

Interpretacja parametrów : zmienne dyskretne

Testowanie hipotez

Stanisław Cichocki

Natalia Nehrebecka

Zajęcia 5-6

Plan zajęć

- ▶ 1. Interpretacja parametrów przy zmiennych zerojedynekowych
 - Zmienne 0-1
 - Interpretacja przy zmiennej 0 – 1 w modelu liniowym względem zmiennych objaśniających
 - Interpretacja przy zmiennej 0 – 1 w modelu, w którym zmienna zależna jest zlogarytmowana
- ▶ 2. Testowanie istotności poszczególnych zmiennych objaśniających w modelu
- ▶ 3. Przedziały ufności
- ▶ 4. Testowanie łącznej istotności zmiennych objaśniających

1. Zmienne 0-1



Zmienne 0 - 1

- ▶ Zmienną zero-jedynkową nazywamy zmienną, która przyjmuje tylko dwie wartości: 0 lub 1
- ▶ płeć: 1 – kobieta, 0 – mężczyzna
- ▶ praca: 1 – pracujący, 0 – niepracujący
- ▶ obecność dzieci: 1 – nie, 0 – tak
- ▶ Uwaga!
- ▶ Ważne jest, że zmienna przyjmuje dwie wartości, nie ma znaczenia ich wielkość.

Interpretacja przy zmiennej 0 – 1 w modelu liniowym względem zmiennych objaśniających

$$D_i = \begin{cases} 1 & \text{poziom badany} \\ 0 & \text{poziom bazowy} \end{cases}$$

- ▶ Niech D_i będzie zmienną zero-jedynkową:

$$y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \gamma D_i + \varepsilon_i$$

- ▶ Dla $D_i = 1$ model ma postać:

$$y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \gamma + \varepsilon_i$$

- ▶ Dla $D_j = 0$ model ma postać:

$$y_j = \beta_1 + \beta_2 X_{2j} + \dots + \beta_K X_{Kj} + \varepsilon_j$$

- ▶ Zatem $\gamma = E(y_i) - E(y_j)$

Interpretacja przy zmiennej 0 – 1 w modelu liniowym względem zmiennych objaśniających

- ▶ Wniosek:
- ▶ Wielkość $\mathcal{V}$ można interpretować jako zmianę oczekiwanej wartości y , jeśli D zmieni się z 0 na 1, przy założeniu pozostałych charakterystyk na niezmiennym poziomie.

Przykład

$$y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \gamma D_i + \varepsilon_i$$

$$\hat{y}_i = b_1 + b_2 X_{2i} + \dots + b_K X_{Ki} + \hat{\gamma} D_i$$

γ – współczynnik przy zmiennej 0-1

INTERPRETACJA: wartość zmiennej zależnej y dla poziomu zmiennej 0-1 $D=1$ jest:

- większa (jeżeli $\hat{\gamma} > 0$) o $|\hat{\gamma}|$ jednostek lub
- mniejsza (jeżeli $\hat{\gamma} < 0$) o $|\hat{\gamma}|$ jednostek

niż wartość zmiennej zależnej y dla poziomu zmiennej 0-1 $D=0$ (dla poziomu bazowego)

Przykład

$$placa_i = \beta_1 + \beta_2 plec_i + \varepsilon_i$$

$$\hat{placa}_i = 9,26 - 503,59 \cdot plec_i$$

► Zmienna $plec_i = \begin{cases} 1 & \text{jesli kobieta} \\ 0 & \text{jesli mezczyzna} \end{cases}$

► Interpretacja:

Oczekiwany poziom płac kobiet jest średnio o 503,59 złotego niższy niż dla mężczyzn, przy założeniu pozostałych charakterystyk na niezmiennym poziomie.

Interpretacja przy zmiennej 0 – 1 w modelu, którym zmienna zależna jest zlogarytmowana

$$\ln y_i = \beta_1 + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \gamma D_i + \varepsilon_i$$

► Wniosek:

Wielkość γ (przemnożoną przez 100%)
można interpretować jako procentową zmianę oczekiwanej
wartości zmiennej zależnej y , jeśli D zmieni się z 0 na 1 .

