Ekonometria

Własności hiperpłaszczyzny regresji
Dekompozycja wariancji zmiennej zależnej, R^2 i jego własności

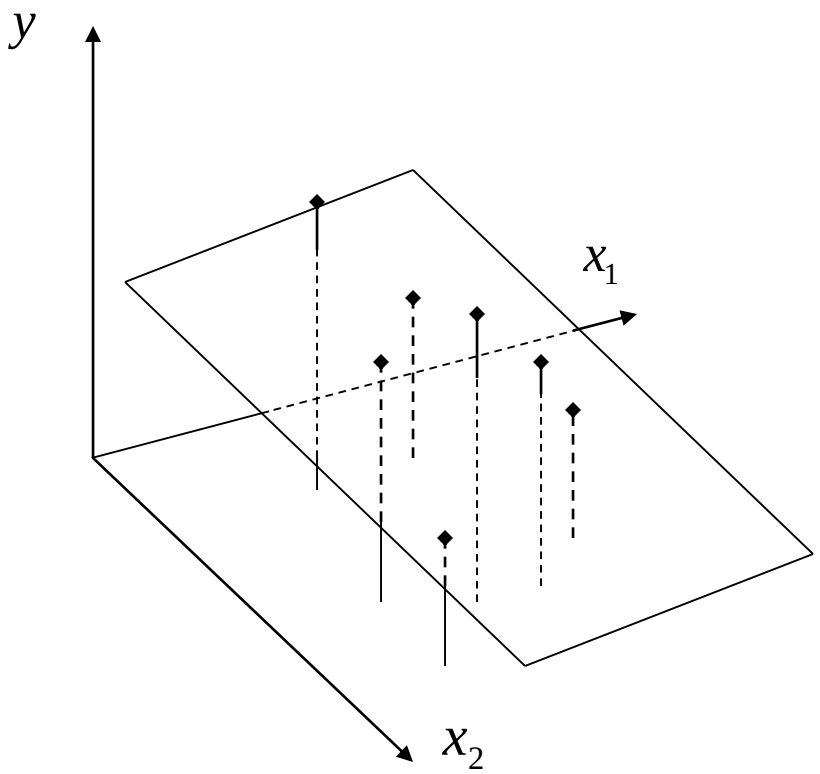
Natalia Nehrebecka
Stanisław Cichocki

Wykład 4

Plan wykładu

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

Hiperpłaszczyzna



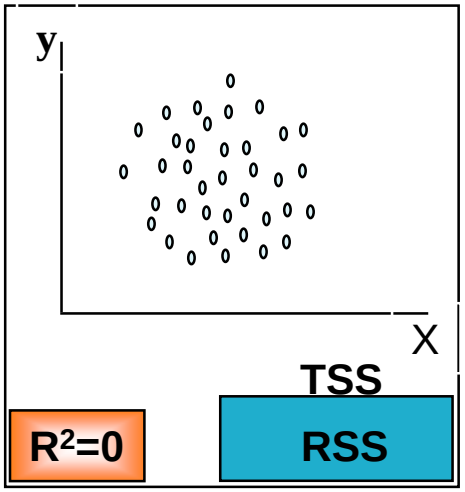
Własności hiperpłaszczyzny regresji

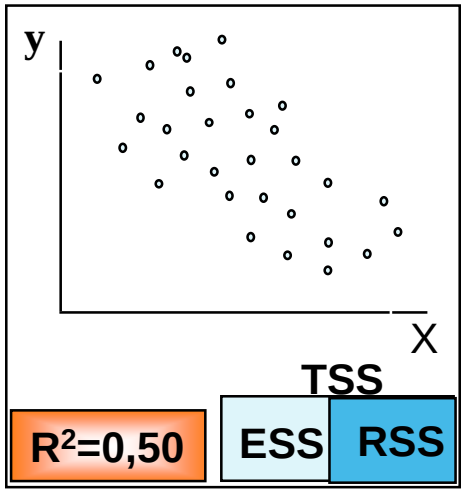
1. $X'e = 0$
2. $\hat{y}'e = 0$

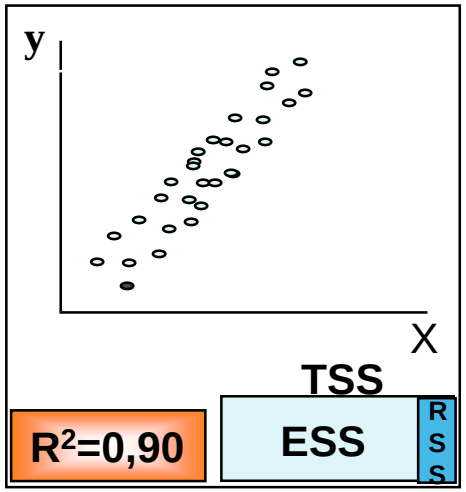
- Dodatkowo dla modelu ze stałą:

3. $\sum_{i=1}^{N} e_i = 0$
4. $\bar{y} = \bar{\hat{y}}$

Współczynnik determinacji R^2







