Problemy z danymi
Heteroskedastyczność i autokorelacja

Stanisław Cichocki

Natalia Nehrebecka

Wykład 14

Plan wykładu

- 1.Problemy z danymi
  - Współliniowość
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
  - Metody radzenia sobie z heteroskedastycznością i autokorelacją

Plan wykładu

- 1.Problemy z danymi
  - Współliniowość
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
  - Metody radzenia sobie z heteroskedastycznością i autokorelacją

Współliniowość

- O współliniowości mówimy w przypadku występowania silnej korelacji między zmiennymi objaśniającymi ⟶ utrudnia to zidentyfikowanie zmiennej, która jest przyczyną zmiennej zależnej

- Wyróżniamy dwa typy współliniowości:

  a) dokładną współliniowość

  b) niedokładną współliniowość

Współliniowość

- O dokładnej współliniowości mówimy, gdy kolumny macierzy obserwacji są współliniowe ⟶ jedna z kolumn macierzy jest kombinacją liniową pozostałych kolumn ⟶ macierz X'X jest osobliwa i wobec tego nieodwracalna

- Oznacza to, że jedna ze zmiennych niezależnych jest kombinacją liniową pozostałych zmiennych niezależnych i nie wnosi żadnej dodatkowej informacji do modelu ⟶ powinniśmy usunąć ja z modelu

- Dokładna współliniowość jest wynikiem błędnej specyfikacji modelu

Współliniowość

Przykład:

zmienne objaśniające w modelu:

a) ln(PKB),

b) ln(Liczba ludności)

c) ln(PKB per capita)

- Zmienna ln(PKB per capita) jest kombinacją zmiennej ln(PKB) i ln(Liczba ludności)

Współliniowość

- O niedokładnej współliniowości mówimy, gdy występuje silna korelacja między zmiennymi objaśniającymi

- W przypadku danych ekonometrycznych występowanie korelacji między zmiennymi objaśniającymi jest regułą ⟶ problemem jest nie samo występowanie korelacji lecz przypadek gdy jest ona bardzo silna ⟶ obniża to precyzję oszacowań

Współliniowość

- Statystyka służąca do wykrywania niedokładnej współliniowości nazywa się współczynnikiem inflacji wariancji:

$$VIF_k = \frac{1}{1 - R_k^2}$$

gdzie:

R_k^2 - R^2 w regresji x_k na pozostalych zmiennych objaśniających

Współliniowość

- Wysokie wartości VIF (>10) dla zmiennych objaśniających sygnalizują występowanie silnej niedokładnej współliniowości między zmiennymi
- Rozwiązaniem problemu silnej niedokładnej współliniowości jest usunięcie zmiennej o najwyższym VIF, co powinno poprawić precyzję oszacowań przy pozostałych zmiennych
- Należy jednak pamiętać, że jeśli usunięta zmienna była istotna w modelu to jej usunięcie może spowodować obciążenie estymatorów przy zmiennych, z którymi jest skorelowana
- Niedokładna współliniowość nie jest wynikiem błędnej specyfikacji modelu lecz wynika z własności konkretnego zbioru danych

Współliniowość

```
reg wydg dochg dochg2 dochg3

      Source |       SS       df       MS              Number of obs =   31679
-------------+------------------------------           F(  3, 31675) = 8591.16
       Model |  2.5996e+10     3  8.6653e+09           Prob > F      =  0.0000
    Residual |  3.1948e+10 31675  1008629.31           R-squared     =  0.4486
-------------+------------------------------           Adj R-squared =  0.4486
       Total |  5.7944e+10 31678  1829163.02           Root MSE      =  1004.3

------------------------------------------------------------------------------
        wydg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dochg |   .8616921   .0080228   107.41   0.000     .8459671     .877417
     dochg2 |  -.0000275   9.05e-07   -30.36   0.000    -.0000292   -.0000257
     dochg3 |   2.76e-10   1.58e-11    17.46   0.000     2.45e-10    3.07e-10
      _cons |   316.3596   13.51433    23.41   0.000      289.871    342.8482
```

Współliniowość

```
  vif

    Variable |       VIF       1/VIF
-------------+----------------------
      dochg2 |     21.44    0.046644
      dochg3 |     13.40    0.074609
       dochg |      4.35    0.229769
-------------+----------------------
    Mean VIF |     13.06
```

Plan wykładu

- 1.Problemy z danymi
  - Współliniowość
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
  - Metody radzenia sobie z heteroskedastycznością i autokorelacją

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

Sferyczność błędów losowych

- Jeżeli założenie o homoskedastyczności i 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

Metody radzenia sobie z heteroskedastycznością i autokorelacją

- Stosowalna Uogólniona Metoda Najmniejszych Kwadratów (SUMNK)
- Odporne estymatory macierzy wariancji i kowariancji: estymator White'a (heteroskedastyczność), estymator Newey'a-Westa (heteroskedastyczność i autokorelacja)

Dziękuję za uwagę