Testy diagnostyczne
Problemy z danymi

Stanisław Cichocki

Natalia Nehrebecka

Wykład 12

Plan wykładu

- 1. Testy diagnostyczne
  - Testowanie heteroskedastyczności
  - Testowanie autokorelacji
- 2.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne

Plan wykładu

- 1. Testy diagnostyczne
  - Testowanie heteroskedastyczności
  - Testowanie autokorelacji
- 2.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne

Testowanie heteroskedastyczności

Przypomnienie: Co to znaczy, że w modelu występuje homoskedastyczność/heteroskedastyczność?

- heteroskedastyczność

$$Var(\varepsilon)=\begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & \cdots & Cov(\varepsilon_1,\varepsilon_n) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2,\varepsilon_n) \\ \vdots & \vdots & & \vdots \\ Cov(\varepsilon_n,\varepsilon_1) & Cov(\varepsilon_n,\varepsilon_1) & \cdots & Var(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Testowanie heteroskedastyczności

- Test Goldfelda-Quandta (Test GQ):

$$H_0 : Var(\varepsilon_i) = \sigma^2 \qquad \text{dla} \quad i = 1, \ldots, N$$

$$H_1 : Var(\varepsilon_i) > Var(\varepsilon_j) \qquad \text{dla} \quad z_i > z_j$$

gdzie z jest pewną zmienną, od której zależy wariancja błędu losowego

- Hipoteza zerowa: homoskedastyczność

- Hipoteza alternatywna: heteroskedastyczność

Testowanie heteroskedastyczności

- Test Goldfelda-Quandta (Test GQ):
  - z jego konstrukcji wynika, iż można go stosować do wykrywania zależności między wariancją błędu losowego a wielkością jednej zmiennej
  - jako jedyny z testów na heteroskedastyczność ma rozkład wyprowadzony dla małych prób

Testowanie heteroskedastyczności

- Test Breuscha-Pagana (Test BP):

$$H_0 : Var(\varepsilon_i) = \sigma^2 \qquad \text{dla } i = 1, \ldots, N$$

$$H_1 : Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 f(\alpha_0 + z_i \alpha)$$

gdzie $f(\bullet)$ **- funkcja różniczkowalna**

z_i - wektor zmiennych, może zawierać zmienne występujące w wektorze zmiennych objaśniających

Testowanie heteroskedastyczności

- Test Breuscha-Pagana (Test BP):

  - Hipoteza zerowa: homoskedastyczność

  - Hipoteza alternatywna: heteroskedastyczność

  - Szczególnie przydatny, jeżeli wariancja błędu losowego zależy od kilku zmiennych

Testowanie heteroskedastyczności

- Test Breuscha-Pagana (Test BP) – sposób przeprowadzenia testu:

1. przeprowadzamy regresję y_i na x_i i uzyskujemy e_i

2. przeprowadzamy regresję pomocniczą:

$$\frac{e_i^2}{\tilde{\sigma}^2} = \hat{\alpha}_0 + z_i \hat{\alpha} + \hat{u}_i$$

i testujemy H0: $\alpha = 0$

Testowanie heteroskedastyczności

- Statystyka testowa:

$$LM = \frac{1}{2} ESS \xrightarrow{D} \chi^2_p$$

gdzie: ESS – wyjaśniona suma kwadratów w regresji pomocniczej

p- ilość zmiennych zawartych w Z

Testowanie heteroskedastyczności

- Inna statystyka testowa:

$$LM = NR^2 \xrightarrow{D} \chi^2_p$$

gdzie: R^2 - współczynnik determinacji z regresji pomocniczej

Testowanie heteroskedastyczności

- Szczególną postacią testu BP jest test White'a ⟶ Z_i zawiera wszystkie kwadraty i iloczyny krzyżowe zmiennych objaśniających

- Stosujemy gdy interesuje nas samo wykrycie heteroskedastyczności a mniej wykrycie zmiennych, od których zależy wariancja błędu losowego

Testowanie heteroskedastyczności

- Test BP i White'a są bardziej uniwersalne niż test GQ jednak rozkłady statystyk testowych dla tych testów są znane tylko dla dużych prób

Jakie założenie KMRL nie jest spełnione przy odrzuceniu H0?

