



Testowanie hipotez Założenia KMRL Tw. Gaussa-Markowa

Stanisław Cichocki
Natalia Nehrebecka

Zajęcia 7-8

Plan zajęć

- ▶ 1. Testowanie łącznej istotności wybranych regresorów
- ▶ 2. Założenia klasycznego modelu regresji liniowej
- ▶ 3. Własności estymatora MNK w KMRL
 - Wartość oczekiwana i nieobciążoność estymatora b
 - Wariancja i efektywność estymatora b : twierdzenie Gaussa-Markowa
- ▶ 4. Estymator wariancji błędu losowego

1. Testowanie łącznej istotności wybranych regresorów



Testowanie łącznej istotności wybranych regresorów

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

- ▶ **Krok 1.** Stawiamy hipotezę zerową H_0 :

$$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$$

- ▶ Brak podstaw do odrzucenia tej hipotezy oznacza, że zmienne x_{2i}, x_{3i}
- ▶ są łącznie nieistotne (powinny zostać usunięte z modelu).

- ▶ **Model bez ograniczeń** $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$
- ▶ **Model z ograniczeniami** $y_i = \beta_1 + \beta_4 x_{4i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$

Testowanie łącznej istotności wybranych regresorów

- ▶ **Krok 2.** Przy założeniu, że postawiona hipoteza zerowa jest prawdziwa, wyznaczamy statystykę testową z rozkładu F:

$$F = \frac{(\mathbf{e}'_R \mathbf{e}_R - \mathbf{e}' \mathbf{e}) / g}{\mathbf{e}' \mathbf{e} / (n - K)} = \frac{(R^2 - R_R^2) / g}{(1 - R^2) / (n - K)}$$

- ▶ Gdzie:
- ▶ $R^2, \mathbf{e}' \mathbf{e}$ oznaczają współczynnik determinacji i sumę kwadratów reszt dla modelu bez ograniczeń
- ▶ $R_R^2, \mathbf{e}'_R \mathbf{e}_R$ to te same wielkości, ale dla modelu z ograniczeniami,
- ▶ g oznacza liczbę ograniczeń,
- ▶ K – ilość szacowanych parametrów w modelu bez ograniczeń,
- ▶ n – liczba obserwacji

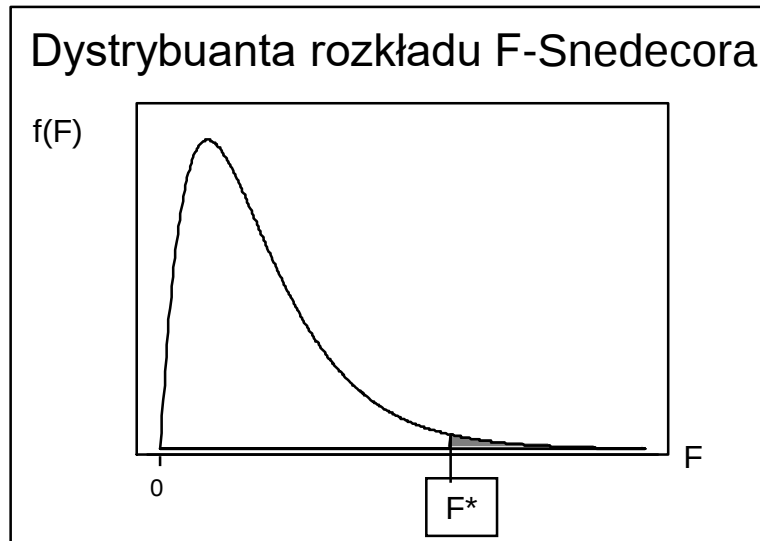
Testowanie łącznej istotności wybranych regresorów

- ▶ **Krok 3.** Odczytujemy z tablic rozkładu F wartość krytyczną (α - poziom istotności)

$$F^* = F(g, n - K)$$

Testowanie łącznej istotności wybranych regresorów

- ▶ **Krok 4.** Podjęcie decyzji



- ▶ $F \geq F^*$ - odrzucamy hipotezę zerową (zmienne są łącznie istotne).
- ▶ $F < F^*$ - nie ma podstaw do odrzucenia hipotezy zerowej (zmienne są łącznie nieistotne).

