Problemy z danymi
Testy diagnostyczne
Konsekwencje niesferyczności błędu losowego

Stanisław Cichocki

Natalia Nehrebecka

Wykład 13

Plan wykładu

- 1.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
- 2. Od ogólnego do szczególnego
- 3. Autokorelacja
  - Testowanie autokorelacji
- 4. Konsekwencje heteroskedastyczności i autokorelacji

Plan wykładu

- 1.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
- 2. Od ogólnego do szczególnego
- 3. Autokorelacja
  - Testowanie autokorelacji
- 4. Konsekwencje heteroskedastyczności i autokorelacji

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

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

Zmienne pominięte

- 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) $X_1^{'} X_2 = 0$ - zmienne pominięte nie są skorelowane ze zmiennymi objaśniającymi, które zostały uwzględnione w modelu

Zmienne pominięte

- Pominięcie istotnych zmiennych jest prawdopodobnie najczęstszym powodem błędów w oszacowaniach

- W praktyce nigdy nie dysponujemy danymi odnośnie wszystkich zmiennych mogących wpływać na zmienną zależną

- W takim przypadku warto umieć określić kierunek ewentualnego obciążenia (trudne w ogólnym przypadku)

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

Zmienne pominięte

- Przykład:

Dla pewnej badanej grupy osób przeprowadzono regresję logarytmu wynagrodzenia na latach nauki (zmienna *latanauki)*. Jaki będzie prawdopodobny kierunek obciążenia parametru przy zmiennej *latanauki* wynikający z pominięcia:

a) wielkości miejscowości, w której zamieszkuje badana osoba;

b) liczby dzieci badanej osoby?

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} - wariancja empiryczna x_1, x_2

$\rho_{x_1 x_2}$ - 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 (β_2)	Korelacja między zmienną pominiętą a zmienną niezależną (ρ)	Znak obciążenia
I	+	+	+ (przeszacowanie)
II	-	-	+
III	+	-	- (niedoszacowanie)
IV	-	+	-

Zmienne pominięte

- Przykład

reg wydg dochg

```
      Source |       SS       df       MS              Number of obs =   31679
-------------+------------------------------           F(  1, 31677) =21732.03
       Model |  2.3577e+10     1  2.3577e+10           Prob > F      =  0.0000
    Residual |  3.4367e+10 31677  1084914.37           R-squared     =  0.4069
-------------+------------------------------           Adj R-squared =  0.4069
       Total |  5.7944e+10 31678  1829163.02           Root MSE      =  1041.6

------------------------------------------------------------------------------
        wydg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dochg |   .5879668   .0039884   147.42   0.000     .5801493    .5957843
       _cons |   712.8104   10.01991    71.14   0.000      693.171    732.4498
------------------------------------------------------------------------------
```

Zmienne pominięte

- Przykład

reg wydg dochg los

```
      Source |       SS       df       MS              Number of obs =   31679
-------------+------------------------------           F(  2, 31676) =11107.42
       Model |  2.3886e+10     2  1.1943e+10           Prob > F      =  0.0000
    Residual |  3.4059e+10 31676  1075214.71           R-squared     =  0.4122
-------------+------------------------------           Adj R-squared =  0.4122
       Total |  5.7944e+10 31678  1829163.02           Root MSE      =  1036.9

------------------------------------------------------------------------------
        wydg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dochg |   .5688205   .0041284   137.78   0.000     .5607287    .5769123
         los |   65.35337   3.859286    16.93   0.000     57.78902    72.91772
       _cons |   548.4807   13.91655    39.41   0.000     521.2037    575.7577
------------------------------------------------------------------------------
```

Zmienne nieistotne

- Mamy 2 modele:

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

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

- 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**

Plan wykładu

- 1.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
- 2. Od ogólnego do szczególnego
- 3. Autokorelacja
  - Testowanie autokorelacji
- 4. Konsekwencje heteroskedastyczności i autokorelacji

Metoda od ogólnego do szczególnego

- Polega na stopniowym upraszczaniu możliwie najogólniejszego modelu początkowego poprzez narzucanie coraz bardziej rozbudowanych ograniczeń