- Homoskedastyczność składnika losowego – wariancja błędu losowego jest stała dla wszystkich obserwacji:

$$\mathrm{var}(\varepsilon_i) = \sigma^2 \quad dla \quad i = 1, 2, \ldots, N$$

Testowanie heteroskedastyczności

- Przykład

xi: reg wydg dochg i.klm

```
      Source |       SS       df       MS              Number of obs =   31705
-------------+------------------------------           F(  6, 31698) = 3651.59
       Model |  2.3693e+10     6  3.9489e+09           Prob > F      =  0.0000
    Residual |  3.4278e+10 31698  1081405.34           R-squared     =  0.4087
-------------+------------------------------           Adj R-squared =  0.4086
       Total |  5.7971e+10 31704  1828523.21           Root MSE      =  1039.9

------------------------------------------------------------------------------
        wydg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dochg |   .5818533   .0040164   144.87   0.000      .573981    .5897256
     _Iklm_2 |  -40.65607   23.26644    -1.75   0.081     -86.2592    4.947067
     _Iklm_3 |  -70.57179   25.89099    -2.73   0.006    -121.3191   -19.82444
     _Iklm_4 |  -109.2499   20.60656    -5.30   0.000    -149.6395   -68.86021
     _Iklm_5 |  -153.3497   22.98153    -6.67   0.000    -198.3944    -108.305
     _Iklm_6 |  -173.5506   18.96167    -9.15   0.000    -210.7162    -136.385
       _cons |   836.1774   18.74554    44.61   0.000     799.4354    872.9194
```

Testowanie heteroskedastyczności

- Przykład

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of wydg
         chi2(1)      =129088.50
         Prob > chi2  =   0.0000