2. Założenia klasycznego modelu regresji liniowej



Założenia klasycznego modelu regresji liniowej

- ▶ 1. Związek pomiędzy zmienną zależną a zmiennymi niezależnymi opisany jest równaniem:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_K x_{Ki} + \varepsilon_i \quad i = 1, 2, 3 \dots n$$

- ▶ 2. Zmienne objaśniające $x_{2i}, x_{3i}, \dots, x_{Ki}$ są niełosowe dla $i = 1, 2, 3 \dots n$
- ▶ 3. Wartość oczekiwana błędu losowego jest równa zero:

$$E(\varepsilon) = \mathbf{0}$$

- ▶ 4. Zaburzenia losowe ε są **sferyczne**. Oznacza to, że warunkowa macierz wariancji-kowariancji wektora zaburzeń przy danej macierzy X ma postać:

$$Var(\varepsilon) = \sigma^2 \mathbf{I}$$

- ▶ gdzie $\mathbf{I}$ oznacza macierz jednostkową.

Założenia klasycznego modelu regresji liniowej

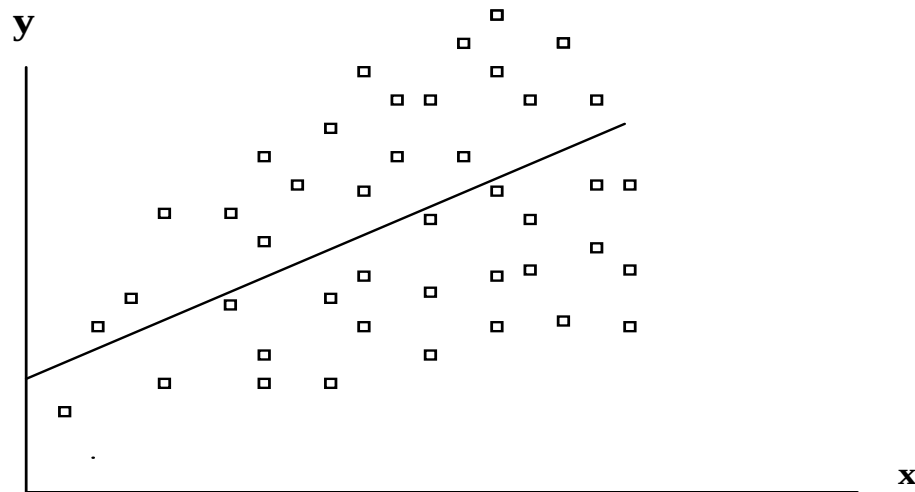
$$\text{Var}(\varepsilon) = \sigma^2 \mathbf{I}$$

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

- ▶ Założenie **sferyczności zaburzeń** oznacza:
- ▶ po pierwsze, że **wariacje** kolejnych zaburzeń (elementy na diagonalnej) są takie same dla wszystkich obserwacji i równe σ^2 , gdzie σ^2 jest nieznaną dodatnią stałą;
- ▶ po drugie, że elementy pozadiagonalne, które są **kowariancjami** zaburzeń dla różnych obserwacji są równe zero, a więc zaburzenia dla różnych obserwacji są ze sobą nieskorelowane.

Założenia klasycznego modelu regresji liniowej

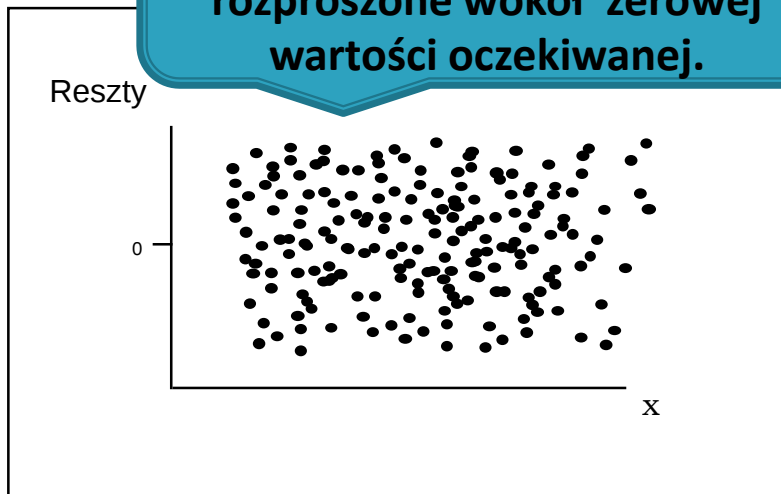
- ▶ Stałość wariancji zaburzeń nazywamy **homoskedastycznością zaburzeń**. Oznacza to, że zaburzenia losowe są **jednakowo rozproszone wokół zerowej wartości oczekiwanej**. Jeśli wariancje nie byłyby jednakowe, to sytuację taką nazywamy **heteroskedastycznością**.



