Regresja liniowa

Ekonometria WNE 2022/23

Sebastian Zalas

Uniwersytet Warszawski
s.zalas@uw.edu.pl

31 października 2023

Ekonometria

- Teoria ekonomii + Metody statystyczne + Dane ekonomiczne

- Najczęściej stosowane narzędzie ekonometryczne: **Metoda Najmniejszych Kwadratów** (MNK), również znana jako **regresja**.
  ang. Ordinary Least Squares, (OLS).

Ekonometria

- Teoria ekonomii + Metody statystyczne + Dane ekonomiczne

- Najczęściej stosowane narzędzie ekonometryczne: **Metoda Najmniejszych Kwadratów** (MNK), również znana jako **regresja**.
  ang. Ordinary Least Squares, (OLS).

Model liniowy

$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y - zmienna zależna
- X - zmienna niezależna
- β_0 - stała, parametr modelu
- β_1 - współczynnik opisujący zależność X i Y, parametr modelu
- ε - błąd losowy

Model liniowy

$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y - zmienna zależna
- X - zmienna niezależna
- β_0 - stała, parametr modelu
- β_1 - współczynnik opisujący zależność X i Y, parametr modelu
- ε - błąd losowy

Model liniowy

$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y - zmienna zależna
- X - zmienna niezależna
- β_0 - stała, parametr modelu
- β_1 - współczynnik opisujący zależność X i Y, parametr modelu
- ε - błąd losowy

Model liniowy

$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y - zmienna zależna
- X - zmienna niezależna
- β_0 - stała, parametr modelu
- β_1 - współczynnik opisujący zależność X i Y, parametr modelu
- ε - błąd losowy

Model liniowy

$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- znamy Y oraz X
- nie znamy β_0, β_1 $\Rightarrow$ jak je znaleźć?
- nie możemy ich wyznaczyć, ale możemy **oszacować**

Model liniowy

$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- znamy Y oraz X
- nie znamy $\beta_0, \beta_1 \Rightarrow$ jak je znaleźć?
- nie możemy ich wyznaczyć, ale możemy **oszacować**

Model liniowy

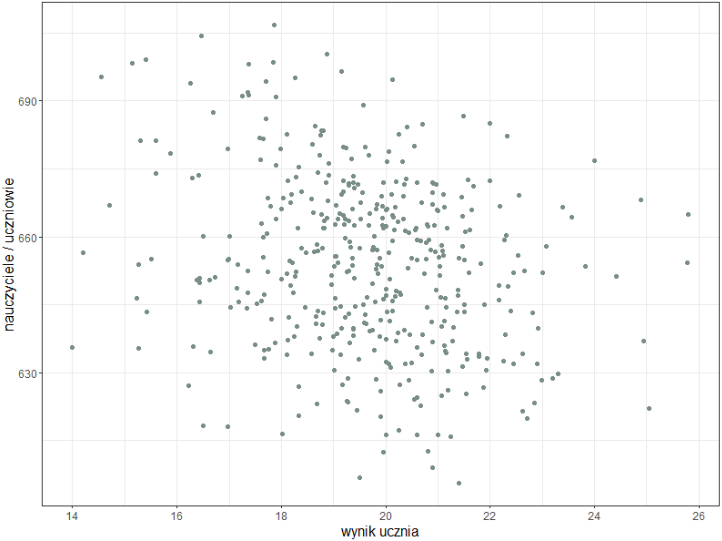
$$Y \sim X$$

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

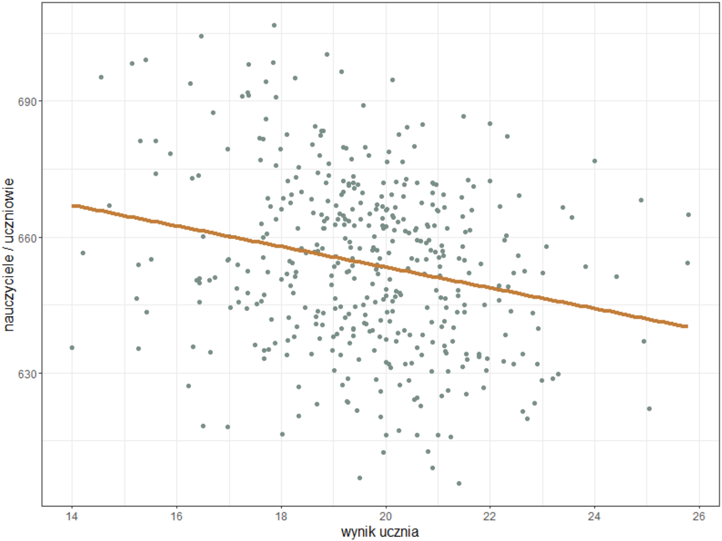
- znamy Y oraz X
- nie znamy β_0, β_1 $\Rightarrow$ jak je znaleźć?
- nie możemy ich wyznaczyć, ale możemy **oszacować**

