Regresja liniowa

Przypadek jednej zmiennej.

Sebastian Zalas

FAME|GRAPE, Uniwersytet Warszawski

Ekonometria 2022/23

Model prostej regresji liniowej

Załóżmy, że zjawisko ekonomiczne można opisać modelem postaci:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

gdzie:

- i - indeks obserwacji, $i = 1, ..., n$;
- y_i - zmienna zależna, objaśniana;
- x_i - zmienna niezależna, objaśniająca;
- β_0 - wyraz wolny;
- β_1 - nachylenie funkcji regresji w populacji
- u_i - składnik losowy.

Wyprowadzenie MNK I

Dany jest model liniowy:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

- zakładamy, że tak wygląda proces generujący dane
- parametry modelu β_0, β_1 są nieznane
- mamy dane zmienne x_i oraz y_i, gdzie $i = 1, \ldots, n$.
- dopiero w procesie estymacji szukamy $\hat{\beta}_0$, $\hat{\beta}_1$, czyli oszacowań prawdziwych parametrów β_0, β_1
- estymacja $\Rightarrow$Metoda Najmniejszych Kwadratów (OLS, Ordinary Least Squares.)

Wyprowadzenie MNK II

Metoda najmniejszych kwadratów polega na znalezieniu takich wartości $\hat{\beta}_0$, $\hat{\beta}_1$, które minimalizują sumę kwadratów reszt:

$$\min \sum_{i=1}^{n} (\hat{u}_i)^2$$

czyli różnic pomiędzy wartościami obserwowanymi i teoretycznymi:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Warunki pierwszego rzędu:

$$\frac{\partial W}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{1}$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} -2x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{2}$$

Wyprowadzenie MNK III

Przekształcimy równanie (1):

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

zauważmy, że $n\bar{x} = \sum_i^n x_i$ oraz $n\bar{y} = \sum_i^n y_i$

$$n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{x} = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{3}$$

otrzymaliśmy wzór na oszacowanie wyrazu wolnego.

Wyprowadzenie MNK IV

Przekształćmy równanie (2):

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \hat{\beta}_0 - \sum_{i=1}^{n} x_i \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i (\bar{y} - \hat{\beta}_1 \bar{x}) - \sum_{i=1}^{n} x_i \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} + \hat{\beta}_1 n\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{4}$$

Otrzymaliśmy wzór na oszacowanie , pokażemy że jest ono równe stosunkowi kowariancji x_i oraz y_i z próby, do wariancji x_i z próby.

Wyprowadzenie MNK V

Przekształcimy licznik równania (4). Pokażemy, że

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

przekształcając prawą stronę:

$$
\begin{aligned}
&= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} - \sum_{i=1}^{n} y_i \bar{x} + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}
\end{aligned}
$$

Wyprowadzenie MNK VI

Przekształcimy mianownik (4). Pokażemy, że

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

przekształcając prawą stronę:

$$\begin{aligned}
&= \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Oszacowanie MNK modelu liniowego

Korzystając z MNK otrzymujemy oszacowanie:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

- oszacowanie nieznanego parametru β_1

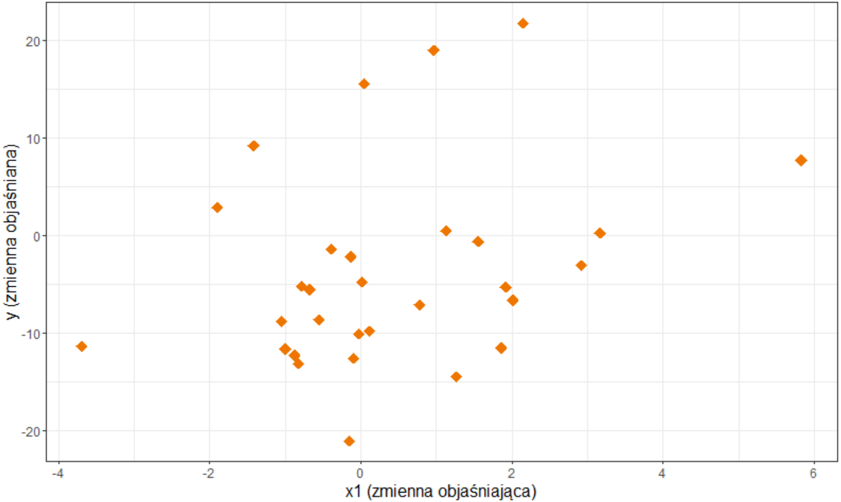
$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{\widehat{\mathrm{Cov}(x_i, y_i)}}{\widehat{\mathrm{Var}(x_i)}}$$

- oszacowanie wyrazu wolnego β_0

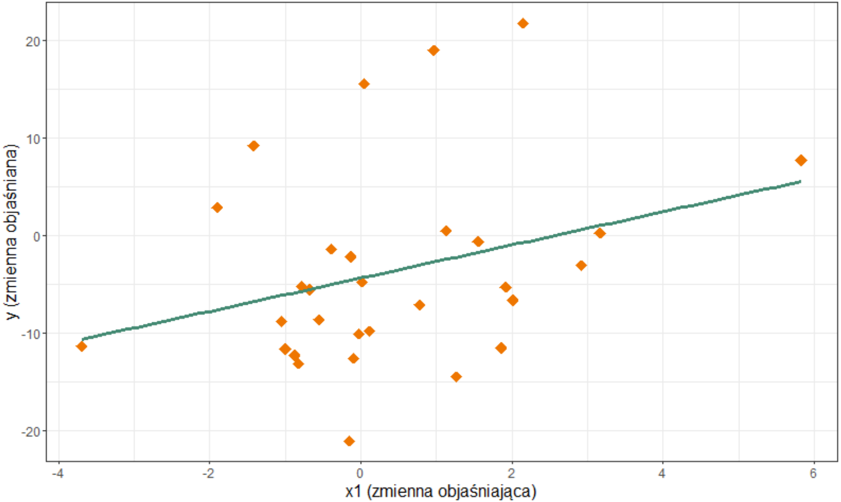
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- wartości teoretyczne: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- reszty: $\hat{u}_i = y_i - \hat{y}_i$

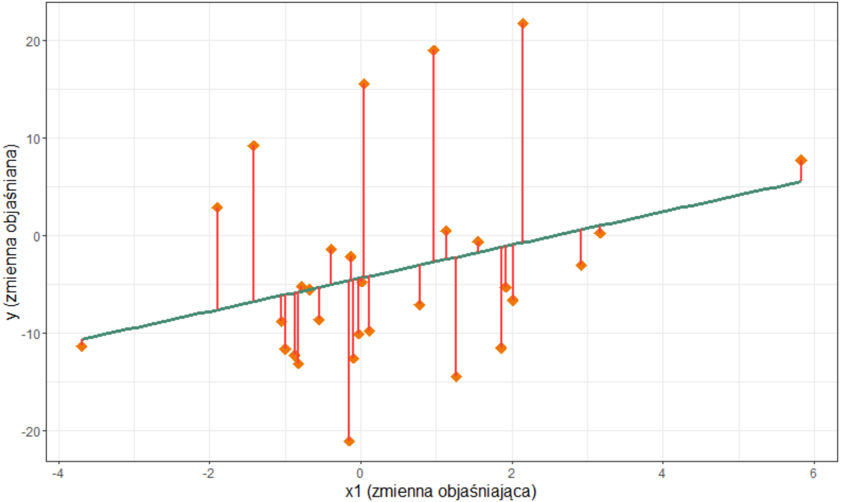
MNK - przedstawienie graficzne



MNK - przedstawienie graficzne



MNK - przedstawienie graficzne



MNK - interpretacja

Przyjmijmy, że oszacowaliśmy parametru modelu liniowego:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

Interpretacja β_1

Jeśli x wzrośnie o jednostkę, to y wzrośnie o β_1 jednostek

- bywa, że β_0 nie ma interpretacji
- uwaga na przyczynowość.

Miara dopasowania R^2

Kiedy oszacowaliśmy model, możemy zastanowić się w jaki sposób ocenić jego dopasowanie do danych $\Rightarrow R^2$

- definiowany jako stosunek wariancji $\hat{y}$ do wariancji z próby y
- interpretacja: jaka część wariancji zmiennej zależnej y została wyjaśniona przez model

Przypomnijmy:

$$y_i = \hat{y}_i + \hat{u}_i$$

Wariancja zmiennej zależnej (*Total Sum of Squares*):

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

Miara dopasowania R^2

Możemy rozłożyć TSS na

- Wariancję $\hat{y}_i$ (*Explained Sum of Squares*)

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

- Wariancję $\hat{u}_i$ (*Residual Sum of Squares*)

$$RSS = \sum_{i=1}^{n} \hat{u}_i{}^2$$

Teraz możemy zdefiniować R^2:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{\sum_{i=1}^{n} (y_i - \bar{y})}$$

ponieważ $TSS = ESS + RSS$.

Pytania? Wątpliwości?
Dziękuję!

e: s.zalas@uw.edu.pl