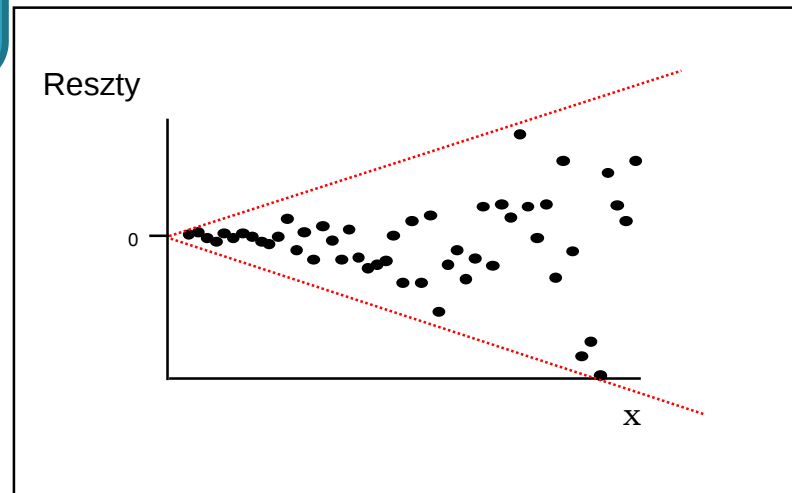
Rys.1. Heteroskedastyczność

Założenia klasycznego modelu regresji liniowej

Oznacza to, że zaburzenia losowe są jednakowo rozproszone wokół zerowej wartości oczekiwanej.



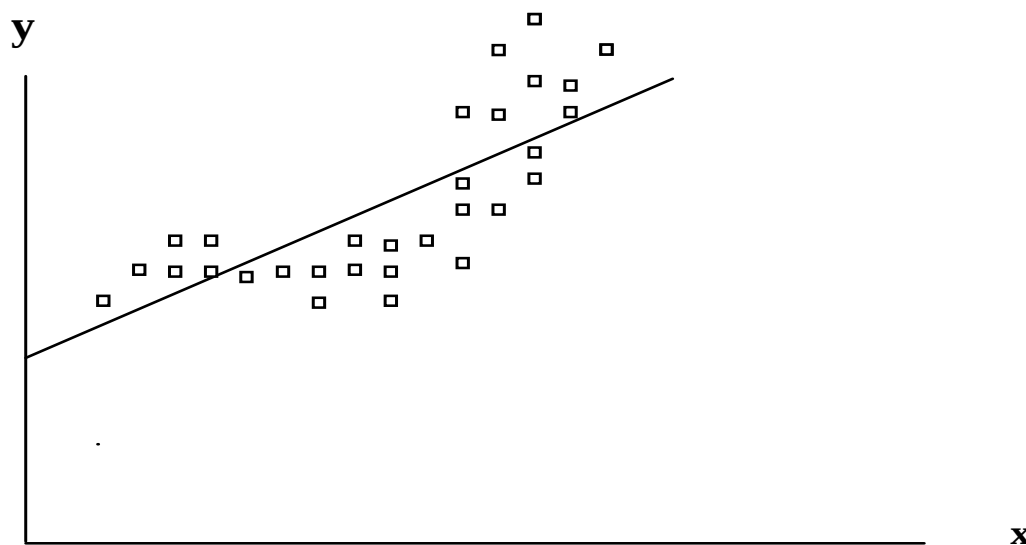
Homoskedastyczność: reszty zachowują się losowo.



Heteroskedastyczność: Wariancja reszt zmienia się wraz ze zmianą zmiennej niezależnej X .