Przykład

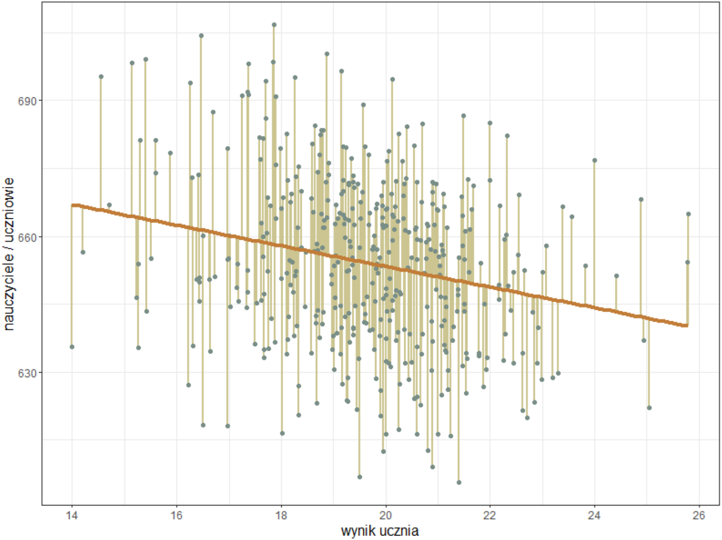


690
660
630
14
16
18
20
22
24
26
wynik ucznia
nauczyciele / uczniowie

Przykład


nauczyciele / uczniowie
690
660
630
14
16
18
20
22
24
26
wynik ucznia

Przykład



690
660
630
nauczyciele / uczniowie
14
16
18
20
22
24
26
wynik ucznia

Metoda Najmniejszych Kwadratów

- szukamy estymatorów: $\hat{\beta}_0 \longrightarrow \beta_0$ oraz $\hat{\beta}_1 \longrightarrow \beta_1$
- minimalizuje sumę kwadratów e_i **reszt** (odchyleń) od danych:

$$\min\left\{\sum_{i=1}^{n} e_i^2\right\}$$

$$\min\left\{\sum_{i=1}^{n}(y_i - (\hat{y}_i))^2\right\}$$

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

- Korzystamy z warunków pierwszego rzędu.

  Dokładne rozwiązanie na końcu prezentacji

Metoda Najmniejszych Kwadratów

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

- estymator MNK parametru β_1, $\hat{\beta}_1$
$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{S_{X,Y}}{S_X}$$
- estymator MNK wyrazu wolnego β_0, $\hat{\beta}_0$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
- estymator jest funkcją danych, więc sam jest zmienną losową.

Oszacowanie modelu liniowego

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

- Y - zmienna zależna
- X - zmienna niezależna
- $\hat{\beta}_0$ - oszacowanie stałej
- $\hat{\beta}_1$ - oszacowanie parametru β_1
- e - reszty

Predykcja

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

- $\hat{Y}$ wartości teoretyczne (dopasowane) zmiennej zależnej:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- e reszty:

$$e = Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$$

- reszty $u \neq \varepsilon$ składnik losowy!

ZAPIS MACIERZOWY

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

- $\boldsymbol{X}_{(n\times 2)}$ - obserwacje zm. niezależnej (objaśniającej)
- $\boldsymbol{y}_{(n\times 1)}$ - obserwacje zm. zależnej (objaśnianej)
- $\varepsilon_{(n\times 1)}$ - wektor składników losowych
- $\beta_{(2\times 1)}$ wektor parametrów β_0, β_1 do oszacowania

Zapis macierzowy

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \\ \vdots & \vdots \\ 1 & x_{n,1} \end{bmatrix}_{n\times 2}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}}_{\boldsymbol{\varepsilon}}$$

Możemy rozszerzyć model do wielu zmiennych

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdot + \beta_k X_k + \varepsilon$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- k - liczba zmiennych niezależnych (objaśniających)
- $\boldsymbol{X}_{(n \times k+1)}$ - obserwacje zm. niezależnych (objaśniających)
- $\boldsymbol{y}_{(n \times 1)}$ - obserwacje zm. zależnej (objaśnianej)
- $\boldsymbol{\varepsilon}_{(n \times 1)}$ - wektor składników losowych
- $\boldsymbol{\beta}_{(k+1 \times 1)}$ wektor $k + 1$ parametrów $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ do oszacowania

