$

z - pominięta zmienna; niech z będzie skorelowane z x

- Co stanie się z $\hat{\beta}$?

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{z}\gamma + \varepsilon) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon
\end{aligned}$$

Błąd zmiennej pominiętej

- Jesteśmy zainteresowani oszacowaniem jak $x \longrightarrow y$

$$y = \beta_0 + \beta_1 x + \varepsilon = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

- Prawdziwy model:

$$y = \beta_0 + \beta_1 x + \gamma z + \varepsilon = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\gamma + \varepsilon \tag{1}$$

z - pominięta zmienna; niech z będzie skorelowane z x

- Co stanie się z $\hat{\beta}$?

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\gamma + \varepsilon) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon
\end{aligned}$$

Błąd zmiennej pominiętej

- obciążenie:

$$\begin{aligned}\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] - \beta &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma \mid \boldsymbol{X}] + \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\gamma\,\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[\varepsilon \mid \boldsymbol{X}]}_{=0} \\ &= \gamma(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] \neq 0 \qquad (2)\end{aligned}$$

- Szacując model bez γz pozostanie on w składniku losowym:

$$y = \beta_0 + \beta_1 x + \underbrace{(\gamma z + \varepsilon)}_{=\nu}$$

- wtedy $\mathbb{E}[\nu \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] + \mathbb{E}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] \neq 0 \longrightarrow$ złamanie założenia KMRL prowadzi do **obciążenia** estymatora MNK

Błąd zmiennej pominiętej

- obciążenie:

$$\begin{aligned}\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] - \beta &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma \mid \boldsymbol{X}] + \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\gamma\,\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[\varepsilon \mid \boldsymbol{X}]}_{=0} \\ &= \gamma(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] \neq 0 \qquad (2)\end{aligned}$$

- Szacując model bez γz pozostanie on w składniku losowym:

$$y = \beta_0 + \beta_1 x + \underbrace{(\gamma z + \varepsilon)}_{=\nu}$$

- wtedy $\mathbb{E}[\nu \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] + \mathbb{E}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] \neq 0 \longrightarrow$ złamanie założenia KMRL prowadzi do **obciążenia** estymatora MNK

Błąd zmiennej pominiętej

- obciążenie:

$$\begin{aligned}\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] - \beta &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma \mid \boldsymbol{X}] + \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\gamma\, \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{\mathbb{E}[\varepsilon \mid \boldsymbol{X}]}_{=0} \\ &= \gamma(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\, \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] \neq 0 \end{aligned} \tag{2}$$

- Szacując model bez γz pozostanie on w składniku losowym:

$$y = \beta_0 + \beta_1 x + \underbrace{(\gamma z + \varepsilon)}_{=\nu}$$

- wtedy $\mathbb{E}[\nu \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] + \mathbb{E}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] \neq 0 \longrightarrow$ złamanie założenia KMRL prowadzi do **obciążenia** estymatora MNK

Błąd zmiennej pominiętej

- obciążenie:

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] - \boldsymbol{\beta} &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}\gamma \mid \boldsymbol{X}] + \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\gamma\, \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{\mathbb{E}[\varepsilon \mid \boldsymbol{X}]}_{=0} \\ &= \gamma(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\, \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}] \neq 0 \end{aligned} \tag{2}$$

- Szacując model bez γz pozostanie on w składniku losowym:

$$y = \beta_0 + \beta_1 x + \underbrace{(\gamma z + \varepsilon)}_{=\nu}$$

- wtedy $\mathbb{E}[\nu \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] + \mathbb{E}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\gamma z \mid \boldsymbol{X}] \neq 0 \longrightarrow$ złamanie założenia KMRL prowadzi do **obciążenia** estymatora MNK

Błąd zmiennej pominiętej

- Kiedy błąd zmiennej pominiętej zachodzi? Przekształcimy (2):

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] - \boldsymbol{\beta} &= \boldsymbol{\gamma} \frac{\mathrm{Cov}[z, x]}{\mathbb{V}[x]} \\ &= \gamma \frac{\sigma_z}{\sigma_x} \rho_{x,z} \end{aligned} \tag{3}$$

- x jest skorelowana z pominiętą zmienną z, czyli

$$\rho_{x,z} \neq 0$$

- pominięta zmienna z jest skorelowana ze zmienną zależną y, czyli:

$$\gamma \neq 0$$

Błąd zmiennej pominiętej

- Kiedy błąd zmiennej pominiętej zachodzi? Przekształcimy (2):

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] - \boldsymbol{\beta} &= \boldsymbol{\gamma} \frac{\mathrm{Cov}[z, x]}{\mathbb{V}[x]} \\ &= \gamma \frac{\sigma_z}{\sigma_x} \rho_{x,z} \end{aligned} \tag{3}$$

- x jest skorelowana z pominiętą zmienną z, czyli

$$\rho_{x,z} \neq 0$$

- pominięta zmienna z jest skorelowana ze zmienną zależną y, czyli:

$$\gamma \neq 0$$

Błąd zmiennej pominiętej

- Kiedy błąd zmiennej pominiętej zachodzi? Przekształcimy (2):

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] - \boldsymbol{\beta} &= \boldsymbol{\gamma} \frac{\mathrm{Cov}[z, x]}{\mathbb{V}[x]} \\ &= \gamma \frac{\sigma_z}{\sigma_x} \rho_{x,z} \end{aligned} \tag{3}$$