- Modele powstałe poprzez narzucenie ograniczeń są szczególnymi przypadkami modelu ogólnego

Metoda od ogólnego do szczególnego

```
      Source |       SS       df       MS              Number of obs =   27047
-------------+------------------------------           F(  7, 27039) =   22.07
       Model |  3.6678e+11     7  5.2397e+10           Prob > F      =  0.0000
    Residual |  6.4198e+13 27039  2.3743e+09           R-squared     =  0.0057
-------------+------------------------------           Adj R-squared =  0.0054
       Total |  6.4565e+13 27046  2.3872e+09           Root MSE      =   48726

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -97.68195   26.49075    -3.69   0.000    -149.6052    -45.7587
        plec |  -2712.252   608.8329    -4.45   0.000    -3905.596   -1518.908
czasnieokres |   4886.704   723.1053     6.76   0.000     3469.381    6304.028
      miasto |   5641.536    617.102     9.14   0.000     4431.984    6851.088
   pelenetat |  -919.0806   1215.741    -0.76   0.450    -3301.997    1463.836
agencjapracy |   231.9189     4022.4     0.06   0.954    -7652.194    8116.031
    prywatny |  -525.3708    668.799    -0.79   0.432    -1836.251    785.5098
       _cons |    54424.2   1857.304    29.30   0.000     50783.78    58064.61
------------------------------------------------------------------------------
```

Metoda od ogólnego do szczególnego

Source	SS	df	MS		
Model	3.6678e+11	7	5.2397e+10	Number of obs	= 27047
Residual	6.4198e+13	27039	2.3743e+09	F(7, 27039)	= 22.07
Total	6.4565e+13	27046	2.3872e+09	Prob > F	= 0.0000
				R-squared	= 0.0057
				Adj R-squared	= 0.0054
				Root MSE	= 48726

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wiek	-97.68195	26.49075	-3.69	0.000	-149.6052	-45.7587
plec	-2712.252	608.8329	-4.45	0.000	-3905.596	-1518.908
czasnieokres	4886.704	723.1053	6.76	0.000	3469.381	6304.028
miasto	5641.536	617.102	9.14	0.000	4431.984	6851.088
pelenetat	-919.0806	1215.741	-0.76	0.450	-3301.997	1463.836
agencjapracy	231.9189	4022.4	0.06	0.954	-7652.194	8116.031
prywatny	-525.3708	668.799	-0.79	0.432	-1836.251	785.5098
_cons	54424.2	1857.304	29.30	0.000	50783.78	58064.61

Metoda od ogólnego do szczególnego

```
test agencjapracy

( 1)  agencjapracy = 0

      F(  1, 27039) =    0.00
           Prob > F =    0.9540


test agencjapracy pelenetat

( 1)  agencjapracy = 0
( 2)  pelenetat = 0

      F(  2, 27039) =    0.29
           Prob > F =    0.7508
```

Metoda od ogólnego do szczególnego

```
test agencjapracy pelenetat prywatny

 ( 1)  agencjapracy = 0
 ( 2)  pelenetat = 0
 ( 3)  prywatny = 0

       F(  3, 27039) =    0.39
            Prob > F =    0.7591
```

Metoda od ogólnego do szczególnego

```
      Source |       SS       df       MS              Number of obs =   27047
-------------+------------------------------           F(  4, 27042) =   38.33
       Model |  3.6399e+11     4  9.0998e+10           Prob > F      =  0.0000
    Residual |  6.4201e+13 27042  2.3741e+09           R-squared     =  0.0056
-------------+------------------------------           Adj R-squared =  0.0055
       Total |  6.4565e+13 27046  2.3872e+09           Root MSE      =   48725

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -92.58908   26.05748    -3.55   0.000    -143.6631   -41.51507
        plec |  -2572.807   593.9404    -4.33   0.000    -3736.961   -1408.653
czasnieokres |   4883.076   693.7924     7.04   0.000     3523.207    6242.945
      miasto |   5676.406   616.2007     9.21   0.000      4468.62    6884.191
       _cons |   52915.06    1181.11    44.80   0.000     50600.02    55230.09
------------------------------------------------------------------------------
```