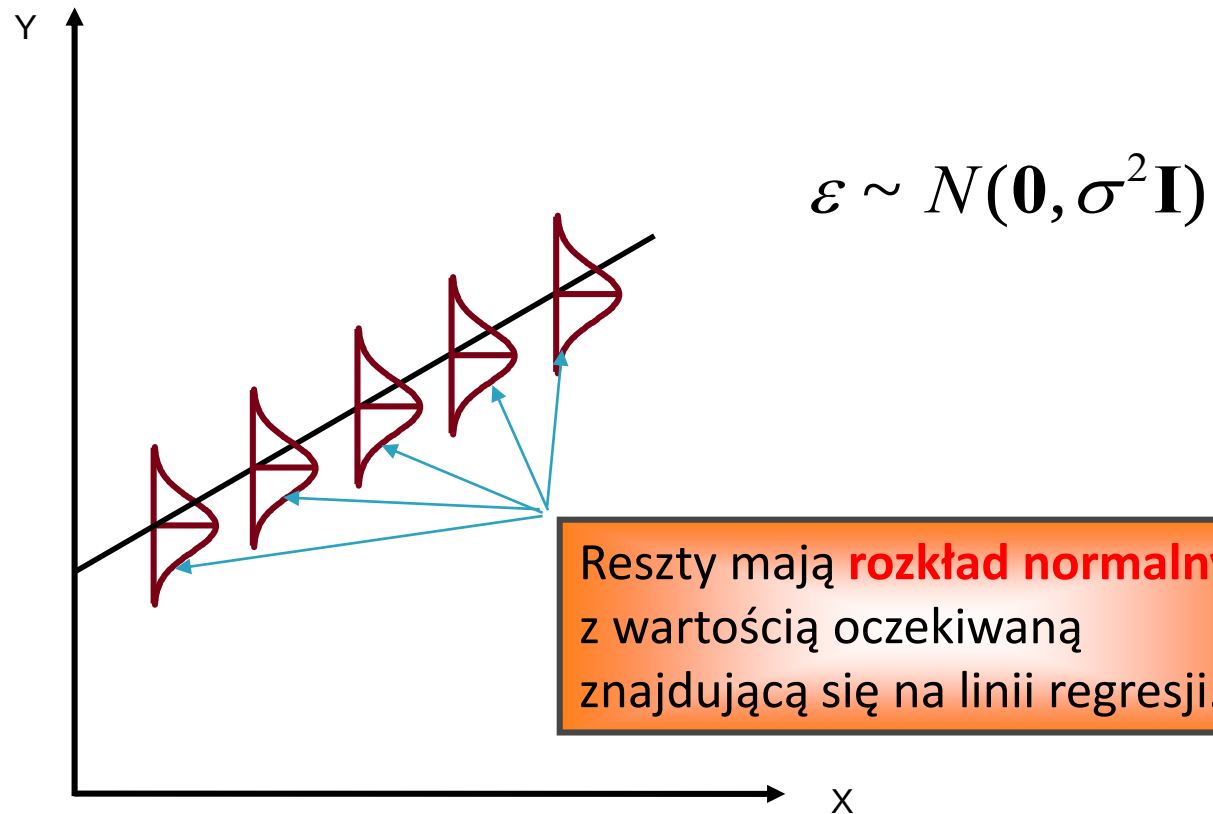
Założenia klasycznego modelu regresji liniowej

- ▶ 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

Dodatkowo założenie klasycznego modelu regresji liniowej



3. Twierdzenie Gaussa–Markowa



Twierdzenie Gaussa-Markowa

W klasycznym modelu regresji liniowej **najlepszym liniowym** i **nieobciążonym** estymatorem wektora parametrów β jest b wyznaczone za pomocą MNK

$$b = (X'X)^{-1} X'y$$

o macierzy wariancji-kowariancji

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

$$Var(b) = \begin{bmatrix} Var(b_1) & Cov(b_2, b_1) & \cdots & Cov(b_n, b_1) \\ Cov(b_1, b_2) & Var(b_2) & \cdots & Cov(b_n, b_2) \\ \vdots & \vdots & & \vdots \\ Cov(b_1, b_n) & Cov(b_2, b_n) & \cdots & Var(b_n) \end{bmatrix}$$

Dyskusja twierdzenia Gaussa-Markowa

- ▶ 1. Estymator $\mathbf{b}$ jest estymatorem **liniowym**, gdyż jest liniową funkcją zmiennej losowej $\mathbf{y}$.