Możemy rozszerzyć model do wielu zmiennych

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdot + \beta_k X_k + \varepsilon$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{bmatrix}_{n\times(k+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

$$\boldsymbol{y} = \boldsymbol{X} \quad \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

MNK - MODEL Z WIELOMA ZMIENNYMI

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

wyznaczamy wektor **reszt**:

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$$

Aby wyznaczyć estymator MNK $\hat{\beta}$, minimalizujemy sumę kwadratów reszt:

$$\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + \cdots + e_n^2 = \boldsymbol{e}'\boldsymbol{e}$$

MNK - MODEL Z WIELOMA ZMIENNYMI

Przekształćmy reszty:

$$\begin{aligned} \boldsymbol{e}'\boldsymbol{e} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \end{aligned}$$

Minimalizujemy reszty:

$$\min_{\hat{\beta}}\left\{ \boldsymbol{y}'\boldsymbol{y} - 2\hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \right\}$$

Warunek pierwszego rzędu:

$$\frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\partial \hat{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = 0$$

MNK - MODEL Z WIELOMA ZMIENNYMI

Przekształćmy reszty:

$$\begin{aligned} \boldsymbol{e}'\boldsymbol{e} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \end{aligned}$$

Minimalizujemy reszty:

$$\min_{\hat{\beta}} \left\{ \boldsymbol{y}'\boldsymbol{y} - 2\hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \right\}$$

Warunek pierwszego rzędu:

$$\frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\partial \hat{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = 0$$

MNK - MODEL Z WIELOMA ZMIENNYMI

Przekształćmy reszty:

$$\begin{aligned} \boldsymbol{e}'\boldsymbol{e} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \end{aligned}$$

Minimalizujemy reszty:

$$\min_{\hat{\beta}} \left\{ \boldsymbol{y}'\boldsymbol{y} - 2\hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \right\}$$

Warunek pierwszego rzędu:

$$\frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\partial \hat{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = 0$$

MNK - MODEL Z WIELOMA ZMIENNYMI

Z warunku pierwszego rzędu otrzymujemy układ równań normalnych:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{X}'\boldsymbol{y}$$

Gdy $\boldsymbol{X}$ ma pełen rząd, $\boldsymbol{X}'\boldsymbol{X}$ jest dodatnio określona, odwracalna $\Rightarrow$ układ równań ma jednoznaczne rozwiązanie

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Wyznaczamy $\hat{\beta}$:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

MNK - MODEL Z WIELOMA ZMIENNYMI

Z warunku pierwszego rzędu otrzymujemy układ równań normalnych:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{X}'\boldsymbol{y}$$

Gdy $\boldsymbol{X}$ ma pełen rząd, $\boldsymbol{X}'\boldsymbol{X}$ jest dodatnio określona, odwracalna $\Rightarrow$ układ równań ma jednoznaczne rozwiązanie

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Wyznaczamy $\hat{\beta}$:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

MNK - MODEL Z WIELOMA ZMIENNYMI

Z warunku pierwszego rzędu otrzymujemy układ równań normalnych:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{X}'\boldsymbol{y}$$

Gdy $\boldsymbol{X}$ ma pełen rząd, $\boldsymbol{X}'\boldsymbol{X}$ jest dodatnio określona, odwracalna $\Rightarrow$ układ równań ma jednoznaczne rozwiązanie

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Wyznaczamy $\hat{\beta}$:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Predykcja - model z wieloma zmiennymi

$$\boldsymbol{y} = \boldsymbol{X}\hat{\beta} + \boldsymbol{e}$$

- $\hat{\boldsymbol{y}}$ wartości teoretyczne (dopasowane) zmiennej zależnej:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$$

- $\boldsymbol{e}$ wektor reszt:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$$

Kluczowe pojęcia

model linowy, metoda najmniejszych kwadratów, parametry, oszacowania, wartości dopasowane (teoretyczne), reszty, składnik losowy, MNK w zapisie macierzowym, minimalizacja sumy kwadratów reszt, estymator MNK, wariancja & kowariancja z próby (empiryczna), własności estymatora MNK (patrz zadanie 1 z listy), macierz $\boldsymbol{X}$, zmienna zależna (objaśniania), zmienna niezależna (objaśniająca)

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl

Wyprowadzenie MNK

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

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

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

Warunki pierwszego rzędu:

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

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

Wyprowadzenie MNK

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

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 $\hat{\beta}_1$, pokażemy że jest ono równe stosunkowi kowariancji x_i oraz y_i z próby, do wariancji x_i z próby.

Wyprowadzenie MNK

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

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:

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