

Problemy z danymi

Testy diagnostyczne

Stanisław Cichocki

Natalia Nehrebecka

Wykład 12

Plan wykładu

- ▶ 1. Problemy z danymi
 - Zmienne pominięte
 - Zmienne nieistotne
- ▶ 2. Autokorelacja
 - Testowanie autokorelacji

Plan wykładu

- ▶ 1. Problemy z danymi
 - Zmienne pominięte
 - Zmienne nieistotne
- ▶ 2. Autokorelacja
 - Testowanie autokorelacji

Zmienne pominięte

- Mamy 2 modele:

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

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

- Potencjalnie każdy z tych modeli może prawidłowo opisywać zmienną $y \longrightarrow$ problemy gdy przy liczeniu estymatorów zastosujemy niewłaściwy model
- Załóżmy, że 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** (omitted variables)

Zmienne pominięte

- $\hat{\beta}_1$ - estymator MNK wektora parametrów w modelu (1)
- Załóżmy, że prawdziwy jest model (2)

$$\begin{aligned}\hat{\beta}_1 &= (X_1' X_1)^{-1} X_1' y = (X_1' X_1)^{-1} X_1' (X_1 \beta_1 + X_2 \beta_2 + \varepsilon) \\ &= \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2 + (X_1' X_1)^{-1} X_1' \varepsilon\end{aligned}$$

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 między 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 zmienną zależną **————→** najgorszy przypadek

b) Przeszacowania/niedoszacowania wpływu zmiennej objaśniającej na zmienną objaśnianą

Zmienne pominięte

► Przykład

reg wydg dochg

Source	SS	df	MS
Model	2.3577e+10	1	2.3577e+10
Residual	3.4367e+10	31677	1084914.37
Total	5.7944e+10	31678	1829163.02

Number of obs = 31679
F(1, 31677) =21732.03
Prob > F = 0.0000
R-squared = 0.4069
Adj R-squared = 0.4069
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		
Model	2.3886e+10	2	1.1943e+10	F(2, 31676) =11107.42		
Residual	3.4059e+10	31676	1075214.71	Prob > F = 0.0000		
Total	5.7944e+10	31678	1829163.02	R-squared = 0.4122		
				Adj R-squared = 0.4122		
				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:

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

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

- Załóżmy, że 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. Autokorelacja
 - Testowanie autokorelacji

Autokorelacja

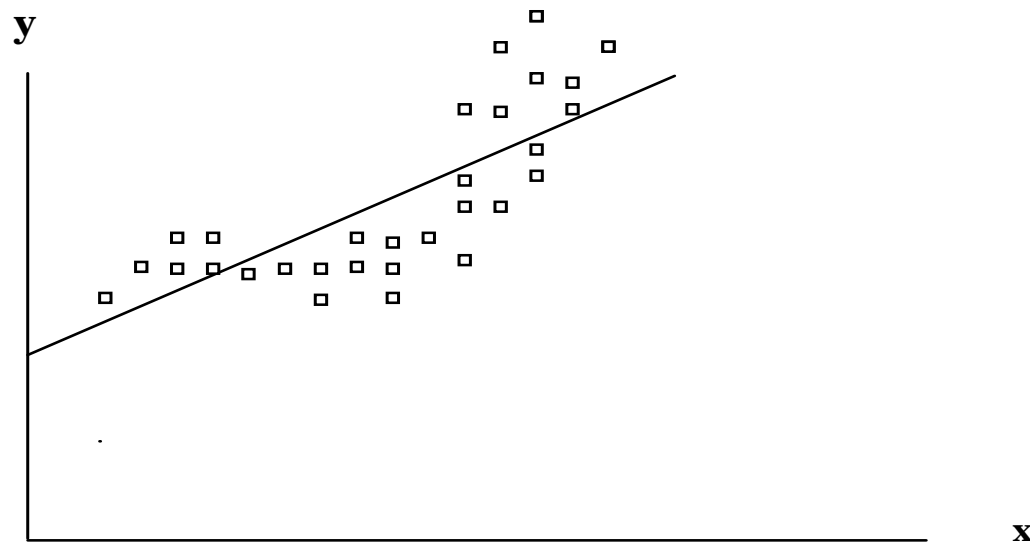
Przypomnienie: Co to znaczy, że w modelu występuje autokorelacja?