Dobroć dopasowania równania regresji Współczynnik determinacji R^2

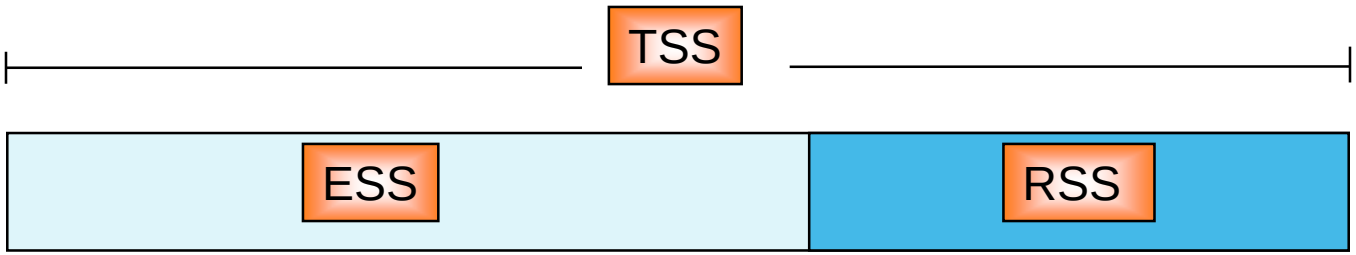
- Dobroć dopasowania prostej regresji (do danych empirycznych) wyrażona jest przez tak zwany **współczynnik determinacji liniowej** oznaczany przez R^2.

- Współczynnik ten określa jaka część zmienności zmiennej objaśnianej y jest wyjaśniona łącznie przez zmienność wszystkich zmiennych objaśniających.

- 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 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 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***).