White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(12)     =   6142.84
         Prob > chi2  =    0.0000
```

Testowanie autokorelacji

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

-Brak autokorelacji

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

Testowanie autokorelacji

- Test Durbina-Watsona (Test DW):

$$H_0 : Cov(\varepsilon_t, \varepsilon_{t-1}) = 0$$ **- brak autokorelacji**

$$H_1 : Cov(\varepsilon_t, \varepsilon_{t-1}) \neq 0$$ **- autokorelacja**

gdzie $t = 1, \ldots, T$

Testowanie autokorelacji

- **Test Durbina-Watsona (Test DW):**

- specjalne tablice z wartościami krytycznymi: d_l, d_u

1. Statystyka DW<2

a) DW < d_l - odrzucamy hipotezę zerową o braku autokorelacji i przyjmujemy hipotezę o dodatniej autokorelacji

b) d_l < DW < d_u - brak konkluzji

c) DW > d_u - nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji

Testowanie autokorelacji

- Test Durbina-Watsona (Test DW):

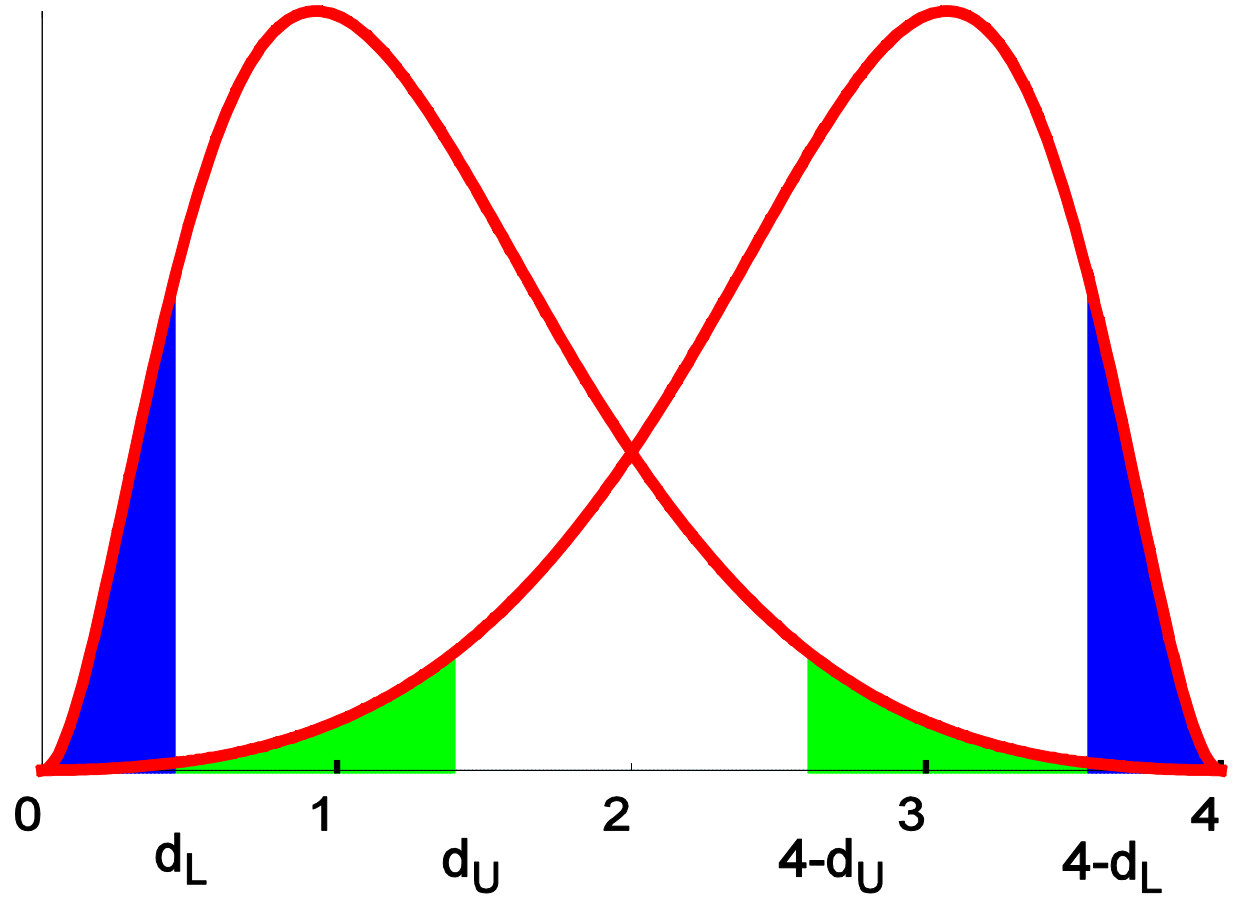
2. Statystyka DW >2

a) DW > $4 - d_l$ - odrzucamy hipotezę zerową o braku autokorelacji i przyjmujemy hipotezę o ujemnej autokorelacji

b) $4 - d_u$ < DW < $4 - d_l$ - brak konkluzji

c) DW < $4 - d_u$ - nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji

Testowanie autokorelacji



Testowanie autokorelacji

- Test Durbina-Watsona (Test BW):

- Do badania autokorelacji I rzędu (między $\varepsilon_t, \varepsilon_{t-1}$)

- Rozkład statystki testowej wyprowadzony dla małych prób

- Nie można go stosować w modelach gdzie jedną ze zmiennych objaśniających jest opóźniona zmienna zależna

- Wada: niestandardowy rozkład i możliwość wystąpienia braku konkluzji

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG):

- Do badania autokorelacji wyższego rzędu

- Można go stosować w modelach gdzie występują opóźnione zmienne zależne

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

Jakie założenie KMRL nie jest spełnione przy odrzuceniu H0?

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

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad dla \quad i \neq j$$

Plan wykładu

- 1. Testy diagnostyczne
  - Testowanie heteroskedastyczności
  - Testowanie autokorelacji
- 2.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne

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

- 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:

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

- 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

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

Dziękuję za uwagę