- ▶ 2. $\mathbf{b}$ jest estymatorem **nieobciążonym**, to znaczy $E(\mathbf{b}) = \beta$.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

- ▶ i podstawiając za $\mathbf{y}$ $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

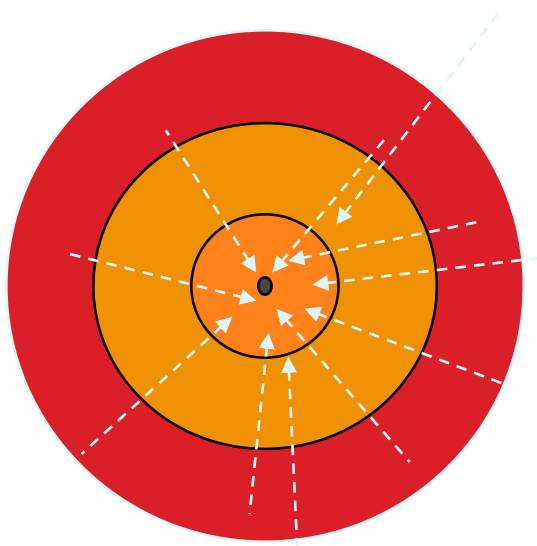
- ▶ otrzymamy:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\beta + \varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon$$

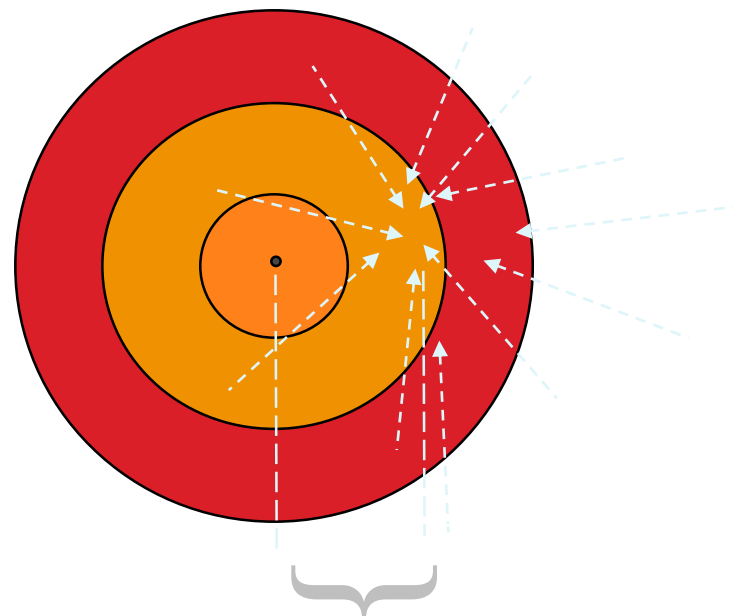
$$E(\mathbf{b}) = \beta + E((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'E(\varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'0 = \beta$$

- ▶ 3. Estymator $\mathbf{b}$ jest estymatorem **najlepszym** w tym sensie, że każdy inny estymator liniowy i nieobciążony ma macierz wariancji-kowariancji większą od tej dla $\mathbf{b}$. Estymator taki nazywamy estymatorem **efektywnym**.