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

STATA

```
. reg w04 dochg if typ==1
```

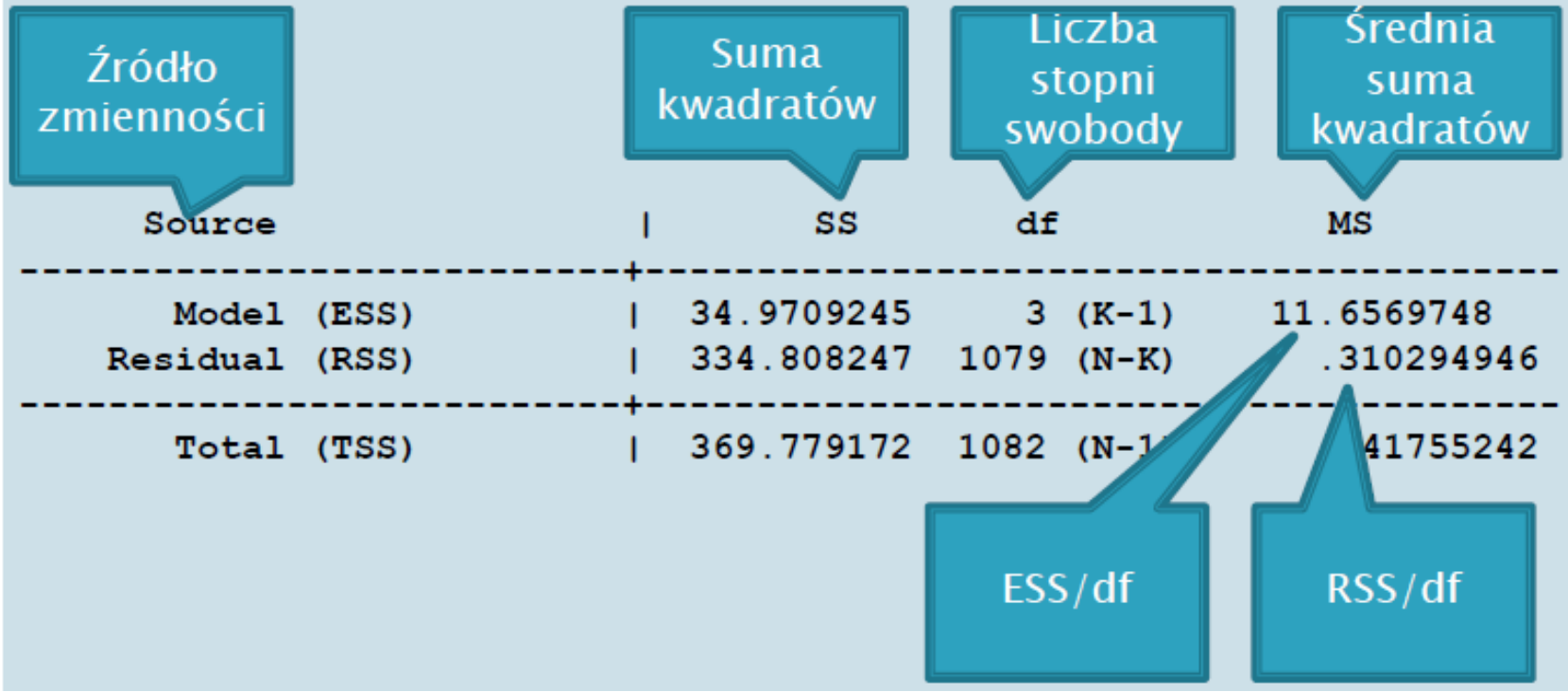
Source	SS	df	MS
Model	203852885	1	203852885
Residual	4.7868e+09	14735	324860.065
Total	4.9907e+09	14736	338671.685

Number of obs	14737
F(1, 14735)	627.51
Prob > F	0.0000
R-squared	0.0408
Adj R-squared	0.0408
Root MSE	569.96

w04	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dochg	.0631331	.0025203	25.05	0.000	.0581931	.0680732
_cons	319.1185	11.04849	28.88	0.000	297.462	340.7749

STATA

Tablica analizy wariancji



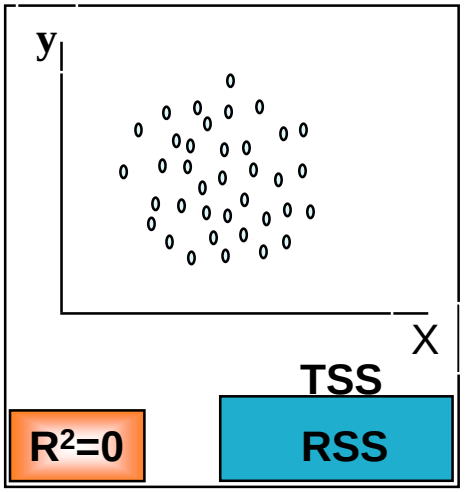
Współczynnik determinacji R^2 (1/3)