Przykład

$$\ln(\text{placa}_i) = \beta_1 + \beta_2 \text{plec}_i + \varepsilon_i$$

$$\ln(\widehat{\text{placa}}_i) = 7,67 - 0,17 \cdot \text{plec}_i$$

- ▶ Zmienna $\text{plec}_i = \begin{cases} 1 & \text{jesli kobieta} \\ 0 & \text{jesli mezczyzna} \end{cases}$
- ▶ Interpretacja:
Oczekiwany poziom płac kobiet jest średnio o 17% niższy niż dla mężczyzn, przy założeniu pozostałych charakterystyk na niezmiennym poziomie.

Zadanie

- ▶ Zbadajmy związek między płacą a stażem pracy, wykształceniem (mierzonym liczbą lat nauki) i płcią respondenta (0 – kobieta, 1 – mężczyzna).

Zmienna zależna: placa

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	-465,485	149,036	-3,123	0,00184 ***
plec	386,015	48,4706	7,964	<0,00001 ***
staz	5,40578	2,38159	2,270	0,02341 **
nauka	79,8292	9,50751	8,396	<0,00001 ***

- ▶ Proszę zinterpretować współczynniki w modelu.

Zmienne dyskretne

- ▶ Nieco bardziej skomplikowana jest sytuacja, gdy mamy do czynienia ze zmienną dyskretną która przyjmuje więcej niż 2 wartości.
- ▶ np. **wykształcenie** (1 – podstawowe, 2 – średnie, 3 - wyższe)
- ▶ W tym przypadku do każdego poziomu **s** zmiennej dyskretnej **X_i** musimy przypisać jedną zmienną zero-jedynkową **$D_{s,i}$**

$$D_{s,i} = 1 \text{ gdy } X_i = s$$

$$D_{s,i} = 0 \text{ gdy } X_i \neq s \text{ dla } s = 1, 2, \dots, S$$

Przykład

$$\text{wyksztalce nie}_i = \begin{cases} 1 & \text{podstawowe} \\ 2 & \text{średnie} \\ 3 & \text{wyzsze} \end{cases}$$

$$\text{podstawowe}_i = \begin{cases} 1 & \text{podstawowe} \\ 0 & \text{w p. p.} \end{cases}$$

$$\text{średnie}_i = \begin{cases} 1 & \text{średnie} \\ 0 & \text{w p. p.} \end{cases}$$

$$\text{wyzsze}_i = \begin{cases} 1 & \text{wyzsze} \\ 0 & \text{w p. p.} \end{cases}$$

Zmienne dyskretne

- ▶ Za **poziom bazowy** uznajemy jeden z poziomów (np. poziom 1), i zmienną zero-jedynkową związaną z tym poziomem usuwamy z modelu **ze stałą**.
- ▶ Np. dla zmiennej wykształcenie
- ▶ Poziom bazowy : wykształcenie podstawowe

$$placa_i = \beta_1 + \beta_2 \acute{s}rednie_i + \beta_3 wyzsze_i + \varepsilon_i$$

- ▶ Dlaczego?
Nie jest możliwe, by w modelu była jednocześnie stała i wszystkie zmienne zero-jedynkowe (dla każdego poziomu zmiennej dyskretnej), ponieważ macierz $X^T X$ byłaby osobliwa!

Interpretacja parametrów przy zmiennych dyskretnych

- ▶ Interpretacja współczynników w modelu z wieloma zmiennymi 0-1 (zmiennymi dyskretnymi) jest analogiczna jak w przypadku modelu z jedną tylko taką zmienną:

dany współczynnik opisuje różnicę między oczekiwaną wartością zmiennej y dla respondenta o charakterystyce bazowej i dla respondenta o charakterystyce s .

Zadanie

- ▶ Szacujemy model, w którym zmienną objaśnianą jest wynagrodzenie, a zmiennymi objaśniającymi są:
 - ▶ płeć pracownika (0 – kobieta, 1 - mężczyzna),
 - ▶ wykształcenie pracownika (1 - podstawowe, 2 - gimnazjum, 3 - średnie, 4 - wyższe),
 - ▶ wiek pracownika,
 - ▶ stan cywilny (1- kawaler/panna, 2-wdowiec/wdowa, 3-rozwiedzony/rozwiedziona, 4-zonaty/ zamężna).
- ▶ Podać prawidłową formę modelu ze stałą i zinterpretować jego współczynniki.