- x jest skorelowana z pominiętą zmienną z, czyli

$$\rho_{x,z} \neq 0$$

- pominięta zmienna z jest skorelowana ze zmienną zależną y, czyli:

$$\gamma \neq 0$$

Błąd zmiennej pominiętej - wnioski

- Błąd zmiennej pominiętej zachodzi gdy oba warunki są spełnione (via (3)):
  1. zmienna w modelu jest skorelowana z pominiętą zmienną
  2. pominięta zmienna jest skorelowana ze zmienną zależną

- Pominięcie zmiennej w modelu prowadzi do obciążenia estymatora MNK z powodu złamania założenia KMRL: $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0$

Błąd zmiennej pominiętej - wnioski

- Błąd zmiennej pominiętej zachodzi gdy oba warunki są spełnione (via (3)):
  1. zmienna w modelu jest skorelowana z pominiętą zmienną
  2. pominięta zmienna jest skorelowana ze zmienną zależną

- Pominięcie zmiennej w modelu prowadzi do obciążenia estymatora MNK z powodu złamania założenia KMRL: $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0$

Błąd zmiennej pominiętej - przykład

- wpływ *profocc* na płace
  - *profocc* =1 gdy osoba pracuje w wysoko wyspecjalizowanym zawodzie

- oszacujmy dwa modele:

$$\log(wage) = profocc + exper + expersq + female + \varepsilon$$

$$\log(wage) = profocc + educ + exper + expersq + female + \varepsilon$$

- Czy edukacja może być pominięta? Czy są inne możliwości?

Błąd zmiennej pominiętej - przykład

- wpływ *profocc* na płace
  - *profocc* =1 gdy osoba pracuje w wysoko wyspecjalizowanym zawodzie

- oszacujmy dwa modele:

$$\log(wage) = profocc + exper + expersq + female + \varepsilon$$

$$\log(wage) = profocc + educ + exper + expersq + female + \varepsilon$$

- Czy edukacja może być pominięta? Czy są inne możliwości?

Błąd zmiennej pominiętej - przykład

- wpływ *profocc* na płace
  - *profocc* =1 gdy osoba pracuje w wysoko wyspecjalizowanym zawodzie

- oszacujmy dwa modele:

$$\log(wage) = profocc + exper + expersq + female + \varepsilon$$

$$\log(wage) = profocc + educ + exper + expersq + female + \varepsilon$$

- Czy edukacja może być pominięta? Czy są inne możliwości?

Błąd zmiennej pominiętej - przykład

	Zmienna zależna:	
	log(*wage*)	
	(1)	(2)
profocc	0.402***	0.227***
	(0.040)	(0.043)
educ		0.063***
		(0.008)
exper	0.038***	0.037***
	(0.005)	(0.005)
expersq	−0.001***	−0.001***
	(0.0001)	(0.0001)
female	−0.312***	−0.310***
	(0.038)	(0.036)
N	526	526
R^2	0.359	0.430
Note:	*p<0.1; **p<0.05; ***p<0.01	

Błąd zmiennej pominiętej - przykład

Edukacja jest pominiętym czynnikiem:

- edukacja jest skorelowana z płacą
- *profocc* jest skorelowane z edukacją

Współczynnik przy *profocc* w kol. (1):

- jest obciążony (przeszacowany):

$$\hat{\beta}_{profocc} - \beta_{profocc} = \beta_{educ} \frac{\sigma_{educ}}{\sigma_{profocc}} \rho_{educ,profocc} > 0$$

$$\hat{\beta}_{profocc} > \beta_{profocc}$$

- mierzy wpływ pracy w wysoko wyspecjalizowanym zawodzie oraz edukacji

Błąd zmiennej pominiętej - przykład

Edukacja jest pominiętym czynnikiem:

- edukacja jest skorelowana z płacą
- *profocc* jest skorelowane z edukacją

Współczynnik przy *profocc* w kol. (1):

- jest obciążony (przeszacowany):

$$\hat{\beta}_{profocc} - \beta_{profocc} = \beta_{educ} \frac{\sigma_{educ}}{\sigma_{profocc}} \rho_{educ,profocc} > 0$$

$$\hat{\beta}_{profocc} > \beta_{profocc}$$

- mierzy wpływ pracy w wysoko wyspecjalizowanym zawodzie oraz edukacji

Błąd zmiennej pominiętej - co zrobić?

Jak go wykryć?

- nie da się przetestować czy pominięto ważną zmienną
- należy bazować na własnej ocenie/ teorii/ literaturze

Co robić?

- warto wyznaczyć, który współczynnik jest interesujący
- włączyć do modelu zmienną *proxy* (zastępczą) dla pominiętej
- można wyznaczyć kierunek obciążenia korzystając z (3) oraz ocenić jego siłę

Błąd zmiennej pominiętej - co zrobić?

Jak go wykryć?

- nie da się przetestować czy pominięto ważną zmienną
- należy bazować na własnej ocenie/ teorii/ literaturze

Co robić?

- warto wyznaczyć, który współczynnik jest interesujący
- włączyć do modelu zmienną *proxy* (zastępczą) dla pominiętej
- można wyznaczyć kierunek obciążenia korzystając z (3) oraz ocenić jego siłę

Błąd zmiennej pominiętej - co zrobić?

Przykład zmiennej *proxy* zastępczej:

- cena samochodu zależy od wieku
- wieku samochodu nie obserwujemy
- może obserwujemy jak długo posiada go obecny właściciel $\leftarrow$ zmienna proxy

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl