

Dekompozycja wariancji, R^2 i jego własności

Interpretacja parametrów

Stanisław Cichocki

Natalia Nehrebecka

Zajęcia 3

Plan zajęć

- ▶ 1. Dobroć dopasowania równania regresji . Współczynnik determinacji R^2
 - Dekompozycja wariancji zmiennej zależnej
 - Współczynnik determinacji R^2
- ▶ 2. Zmienne ciągłe a zmienne dyskretne
- ▶ 3. Liniowość w modelu
- ▶ 4. Interpretacja parametrów przy zmiennych objaśniających ciągłych
 - Interpretacja współczynników regresji w modelu liniowym względem zmiennych objaśniających
 - Elastyczność
 - Semielastyczność

1. Dobroć dopasowania równania regresji .
Współczynnik determinacji R^2



Dobroć dopasowania równania regresji

Współczynnik determinacji R^2

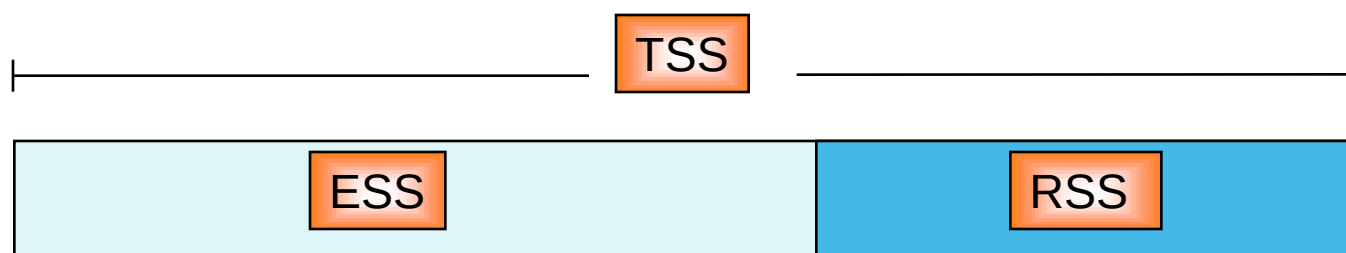
- ▶ Dobroć dopasowania równania regresji (do danych empirycznych) wyrażona jest przez tak zwany **współczynnik determinacji liniowej** oznaczany przez R^2 .



- ▶ Współczynnik ten określa jaka część zmienności zmiennej objaśnianej y jest wyjaśniona łącznie przez zmienność wszystkich zmiennych objaśniających $X_2, \dots, X_K$.
- ▶ Jedną z miar zmienności zmiennej jest WARIANCJA.

Dekompozycja wariancji zmiennej zależnej

Wariancje zmiennej zależnej y można przedstawić jako dekompozycję (podział) na część wyjaśnioną przez model i na część niewyjaśnioną przez model.



Dekompozycja wariancji jest możliwa JEDYNIE dla
modelu ze stałą

Oznaczenia (1/3)

► Całkowita suma kwadratów:

Zmienność całkowitą zmiennej objaśnianej y , oznaczaną w literaturze angielskim skrótem TSS (Total Sum of Squares), mierzymy za pomocą sumy kwadratów odchyleń obserwacji zmiennej objaśnianej od średniej:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Oznaczenia (2/3)

► Wyjaśniona suma kwadratów:

Jeśli model zawiera stałą, to całkowitą sumę kwadratów możemy zdekomponować na dwa składniki, na wyjaśnioną (równaniem regresji) sumę kwadratów, oznaczaną przez ESS (Explained Sum of Squares)

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Oznaczenia (3/3)

- ▶ **Resztowa suma kwadratów:**
- ▶ i resztową (niewyjaśnioną) sumę kwadratów, oznaczaną przez RSS (Residual Sum of Squares).

$$RSS = \sum_{i=1}^n e_i^2$$

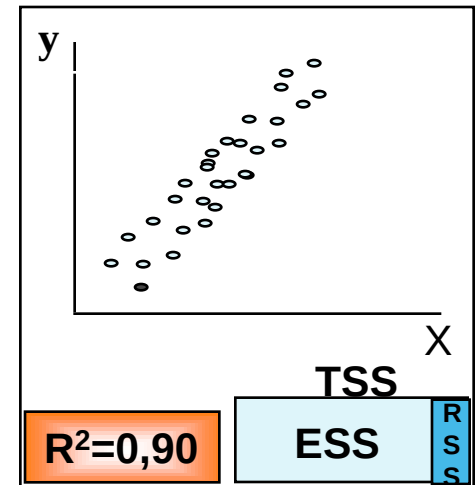
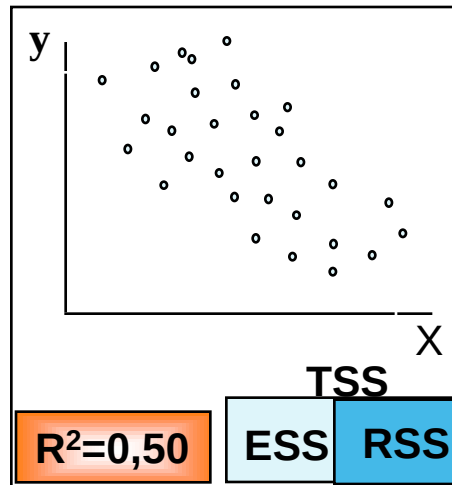
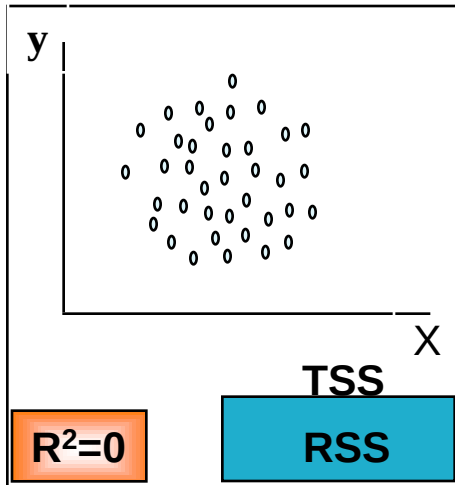
Współczynnik determinacji R^2 (1/3)

$$R^2 = \frac{\text{wyjasniona suma kwadratów}}{\text{całkowita suma kwadratów}} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Dla
modelu ze stałą

$$0 \leq R^2 \leq 1$$

Współczynnik determinacji R^2 (2/3)



- Współczynnik determinacji jest jedną z podstawowych miar jakości dopasowania modelu.
- Informuje o tym, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez model.
- Współczynnik determinacji przyjmuje wartości z przedziału $[0;1]$ w modelu ze stałą.
- Jego wartości najczęściej są wyrażane w procentach.
- Dopasowanie modelu jest tym lepsze, im jego wartość jest bliższa jedności.

Współczynnik determinacji R^2 (3/3)

- ▶ R^2 przyjmuje wartości z przedziału domkniętego między 0 i 1.
- ▶ Jeśli $R^2 = 1$, to model regresji w 100% wyjaśnia zmienność y ,

$$y = \hat{y}, \quad e = 0 \quad \text{oraz} \quad RSS = 0$$

- ▶ a jeśli $R^2 = 0$, to model regresji w ogóle nie wyjaśnia zmienności y .

$$\hat{y} = \bar{y}, \quad ESS = 0$$

- ▶ Jeśli na przykład wynosi $R^2 = 0,7$ to możemy powiedzieć, że
- ▶ 70% zmienności zmiennej objaśnianej y jest wyjaśnione przez łączną zmienność wszystkich zmiennych objaśniających, a 30% zmienności jest niewyjaśnione (jest zmiennością resztową).

Uwaga

- ▶ R^2 jest WYŁĄCZNIE statystyką opisową i nie należy jej stosować do porównywania modeli.
- ▶ Przy szacowaniu kilku modeli dla danej zmiennej zależnej z różną liczbą zmiennych objaśniających na podstawie identycznego zbioru danych, korzystanie ze współczynnika determinacji R^2 dla wyboru modelu lepiej dopasowanego do danych empirycznych staje się problematyczne.

$$placa_i = \beta_1 + \beta_2 wiek_i + \varepsilon_i \quad R^2 = 5\%$$

$$placa_i = \beta_1 + \beta_2 wiek_i + \beta_3 plec_i + \varepsilon_i \quad R^2 = 7\%$$

Uwaga

- ▶ Gdy bowiem dodajemy do równania dalsze zmienne objaśniające to zawsze wzrasta R^2 niezależnie od prawdziwej ważności tych nowododanych zmiennych.
- ▶ Wiąże się to z ogólnymi własnościami optymalizacji.
- ▶ Jeśli, poprzez narzucenia ograniczeń, zmniejszymy zbiór, na którym minimalizujemy funkcję celu, to uzyskana w minimum wartość funkcji celu będzie większa lub równa wartości funkcji w minimum dla minimalizacji bez ograniczeń.

Uwaga

- ▶ Nawet dodając do modelu całkowicie bezsensowną zmienną uzyskujemy polepszenie dopasowania.
- ▶ Co więcej, jeśli $K=N$ (liczba parametrów =liczbie obserwacji), to $R^2=1$
- ▶ Z tego powodu za miarę dobroci dopasowania zaproponowano nie R^2 , a tak zwany „skorygowany współczynnik determinacji” $\bar{R}^2$.

Skorygowany współczynnik determinacji $\bar{R}^2$

- ▶ $\bar{R}^2$ jest skorygowany ze względu na tak zwaną liczbę stopni swobody, to znaczy ze względu na różnicę między liczbą obserwacji **N** a liczbą zmiennych objaśniających **K**.

$$\bar{R}^2 = 1 - \frac{N-1}{N-K} (1 - R^2)$$

Zadanie

Oszacowano dwa modele MNK na próbie liczącej 12 obserwacji i otrzymano następujące wyniki:

a) $\hat{y}_i = 2 + 1,5x_{1i} + 3x_{2i}$, $R^2 = 0,8$

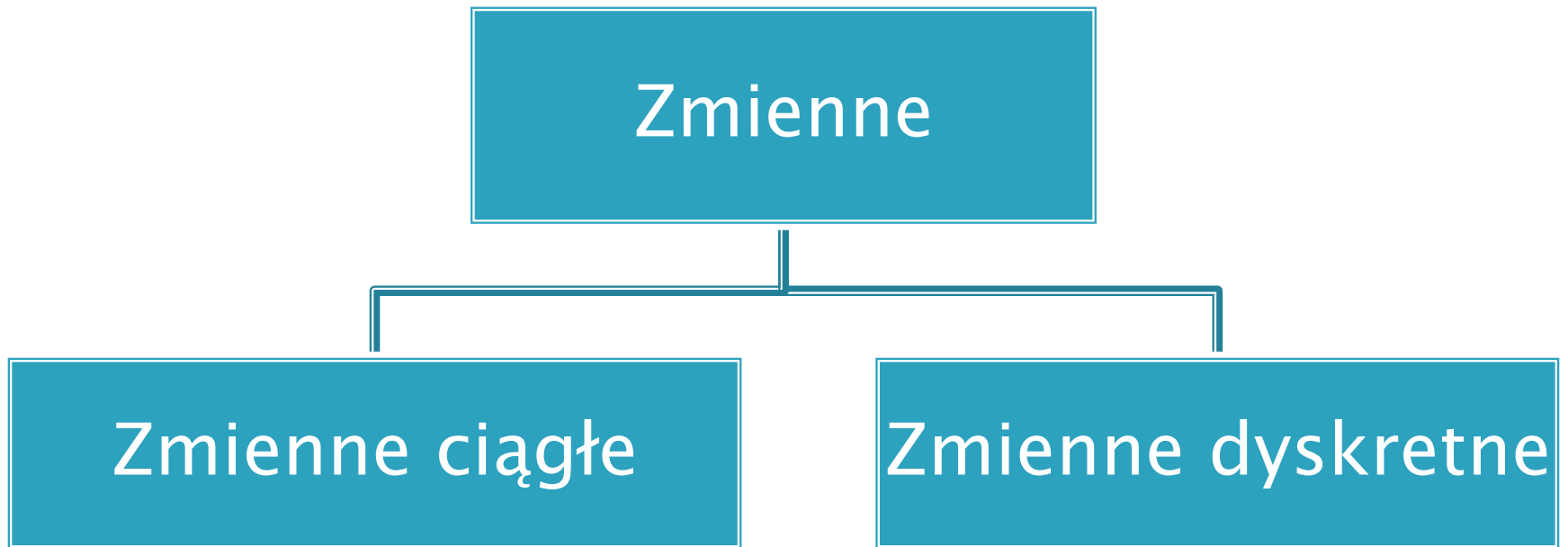
b) $\hat{y}_i = 1 + 0,8x_{1i} + 4x_{2i} + 6x_{3i}$, $R^2 = 0,82$

Który z powyższych modeli należy wybrać i dlaczego?

2. Zmienne ciągłe a zmienne dyskretne



Zmienne



Zmienne ciągłe

- ▶ Zmienną ciągłą nazywamy zmienną, która przyjmuje wartości ze zbioru liczb rzeczywistych.
- ▶ zmiennymi posiadającymi charakter ilościowy
- ▶ Np. dochody, wydatki, cena nieruchomości itd.

Zmienne dyskretne

- ▶ Zmienną dyskretną nazywamy zmienną, która przyjmuje wartości ze skończonego podzbioru liczb naturalnych.
- ▶ Zazwyczaj podzbiór ten jest stosunkowo mało liczny – obejmuje kilka czy kilkanaście elementów.
- ▶ zmiennymi posiadającymi charakter jakościowy
- ▶ np. płeć, wykształcenie, miejsce zamieszkania, stan cywilny i itd.

3. Liniowość w modelu



Liniowość w modelu

- ▶ Liniowość w modelu względem:
- ▶ Po pierwsze - **zmiennych objaśniających** X_{Ki} , które są w pierwszej potęgze,

$$y_i = \beta_1 + \beta_2 X_{2i}^{(1)} + \dots + \beta_K X_{Ki}^{(1)} + \varepsilon_i$$

- ▶ a po drugie - **względem parametrów** β_k , które są również w pierwszej potęgze.

$$y_i = \beta_1^{(1)} + \beta_2^{(1)} X_{2i} + \dots + \beta_K^{(1)} X_{Ki} + \varepsilon_i$$

4. Interpretacja parametrów



- Interpretacja współczynników regresji w modelu liniowym względem zmiennych objaśniających

Interpretacja współczynników regresji w modelu liniowym względem zmiennych objaśniających

- ▶ Model regresji liniowej dla i-tej obserwacji:

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

- ▶ Wartość oczekiwana zmiennej objaśnianej przy danych wartościach zmiennych objaśniających wynosi:

$$\mathbf{E}(y_i) = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} \dots + \beta_K X_{Ki}$$

- ▶ Pochodną cząstkową warunkowej wartości oczekiwanej po X_{Ki}

$$\frac{\partial E(y_i)}{\partial X_{ki}} = \beta_k$$

- ▶ β_k mierzy oczekiwaną zmianę y_i jako efekt zmiany X_{Ki} o jedną jednostkę, **gdy wartości innych zmiennych objaśniających modelu pozostają niezmiennione** (ceteris paribus).

Przykład

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

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

β_2 – współczynnik

INTERPRETACJA: jeżeli wartość zmiennej niezależnej X_{2i} wzrośnie o jednostkę, to wartość zmiennej zależnej y :

- wzrośnie (jeżeli $b_2 > 0$) o $|b_2|$ jednostek lub
- spadnie (jeżeli $b_2 < 0$) o $|b_2|$ jednostek

ceteris paribus.

β_1 – wyraz wolny

Uwaga ! Wyraz wolny nie interpretujemy.

Przykład

$$placa_i = \beta_1 + \beta_2 nauka_i + \beta_3 wiek_i + \varepsilon_i$$

$$\hat{placa}_i = -993,26 + 73,59 \cdot nauka_i + 35,09 \cdot wiek_i$$

- ▶ Zmienna ***nauka_i*** - lata nauki i-tej osoby
- ▶ Interpretacja:
- ▶ Miesięczne wynagrodzenie wzrasta przeciętnie o 73,59 zł przy wzroście liczby lat nauki o jeden rok, przy założeniu pozostałych charakterystyk na nie zmienionym poziomie.

2. Interpretacja parametrów

» Elastyczność

Elastyczność

- ▶ Elastyczności mogą być wyznaczone z modelu - **LOGLINIOWYM**, w którym zarówno zmienna objaśniana jak i zmienne objaśniające są logarytmami zmiennych pierwotnych.

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

$$\ln E(y_i) = \beta_1 + \beta_2 \ln X_{2i} + \dots + \beta_K \ln X_{Ki}$$

$$\frac{\partial \ln E(y_i)}{\partial \ln X_{ki}} = \beta_k$$

▶

β_k mierzy o ile procent zmieni się zmienna objaśniana, gdy zmienna objaśniająca zmieni się o jeden procent, **gdy wartości innych zmiennych objaśniających modelu pozostają niezmiennione** (ceteris paribus).

Przykład

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

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

β_2 – współczynnik

INTERPRETACJA: jeżeli wartość zmiennej niezależnej X_{2i} wzrośnie o 1% ,
to wartość zmiennej zależnej y :

- wzrośnie (jeżeli $b_2 > 0$) o $|b_2|$ % lub
- spadnie (jeżeli $b_2 < 0$) o $|b_2|$ %.

ceteris paribus.

β_1 – wyraz wolny

Uwaga ! Wyraz wolny nie interpretujemy.

Przykład

$$\ln(\text{wydatki}_i) = \beta_1 + \beta_2 \ln(\text{dochód}_i) + \varepsilon_i$$

$$\ln(\widehat{\text{wydatki}}_i) = 2 + 0,25 \cdot \ln(\text{dochód}_i)$$

- ▶ Interpretacja:
- ▶ Miesięczne wydatki wzrastają przeciętnie o 0,25% przy wzroście miesięcznego wynagrodzenia o 1%, przy założeniu pozostałych charakterystyk na nie zmienionym poziomie.

2. Interpretacja parametrów

» Semielastyczność

Semielastyczność

- ▶ Semielastyczności mogą być wyznaczone z modelu, w którym zmienna objaśniana jest zlogarytmowana a zmienne objaśniające nie są logarytmami zmiennych pierwotnych.

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

$$\ln E(y_i) = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki}$$

$$\frac{\partial \ln E(y_i)}{\partial X_{ki}} = \beta_k$$

▶ $\beta_k * 100\%$ mierzy o ile procent zmieni się zmienna objaśniana, gdy zmienna objaśniająca zmieni się o jedną jednostkę, **gdy wartości innych zmiennych objaśniających modelu pozostają niezmiennione** (ceteris paribus).

Przykład

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

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

β_2 – współczynnik

INTERPRETACJA: jeżeli wartość zmiennej niezależnej X_{2i} wzrośnie o 1 jednostkę, to wartość zmiennej zależnej y :

- wzrośnie (jeżeli $b_2 > 0$) o $|b_2| * 100\%$ lub
- spadnie (jeżeli $b_2 < 0$) o $|b_2| * 100\%$.

ceteris paribus.

β_1 – wyraz wolny

Uwaga ! Wyraz wolny nie interpretujemy.

Przykład

$$\ln(placa_i) = \beta_1 + \beta_2 wiek_i + \varepsilon_i$$

$$\ln(\hat{placa}_i) = 2,34 + 0,04 \cdot wiek_i$$

- ▶ Interpretacja:
- ▶ Płaca wzrasta przeciętnie o 4% przy wzroście wieku o 1 rok, przy założeniu pozostałych charakterystyk na nie zmienionym poziomie.

Zadanie

$$\ln(\text{wydatki}_i) = 3,6 + 0,35 \cdot \ln(\text{dochód})_i + 0,11 \cdot \text{dzieci}_i$$

- ▶ Interpretacja:
- ▶ Elastyczność, wzrost dochodu o 1% powoduje wzrost wydatków o 0,35% przy założeniu pozostałych charakterystyk na niezmiennym poziomie.
- ▶ Semielastyczność, wzrost liczby dzieci o 1 powoduje wzrost wydatków o 11%=0,11*100% przy założeniu pozostałych charakterystyk na niezmiennym poziomie.

Pytania teoretyczne



Pytania teoretyczne

1. Proszę opisać wzór dekompozycji zmienności całkowitej zmiennej objaśnianej y na zmienność wyjaśnioną i niewyjaśnioną.
2. Podać interpretacje R^2 .
3. Wyjaśnić, dlaczego R^2 nie można używać do porównania modeli.
4. Jaką interpretację mają współczynniki regresji przy zmiennych objaśniających ciągłych w modelu liniowym względem zmiennych objaśniających?
5. Jaką interpretację mają współczynniki regresji przy zmiennych objaśniających ciągłych w modelu logliniowym względem zmiennych objaśniających?

Dziękuję za uwagę