2. Testowanie istotności poszczególnych zmiennych



Testowanie istotności poszczególnych zmiennych

- ▶ Testowanie prostych hipotez przebiega w następujących krokach:

- ▶ Dla modelu:

$$y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \varepsilon_i$$

- ▶ którego oszacowaniem jest:

$$\hat{y}_i = b_1 + b_2 X_{2i} + \dots + b_K X_{Ki}$$

- ▶ **Krok 1.** Stawiamy tak zwaną hipotezę zerową co do wartości nieznanego parametru β_K

$$H_0 : \beta_K = 0 \quad (\text{zmienna } X_{Ki} \text{ jest nieistotna})$$

- ▶ Hipotezie tej towarzyszy hipoteza alternatywna:

$$H_1 : \beta_K \neq 0 \quad (\text{zmienna } X_{Ki} \text{ jest istotna})$$

Testowanie istotności poszczególnych zmiennych

Krok 2. Przy założeniu, że postawiona hipoteza zerowa jest prawdziwa, wyznaczamy statystykę testową z rozkładu *t - Studenta* o **N - K** stopniach swobody postaci:

$$t = \frac{b_K}{se(b_K)}$$

Gdzie:

$se(b_K)$ - odchylenie standardowe estymatora b_K

Przykład

Zmienna zależna: placa

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	-465,485	149,036	-3,123	0,00184 ***
plec	386,015	48,4706	7,964	<0,00001 ***
staz	5,40578	2,38159	2,270	0,02341 **
nauka	79,8292	9,50751	8,396	<0,00001 ***

b

$se(b)$

$$t = \frac{b_K}{se(b_K)}$$

Testowanie istotności poszczególnych zmiennych

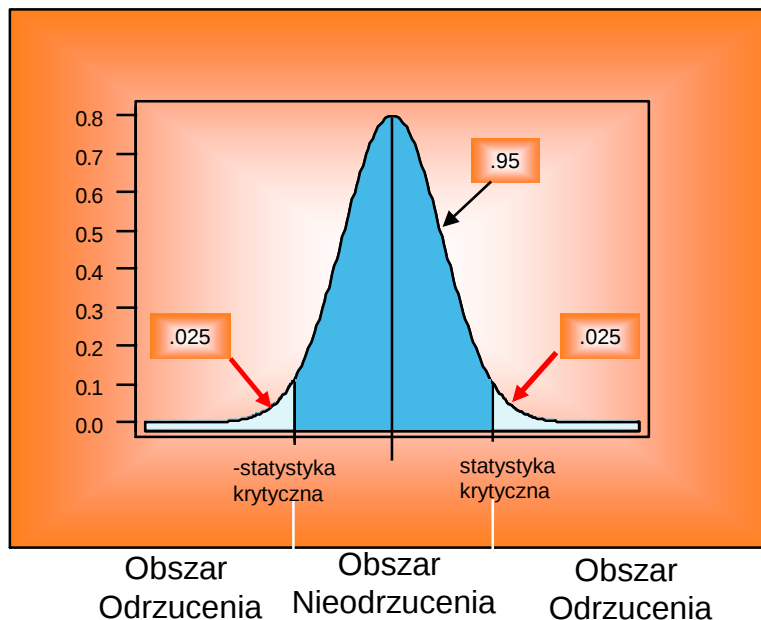
Krok 3. Odczytujemy z tablic rozkładu *t-Studenta* wartość krytyczną (α - poziom istotności¹⁾)

$$t^* = t \left(\underbrace{N - K}_{\text{Stopni swobody}}; \underbrace{1 - \frac{\alpha}{2}}_{\text{Rząd kwantyla}} \right)$$

¹⁾ maksymalne dopuszczalne prawdopodobieństwo popełnienia błędu polegającego na odrzuceniu prawdziwej hipotezy zerowej

Testowanie istotności poszczególnych zmiennych

Krok 4. Podjęcie decyzji



- ✓ $|t| \geq t^*$ - odrzucamy hipotezę zerową, czyli zmienna X_{ki} jest istotna.
- ✓ $|t| < t^*$ - nie ma podstaw do odrzucenia hipotezy zerowej, czyli zmienna X_{ki} jest nieistotna.

Wartość p (p-value)

- ▶ W popularnych pakietach ekonometrycznych obok wyliczonej wartości statystyki t podawane jest również odpowiadające mu **prawdopodobieństwo p**, że $\beta_k = 0$. Oznaczone ono jest z angielskiego przez *p-value*.
 - ▶ Jest to empirycznie wyliczony poziom istotności dla statystyki testowej t przy hipotezie zerowej $\beta_k = 0$
- ✓ Jeśli $p\text{-value} \leq \alpha$ (poziomu istotności), to odrzucamy hipotezę zerową.
 - ✓ Jeśli $p\text{-value} > \alpha$ (poziomu istotności), to brak podstaw do odrzucenia hipotezy zerowej.

Przykład

Zmienna zależna: placa

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	-465,485	149,036	-3,123	0,00184 ***
plec	386,015	48,4706	7,964	<0,00001 ***
staz	5,40578	2,38159	2,270	0,02341 **
nauka	79,8292	9,50751	8,396	<0,00001 ***

p – value

3. Przedziały ufności dla parametrów



Przedziały ufności dla parametrów

- ▶ Jaki jest przedział, w którym z określonym prawdopodobieństwem znajdzie się nieznana wartość parametru β_K . Odpowiedź na to pytanie uzyskamy wyznaczając tak zwany przedział ufności.
- ▶ Przedział ufności dla nieznanego parametru β_K na poziomie ufności $1 - \alpha$:

$$P(b_K - t_{1-\alpha/2} se(b_K) \leq \beta_K \leq b_K + t_{1-\alpha/2} se(b_K)) = 1 - \alpha$$

4. Test na łączną istotność zmiennych objaśniających



Test na łączną istotność zmiennych objaśniających

- ▶ Często stawianym pytaniem jest kwestia, czy równanie regresji jest statystycznie istotne? Jest ono równoważne pytaniu, czy łącznie współczynniki regresji, z wyjątkiem stałej, są równe zero.
- ▶ **Krok 1.** Stawiamy hipotezę zerową H_0 :

$$H_0 : \begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- ▶ H_1 : nie wszystkie współczynniki β_K są jednocześnie równe zero
(przynajmniej jedna zmienna jest istotna)

Test na łączną istotność zmiennych objaśniających

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

$$F = \frac{ESS / (K - 1)}{RSS / (n - K)} = \frac{R^2 / (K - 1)}{(1 - R^2) / (n - K)}$$

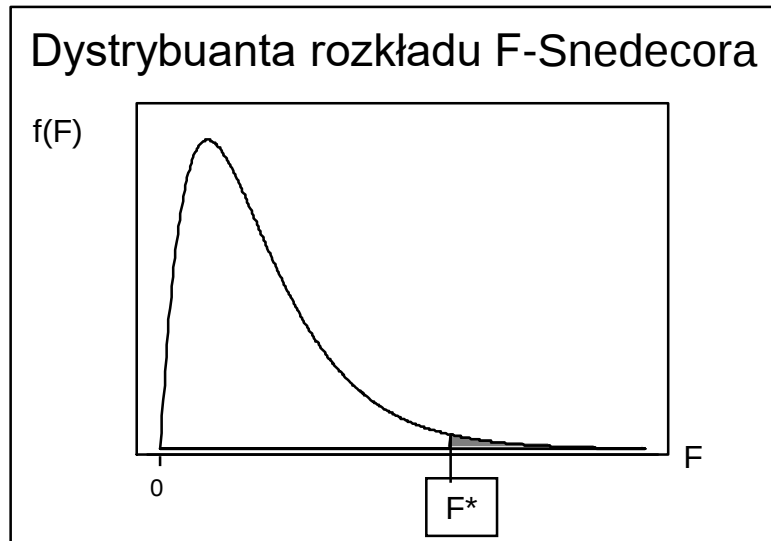
Test na łączną istotność zmiennych objaśniających

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

$$F^* = F(K - 1, n - K)$$

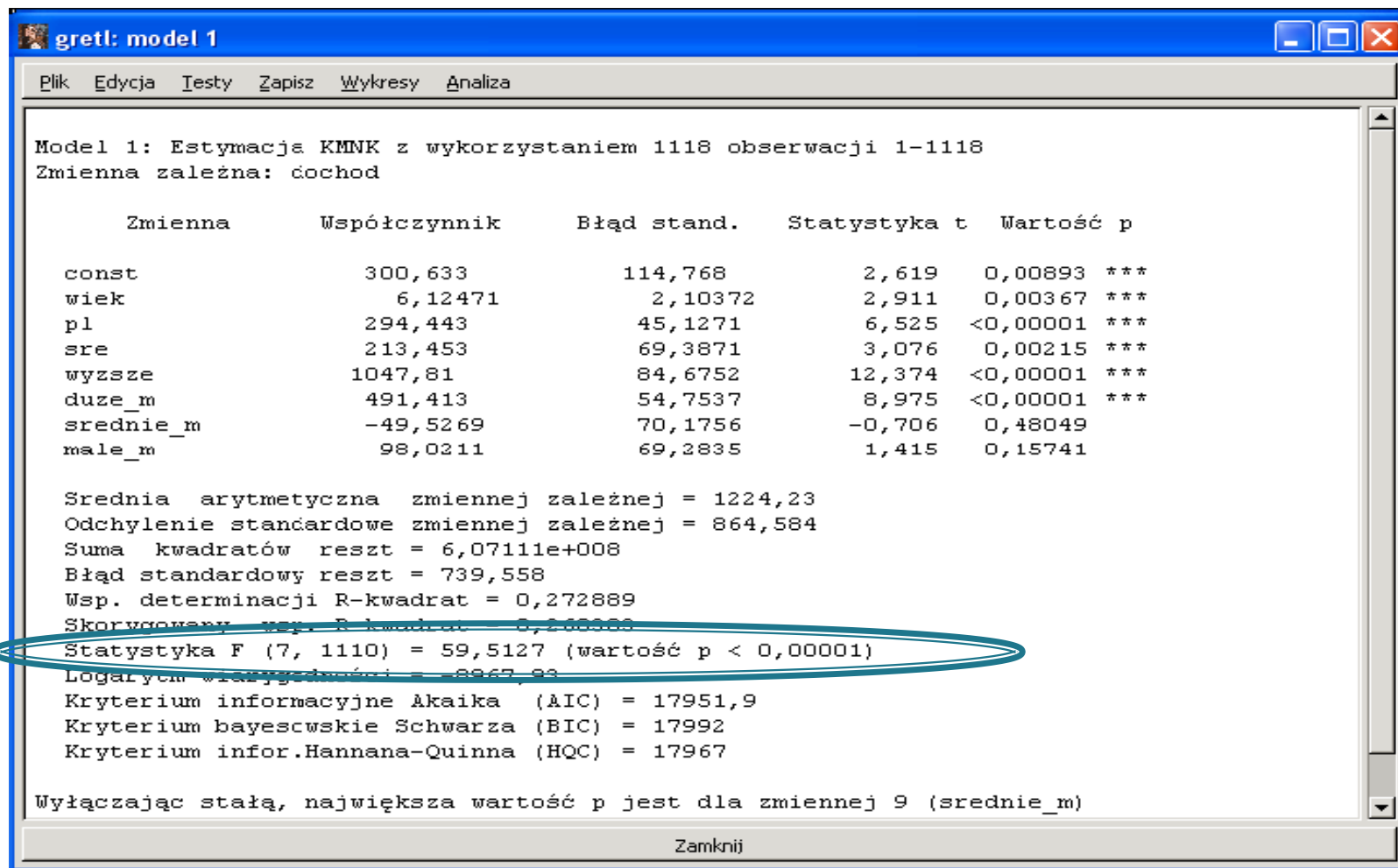
Test na łączną istotność zmiennych objaśniających

► Krok 4. Podjęcie decyzji



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

Przykład



The screenshot shows the 'gretl: model 1' window. The title bar is blue with standard window controls. The menu bar includes 'Plik', 'Edycja', 'Testy', 'Zapisz', 'Wykresy', and 'Analiza'. The main text area displays the following information:

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

Summary statistics:

- Srednia arytmetyczna zmiennej zależnej = 1224,23
- Odchylenie standardowe zmiennej zależnej = 864,584
- Suma kwadratów reszt = 6,07111e+008
- Błąd standardowy reszt = 739,558
- Wsp. determinacji R-kwadrat = 0,272889
- Skorygowany wsp. R-kwadrat = 0,268388
- 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

Pytania teoretyczne

1. Omówić zasady wprowadzania do równania regresji zmiennych objaśniających dyskretnych.
2. Na czym polega statystyczna istotność zmiennej objaśniającej?
3. Co to jest przedział ufności dla nieznanego parametru β_K ?
4. Proszę przedstawić test istotności równania regresji.

Dziękuję za uwagę