Problemy z danymi (cz. I)

Natalia Nehrebecka
Stanisław Cichocki

Wykład 12

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:

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

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

- 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 zmienna objaśnianą

Przykład

- **Zmienna objaśniana:** $\ln\_dochod_i$ - logarytm naturalny płacy miesięcznej i–tej osoby.
- Zmiennymi objaśniającymi są:
  - **plec**$_i$ - płeć i-tej osoby.
  - **wiek**$_i$ – wiek i-tej osoby mierzony w latach.
  - **wiek**^{2} - wiek do kwadratu i-tej osoby mierzony w latach.
  - **srednie**$_i$ = 1 jeśli i–ta osoba ma wykształcenie średnie oraz $srednie_i = 0$ jeśli i – ta osoba ma wykształcenie inne niż średnie;
  - **wyzsze**$_i$ = 1 jeśli i–ta osoba ma wykształcenie wyższe oraz $wyzsze_i = 0$ jeśli i–ta osoba ma wykształcenie inne niż wyższe;
  - **malemiasto** = 1 jeśli i-ta osoba mieszka w mieście do 20 tys. mieszkańców;
  - **sredniemiasto** = 1 jeśli i-ta osoba mieszka w mieście od 20 tys. do 100 tys. mieszkanców;
  - **duzemiasto** = 1 jeśli i-ta osoba mieszka w mieście powyżej 100 tys. mieszkańców.

Przykład

```
xi: regress ln_dochod wiek wiek_2 plec  miasto_male miasto_srednie miasto_duze srednie
            wyzsze
      Source |       SS       df       MS              Number of obs =    1083
-------------+------------------------------           F(  8,  1074) =   39.86
       Model |  84.6530708     8  10.5816339           Prob > F      =  0.0000
    Residual |  285.126101  1074  .265480541           R-squared     =  0.2289
-------------+------------------------------           Adj R-squared =  0.2232
       Total |  369.779172  1082  .341755242           Root MSE      =  .51525

------------------------------------------------------------------------------
   ln_dochod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .0482201   .0100829     4.78   0.000     .0284357    .0680045
      wiek_2 |  -.0005786   .0001286    -4.50   0.000    -.0008309   -.0003263
        plec |  -.3482293   .0315385   -11.04   0.000    -.4101134   -.2863453
 miasto_male |   .0203816   .0179177     4.25   0.000     .0109793     .297839
miasto_sre~e |   .0498977   .0460108     1.08   0.278    -.0403835    .1401789
 miasto_duze |   .0960403   .0450749     2.13   0.033     .0075954    .1844851
     srednie |   .1578171   .0522706     3.02   0.003      .055253    .2603813
     wyzsze |   .5334086   .0681366     7.83   0.000     .3997126    .6671046
       _cons |   5.357663   .1943577    27.57   0.000     4.976299    5.739027
------------------------------------------------------------------------------
```

Przykład

Sprawdzimy jaki wpływ na wyniki oszacowań ma pominiecie istotnych zmiennych objaśniających – *usuniemy z modelu zmienne dotyczące wykształcenia*.

Przykład

```
xi: regress ln_dochod wiek wiek_2 plec  miasto_male miasto_srednie miasto_duze
      Source |       SS       df       MS              Number of obs =    1083
-------------+------------------------------           F(  6,  1076) =   38.47
       Model |   65.308758     6   10.884793            Prob > F      =  0.0000
    Residual |  304.470414  1076  .282965068            R-squared     =  0.1766
-------------+------------------------------           Adj R-squared =  0.1720
       Total |  369.779172  1082  .341755242            Root MSE      =  .53194

------------------------------------------------------------------------------
   ln_dochod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .0518353   .0103932     4.99   0.000      .031442    .0722286
      wiek_2 |  -.0006269   .0001324    -4.74   0.000    -.0008866   -.0003672
        plec |  -.3519787   .0325573   -10.81   0.000    -.4158616   -.2880957
 miasto_male |   .0457722   .0484341     5.49   0.000     .02707362    .2608083
miasto_sre~e |   .0698393   .0474865     1.26   0.208    -.0333373    .1530158
 miasto_duze |   .1271572   .0463735     2.74   0.006     .0361644      .21815
       _cons |   5.428209   .1972393    27.52   0.000     5.041192    5.815227
------------------------------------------------------------------------------
```

Przykład

- Usunięcie z modelu zmiennych dotyczących wykształcenia spowodowało, iż uzyskaliśmy inne wartości oszacowanych parametrów.
- Największe różnice można zaobserwować w przypadku ocen przy zmiennych dotyczących **miejsca zamieszkania**.
  - Ponieważ wiemy, iż wykształcenie ma istotny wpływ na płace, więc uzyskane oceny dla modelu z restrykcjami z teoretycznego punktu widzenia należy traktować jako **obciążone**.

Oceny przy zmiennych	w modelu bez restrykcji	w modelu z restrykcjami
malemiasto	.0203816	.0457722
sredniemiasto	.0498977	.0698393
duzemiasto	.0960403	.1271572

  - Dodatnie obciążenie wynika z faktu, iż wykształcenie jest dodatnio skorelowane z miejscem zamieszkania – największy odsetek osób z wyższym wykształceniem jest w dużych miastach.

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)

Przykład 1

Source	SS	df	MS	Number of obs =	371
Model	10.6867044	1	10.6867044	F(1, 369) =	45.48
Residual	86.7135186	369	.234995985	Prob > F =	0.0000
Total	97.400223	370	.263243846	R-squared =	0.1097
				Adj R-squared =	0.1073
				Root MSE =	.48476

logrincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0650681	.0096489	6.74	0.000	.0460944	.0840417
_cons	5.549348	.1175245	47.22	0.000	5.318247	5.78045

2. Jaki będzie prawdopodobny kierunek obciążenia oszacowania parametru przy zmiennej educ wynikły z pominięcia:

   (a) inteligencji respondentki.

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 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