-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

Testowanie autokorelacji

- Test Durbina-Watsona (Test DW):

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

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

gdzie $t = 1, \dots, T$

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

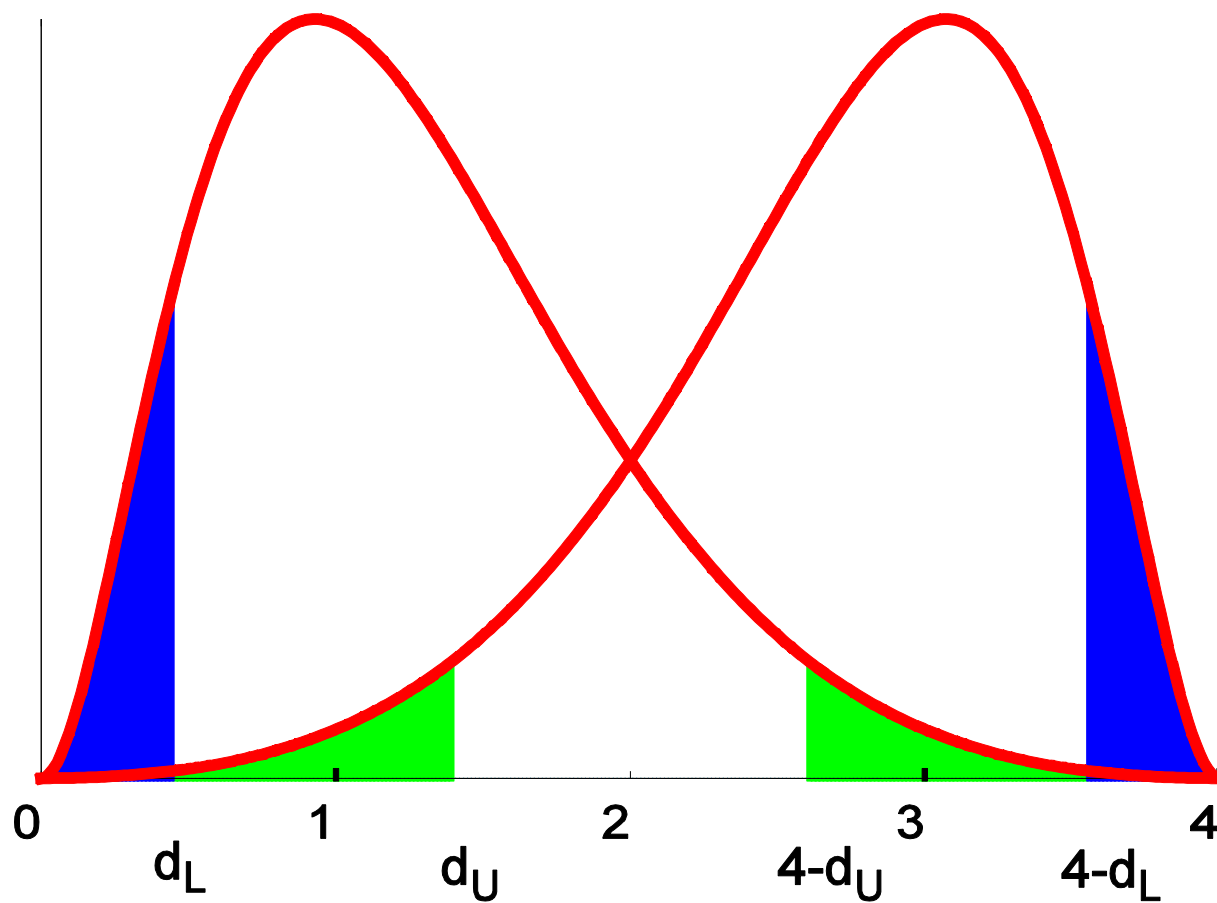
Testowanie autokorelacji

- 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 statystyki testowej wyprowadzony dla małych prób
 - Wada: 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-Godfrey (Test BG):
 - Do badania autokorelacji wyższego rzędu
 - Można go stosować w modelach gdzie występują opóźnione zmienne zależne

Testowanie autokorelacji

- Test Breuscha-Godfrey (Test BG):

$$H_0 : Cov(\varepsilon_t, \varepsilon_{t-i}) = 0 \quad \text{gdzie} \quad i = 1, \dots, s$$

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

- Hipoteza zerowa: brak autokorelacji
- Hipoteza alternatywna: autokorelacja

Testowanie autokorelacji

- ▶ Test Breuscha-Godfrey (Test BG) – sposób przeprowadzenia testu:

1. przeprowadzamy regresję y_i na x_i i uzyskujemy reszty

2. przeprowadzamy regresję pomocniczą:

$$e_t = x_t \mu + \gamma_1 e_{t-1} + \dots + \gamma_s e_{t-s} + u_t$$

i testujemy $H_0: \gamma_1 = \dots = \gamma_s = 0$

Testowanie autokorelacji

- ▶ Statystyka testowa:

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

lub statystyka F

Jakie założenie KMRL nie jest spełnione przy odrzuceniu H_0 ?

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

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

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 jakich testów testuje się autokorelację? Jakiemu założeniu KMRL odpowiada H_0 w tych testach? Jakie są hipotezy alternatywne w tych testach?

Dziękuję za uwagę