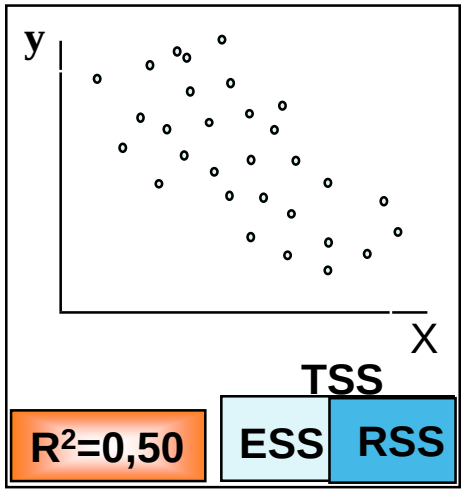
$$R^2 = \frac{wyjasniona\ suma\ kwadratów}{calkowita\ suma\ kwadratów} = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

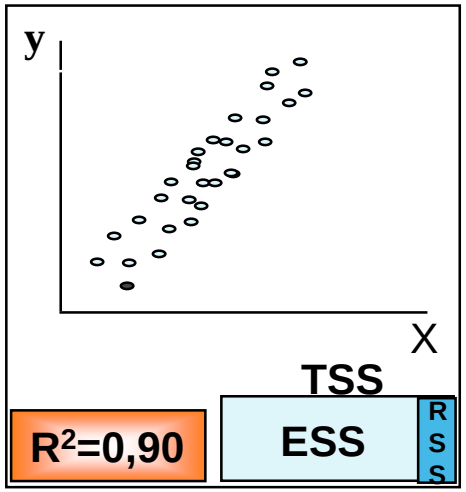
Dla
modelu ze stałą

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

Współczynnik determinacji R^2 (2/3)







Współczynnik determinacji 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,

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

- 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).

Współczynnik determinacji R^2 (1/3)

- Współczynnik determinacji ***R²*** nazywany jest niekiedy ***współczynnikiem korelacji wielorakiej***.

- ***R²*** jest równy kwadratowi współczynnika korelacji między y_i i $\hat{y}_i$

- **W modelu ze stałą i jedną** zmienną współczynnik determinacji ***R²*** jest równy kwadratowi współczynnika korelacji $\hat{\rho}_{xy}$

Współczynnik determinacji R^2 (2/3)

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(b_1 + x_i b_2 - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \left\{\bar{y} = b_1 + b_2\bar{x}\right\} =$$

$$= \frac{\sum_{i=1}^{n}(b_1 + x_i b_2 - b_1 - b_2\bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(x_i b_2 - b_2\bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = b_2^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$b_2 = \frac{S_{yx}}{S_x^2}$$

Współczynnik determinacji R^2 (3/3)

$$R^2 = b_2^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \left[\frac{S_{yx}}{S_x^2}\right]^2 \left[\frac{S_x^2}{S_y^2}\right] = \frac{S_{yx}^2}{S_x^2 S_y^2} = \hat{\rho}_{xy}^2$$

Pytanie?

$$\boldsymbol{placa_i} = \beta_1 + \beta_2 \boldsymbol{wiek_i} + \varepsilon_{1i} \qquad \boldsymbol{R}^2 = 5\%$$

$$\boldsymbol{placa_i} = \alpha_1 + \alpha_2 \boldsymbol{wiek_i} + \alpha_3 \boldsymbol{plec_i} + \varepsilon_{2i} \qquad \boldsymbol{R}^2 = 7\%$$

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 dalsze zmienne objaśniające to zawsze wzrasta R^2 niezależnie od prawdziwej ważności tych nowododanych zmiennych.

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.

Pytania teoretyczne

1. Pokazać , że w modelu ze stałą suma reszt jest równa zero.
2. Pokazać, że w modelu ze stałą średnia wartość zmiennej zależnej równa jest średniej z wartości dopasowanych.
3. Udowodnić, że w modelu ze stałą TSS = ESS + RSS.
4. Podać interpretacje R^2.
5. Wyjaśnić, dlaczego R^2 nie można używać do porównania modeli.

Dziękuję za uwagę