Metoda od ogólnego do szczególnego

```
test agencjapracy pelenetat prywatny plec

 ( 1)  agencjapracy = 0
 ( 2)  pelenetat = 0
 ( 3)  prywatny = 0
 ( 4)  plec = 0

       F(  4, 27039) =    4.98
            Prob > F =    0.0005
```

Plan wykładu

- 1.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
- 2. Od ogólnego do szczególnego
- 3. Autokorelacja
  - Testowanie autokorelacji
- 4. Konsekwencje heteroskedastyczności i autokorelacji

Autokorelacja

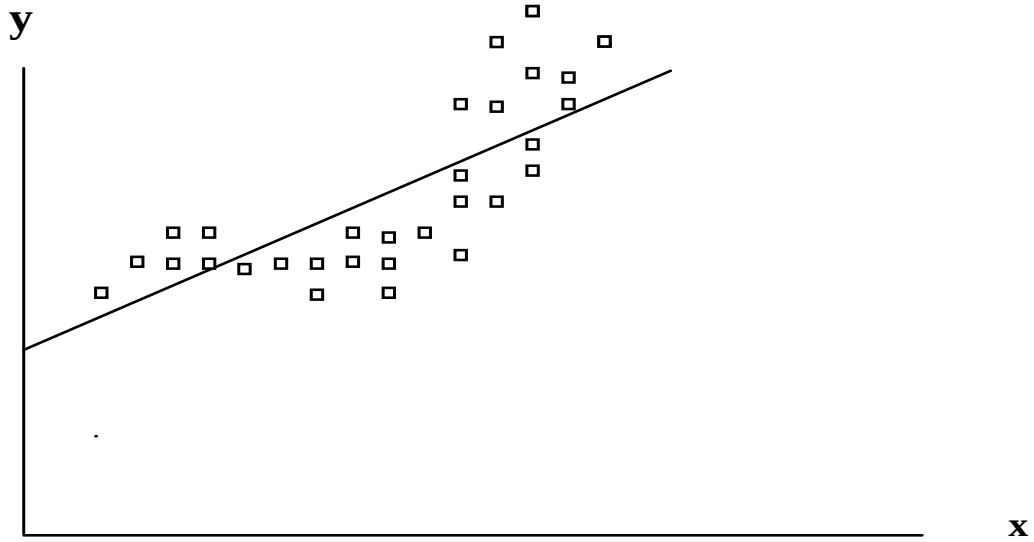
Przypomnienie: Co to znaczy, że w modelu występuje autokorelacja?

-Brak autokorelacji

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_2, \varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & Cov(\varepsilon_n, \varepsilon_1) & \cdots & Var(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Autokorelacja

- Przypadek zerowych kowariancji dla różnych zaburzeń losowych ε_i oraz ε_j nazywamy brakiem autokorelacji zaburzeń. Oznacza to, że **zaburzenia losowe dla różnych obserwacji są niezależne**, a przez to nieskorelowane, a więc nie mają tendencji do gromadzenia się np. wokół dodatnich lub ujemnych (lub naprzemiennie dodatnich i ujemnych) wartości



Rys. 2. **Autokorelacja**

Autokorelacja

$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) > 0$ dla $i \neq j$ - dodatnia autokorelacja

$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) < 0$ dla $i \neq j$ - ujemna autokorelacja

Testy diagnostyczne

- Dla każdego testu:

1. Nazwa testu
2. Hipotezy
3. Jakie założenie KMRL nie jest spełnione w przypadku odrzucenia H0?
4. Jakie są konsekwencje niespełnienia założenia KMRL?
5. W jaki sposób można rozwiązać problemy zasygnalizowane przez wynik testu?

