

Dekompozycja wariancji, R^2 i jego własności

Stanisław Cichocki

Natalia Nehrebecka

Wykład 4

Plan wykładu

- ▶ 1. Własności hiperpłaszczyzny regresji
- ▶ 2. Dobroć dopasowania równania regresji. Współczynnik determinacji R^2
 - Dekompozycja wariancji zmiennej zależnej
 - Współczynnik determinacji R^2

Plan wykładu

- ▶ 1. Własności hiperpłaszczyzny regresji
- ▶ 2. Dobroć dopasowania równania regresji. Współczynnik determinacji R^2
 - Dekompozycja wariancji zmiennej zależnej
 - Współczynnik determinacji R^2

Własności hiperpłaszczyzny regresji

1. $X'e = 0$

2. $\hat{y}'e = 0$

▶ Dodatkowo dla modelu ze stałą:

3. $\sum_{i=1}^N e_i = 0$

4. $\bar{y} = \bar{\hat{y}}$

Plan wykładu

- ▶ 1. Własności hiperpłaszczyzny regresji
- ▶ 2. Dobroć dopasowania równania regresji. Współczynnik determinacji R^2
 - Dekompozycja wariancji zmiennej zależnej
 - Współczynnik determinacji R^2

Dobroć dopasowania równania regresji

Współczynnik determinacji R^2

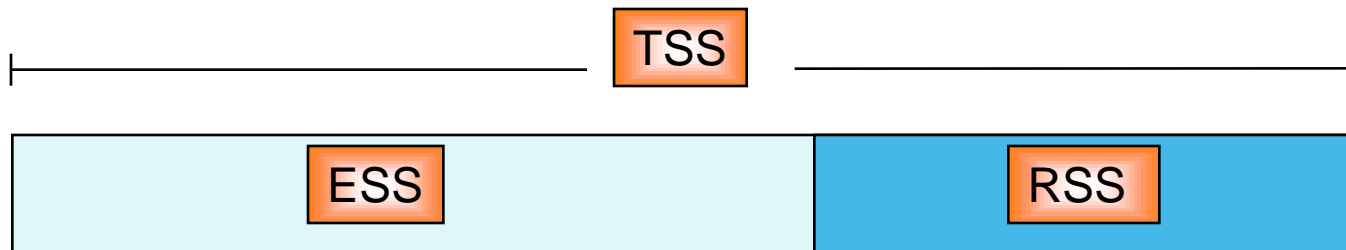
- ▶ Dobroć dopasowania równania regresji (do danych empirycznych) wyrażona jest przez tak zwany **współczynnik determinacji liniowej** oznaczany przez R^2 .



- ▶ 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 dekompozycje (podział) na część wyjaśnioną przez model i na część niewyjaśnioną przez model.



Dekompozycja wariancji jest możliwa **JEDYNYE** 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)

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

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 (1/2)

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

Dla
modelu ze stałą

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

Współczynnik determinacji R^2 (1/2)

Model: $wydatki_i = \beta_1 + \beta_2 \text{dochod}_i + \varepsilon_i$

| Source | SS | df | MS |
|----------|------------|-------|------------|
| Model | 14500.6608 | 1 | 14500.6608 |
| Residual | 56033.0789 | 31845 | 1.75955657 |
| Total | 70533.7398 | 31846 | 2.21483828 |

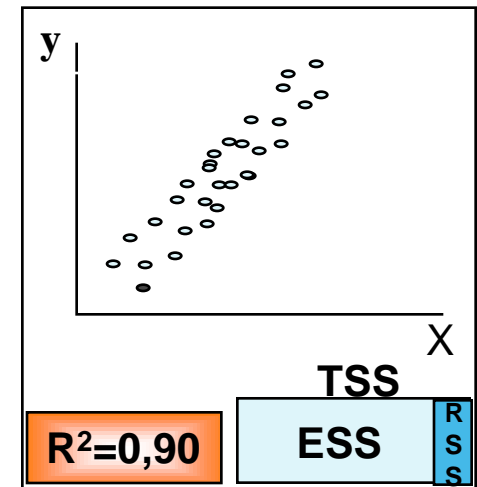
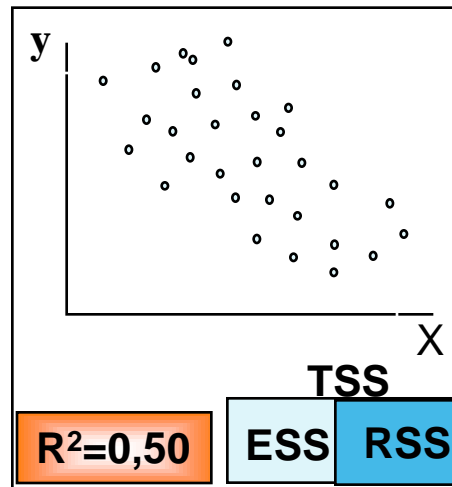
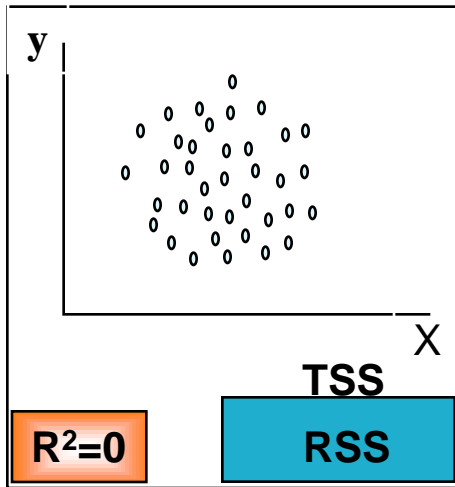
Number of obs = 31847
F(1, 31845) = 8241.09
Prob > F = 0.0000
R-squared = 0.2056
Adj R-squared = 0.2056
Root MSE = 1.3265

TSS

RSS

$$R^2 = \frac{14500,66}{70533,74} = 1 - \frac{56033,08}{70533,74} \approx 0,205$$

Współczynnik determinacji R^2 (2/2)



np. jeśli na przykład $R^2 = 0,7$ 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.
- ▶ Gdy bowiem dodajemy do równania kolejne zmienne objaśniające to zawsze wzrasta R^2 .

$$placa_i = \beta_1 + \beta_2 \text{wiek}_i + \varepsilon_i \quad R^2 = 5\%$$

$$placa_i = \beta_1 + \beta_2 \text{wiek}_i + \beta_3 \text{plec}_i + \varepsilon_i \quad R^2 = 7\%$$

Dziękuję za uwagę