Obciążony i nieobciążony estymator



**Estymator
nieobciążony**

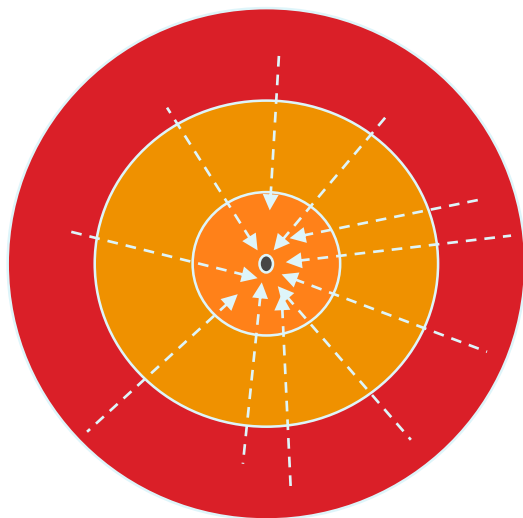


Obciążenie

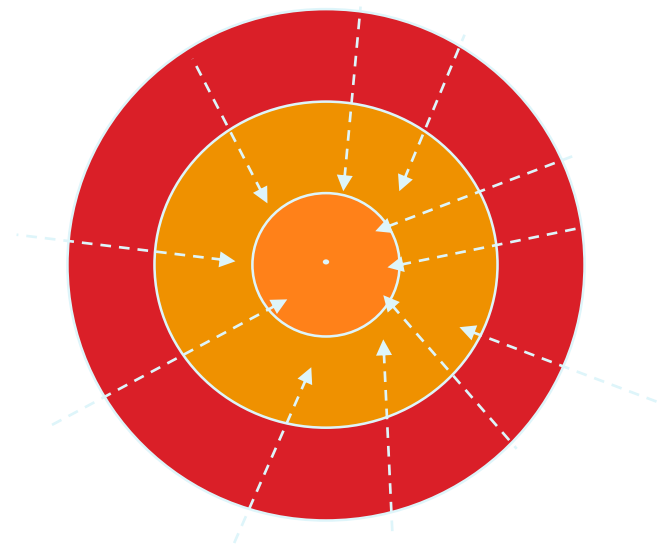
**Estymator
obciążony**

Efektywność

Estymator jest efektywnym, jeśli ma najniższą wariancję i odchylenie standardowe.



**Estymator
efektywny**



**Estymator
nieefektywny**

4. Estymator wariancji błędu losowego



Estymator wariancji błędu losowego

- ▶ Ponieważ zaburzenia losowe ε są nieznane, to i nieznana jest ich wariancja.
- ▶ Wektor reszt $\mathbf{e}$ w wyznaczony z próby traktujemy jako wektor estymatorów zaburzeń losowych.
- ▶ Istnieje więc przesłanka dla wyznaczenia estymatora wariancji $\hat{\sigma}^2$ na podstawie wektora reszt $\mathbf{e}$.
- ▶ W teorii ekonometrii dowodzi się twierdzenie, że

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - K} = \frac{\mathbf{e}'\mathbf{e}}{n - K}$$

Przykład

gretl: model 1

Plik Edycja Testy Zapisz Wykresy Analiza

Model 1: Estymacja KMNK z wykorzystaniem 1118 obserwacji 1-1118
Zmienna zależna: dochod

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	300,633	114,768	2,619	0,00893 ***
wiek	6,12471	2,10372	2,911	0,00367 ***
pl	294,443	45,1271	6,525	<0,00001 ***
sre	213,453	69,3871	3,076	0,00215 ***
wyzsze	1047,81	84,6752	12,374	<0,00001 ***
duze_m	491,413	54,7537	8,975	<0,00001 ***
srednie_m	-49,5269	70,1756	-0,706	0,48049
male_m	98,0211	69,2835	1,415	0,15741

Srednia arytmetyczna zmiennej zależnej = 1224,23
Odchylenie standardowe zmiennej zależnej = 864,584
~~Suma kwadratów reszt = 6,87111E+008~~
Błąd standardowy reszt = 739,558
Wsp. determinacji R-kwadrat = 0,278889
Skorygowany wsp. R-kwadrat = 0,268303
Statystyka F (7, 1110) = 59,5127 (wartość p < 0,00001)
Logarytm wiarygodności = -8967,93
Kryterium informacyjne Akaika (AIC) = 17951,9
Kryterium bayesowskie Schwarz (BIC) = 17992
Kryterium infor.Hannana-Quinna (HQC) = 17967

Wyłączając stałą, największa wartość p jest dla zmiennej 9 (srednie_m)

Zamknij

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{N - K}}$$

Standardowy błąd estymatora b_i

$$se(b_K) = \sqrt{Var(b_K)}$$

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

$$Var(b) = \begin{bmatrix} Var(b_1) & Cov(b_2, b_1) & \cdots & Cov(b_n, b_1) \\ Cov(b_1, b_2) & Var(b_2) & \cdots & Cov(b_n, b_2) \\ \vdots & \vdots & & \vdots \\ Cov(b_1, b_n) & Cov(b_2, b_n) & \cdots & Var(b_n) \end{bmatrix}$$

- ▶ Standardowy błąd szacunku parametru β_i mówi o ile jednostek wartość b_i różni się od nieznaney wielkości parametru β_i .

Pytania teoretyczne

1. Jak testujemy hipotezę o nieistotności określonego podzbioru regresorów?
2. Proszę wymienić założenia klasycznego modelu regresji liniowej.
3. Proszę sformułować twierdzenie Gaussa-Markowa i je zinterpretować.
4. Co to jest błąd standardowy estymatora? Proszę podać wzór dla przypadku regresji wielu zmiennych i go zinterpretować.

Dziękuję za uwagę