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG):

$$H_0 : Cov(\varepsilon_t, \varepsilon_{t-i}) = 0 \qquad \textbf{gdzie} \qquad i = 1, \ldots, s$$

$$H_1 : \varepsilon_t = \gamma_1 \varepsilon_{t-1} + \ldots + \gamma_s \varepsilon_{t-s} + u_t \quad \textbf{gdzie} \quad Var(u) = \sigma_u^2 I$$

- Hipoteza zerowa: brak autokorelacji

- Hipoteza alternatywna: autokorelacja

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG) – sposób przeprowadzenia testu:

1. przeprowadzamy regresję y_i na x_i i uzyskujemy reszty

2. przeprowadzamy regresję pomocniczą:

$$e_t = x_t \hat{\mu} + \hat{\gamma}_1 e_{t-1} + \ldots + \hat{\gamma}_s e_{t-s} + \hat{u}_t$$

i testujemy H0: $\gamma_1 = \ldots = \gamma_s = 0$

Testowanie autokorelacji

- Statystyka testowa:

$$LM = TR^2 \xrightarrow{D} \chi_p^2$$

lub statystyka F

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG):

- Do badania autokorelacji wyższego rzędu

- Można go stosować w modelach gdzie występują opóźnione zmienne zależne

- Dla tego testu znany jest jedynie asymptotyczny rozkład statystyki testowej

Jakie założenie KMRL nie jest spełnione przy odrzuceniu H0?

- Brak autokorelacji błędu losowego – kowariancja dwóch różnych błędów losowych jest zerowa:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad dla \quad i \neq j$$

Plan wykładu

- 1.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
- 2. Od ogólnego do szczególnego
- 3. Autokorelacja
  - Testowanie autokorelacji
- 4. Konsekwencje heteroskedastyczności i autokorelacji

Sferyczność błędów losowych

- Jeżeli założenie o homoskedastyczności i braku autokorelacji jest spełnione to błędy losowe są **sferyczne**

- Jeżeli, któreś z tych założeń nie jest spełnione to błędy losowe są **niesferyczne** a macierz wariancji i kowariancji ma postać dowolnej macierzy symetrycznej i dodatnio półokreślonej:

$$Var(\varepsilon) = \Omega = \sigma^2 V$$

Konsekwencje heteroskedastyczności i autokorelacji

- Estymator *b* jest nadal **nieobciążony**:

$$E(b) = E\left[(X'X)^{-1}X'y\right] =$$

$$E\left[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\right] =$$

$$\beta + (X'X)^{-1}X'E(\varepsilon) = \beta$$

- Nie będzie on jednak **efektywny** ⟶ można znaleźć estymator o mniejszej wariancji

Konsekwencje heteroskedastyczności i autokorelacji

- Macierz wariancji i kowariancji *b*:

$$Var(b) = E\left((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right) =$$

$$(X'X)^{-1}X'\Omega X(X'X)^{-1} =$$

$$\sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$$

- Wzór ten różni się znacznie od prawidłowego wzoru na wariancję MNK:

$$Var(b) = \sigma^2(X'X)^{-1}$$

Konsekwencje heteroskedastyczności i autokorelacji

- W rezultacie **estymator macierzy wariancji i kowariancji *b***, którym posługiwaliśmy się do tej pory, **nie będzie dobrym** oszacowaniem macierzy wariancji i kowariancji *b*

- Konsekwencje **braku zgodności** estymatora macierzy wariancji i kowariancji *b* mogą być poważne:
  - estymatora macierzy wariancji i kowariancji *b* używamy przy konstruowaniu praktycznie wszystkich statystyk testowych
  - brak jego zgodności implikuje, że używanie standardowych statystyk testowych może doprowadzić do **błędnych wyników wnioskowania statystycznego**

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 oszacowania parametrów przy zmiennej x_2?

Pytania teoretyczne

5. Za pomocą jakiego testu testuje się autokorelację? Jakiemu założeniu KMRL odpowiada H0 w tym teście? Jaka jest hipoteza alternatywna w tym teście?
6. Jak niesferyczność błędów losowych wpływa na własnosći MNK?

Dziękuję za uwagę