Ekonometria

Metoda Najmniejszych Kwadratów cz. 1

Natalia Nehrebecka
Stanisław Cichocki

Wykład 2

Plan wykładu

- 1. Model liniowy
  - Postać modelu liniowego
  - Zapis macierzowy modelu liniowego
- 2. Estymacja modelu
  - Wstęp
  - Wartość teoretyczna (dopasowana)
  - Reszty
  - Metoda Najmniejszych Kwadratów
- 3. MNK – przypadek jednej zmiennej

Plan wykładu

- 1. Model liniowy
  - Postać modelu liniowego
  - Zapis macierzowy modelu liniowego
- 2. Estymacja modelu
  - Wstęp
  - Wartość teoretyczna (dopasowana)
  - Reszty
  - Metoda Najmniejszych Kwadratów
- 3. MNK – przypadek jednej zmiennej

Postać modelu liniowego

teoria ekonomiczna  dane empiryczne

zależności ilościowe między zmiennymi

badanie ekonometryczne

Postać modelu liniowego – równanie regresji w populacji

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + \varepsilon_i$$

- y_i – zmienna objaśniana (*zależna, endogeniczna*),
- $x_1, \ldots, x_K$ – zmienne objaśniające (*niezależne, egzogeniczne*),
- ε_i – błąd losowy,
- $\beta_1, \ldots, \beta_K$ – nieznane parametry,
- $i = 1, \ldots, N$
- i – indeks obserwacji,
- N – liczba obserwacji.

Postać modelu – równanie regresji w populacji

y	x
Zmienna objaśniana	Zmienna objaśniająca
Zmienna zależna	Zmienna niezależna
Zmienna endogeniczna	Zmienna egzogeniczna
Regresant	Regresor
	Zmienna kontrolna
	Predyktor

Pytanie

- Który z modeli jest poprawny i dlaczego?
- Co jest zmienną objaśnianą a co objaśniającą?

$$wydatki_i = \beta_1 + \beta_2 dochód_i + \varepsilon_i$$

$$dochód_i = \beta_1 + \beta_2 wydatki_i + \varepsilon_i$$

Przykład

- Związek przyczynowo-skutkowy ≠ korelacja

Przykład

Stwierdzono dodatnią korelację między wielkością spożycia lodów w danym dniu i liczbą utonięć w tym dniu. Czy po zjedzeniu lodów nie powinno się wchodzić do wody?

Odpowiedź

Więcej utonięć zdarza się w ciepłe dni (kąpie się wtedy więcej osób). W takie dni jest też większe spożycie lodów. Jednak spożycie lodów nie powoduje utonięcia

Czyli: *występuje korelacja między zdarzeniami ale nie ma między nimi związku przyczynowo-skutkowego*

Postać modelu – równanie regresji w populacji

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + \varepsilon_i$$

- y_i – zmienna objaśniana (*zależna, endogeniczna*),
- $x_1, \ldots, x_K$ – zmienne objaśniające (*niezależne, egzogeniczne*),
- ε_i – błąd losowy,
- $\beta_1, \ldots, \beta_K$ – nieznane parametry,
- $i = 1, \ldots, N$
- i – indeks obserwacji,
- N – liczba obserwacji.

Postać modelu - zapis macierzowy

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{K1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{KN} \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

Stąd równanie macierzowe ma postać:

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

Przykład - zapis macierzowy

$$wydatki_i = \beta_1 + \beta_2 dochód_i + \varepsilon_i$$

id	wydatki	dochód
1	639,09	890,6
2	664,47	2300
3	467,55	1814,5
…	…	…

- $y - wydatki$
- $X - dochód$

$$\mathbf{y} = \begin{bmatrix} 639.09 \\ 664.47 \\ 467.55 \\ \vdots \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 890.6 \\ 1 & 2300 \\ 1 & 1814.5 \\ \vdots & \vdots \end{bmatrix}$$

Zatem…

- Zależność między analizowanymi zmiennymi jest **liniowa**
  - (równanie regresji liniowej wyznacza hiperpłaszczyznę regresji)

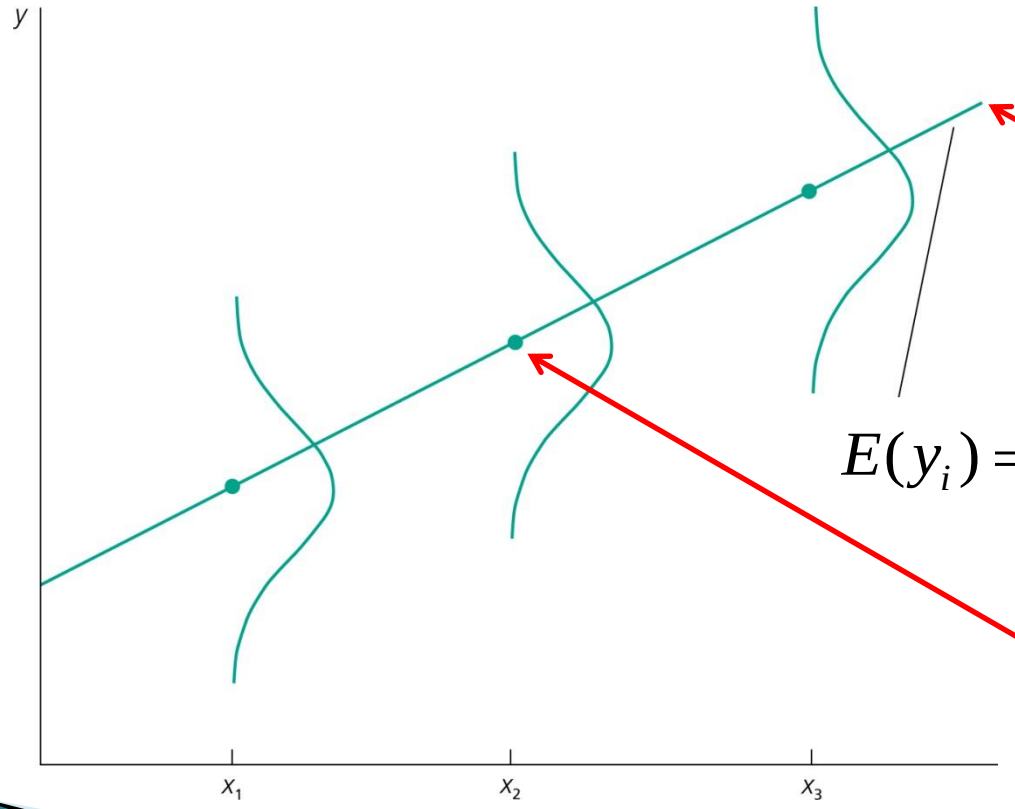
- Istnieje zależność przyczynowo-skutkowa między zmiennymi (≠korelacja)
  - zmienne objaśniające są przyczyną zmienności zmiennej objaśnianej
  - zależność zwykle wynika z teorii (powinna)

- Pewna część zmienności zmiennej objaśnianej pozostaje niewyjaśniona, bo:
  - nieuwzględnienie pewnych zmiennych objaśniających
  - losowy charakter czynników wpływających na zmienną objaśnianą

Plan wykładu

- 1. Model liniowy
  - Postać modelu liniowego
  - Zapis macierzowy modelu liniowego
- 2. Estymacja modelu
  - Wstęp
  - Wartość teoretyczna (dopasowana)
  - Reszty
  - Metoda Najmniejszych Kwadratów
- 3. MNK – przypadek jednej zmiennej

Prosta regresji (*populacja*)

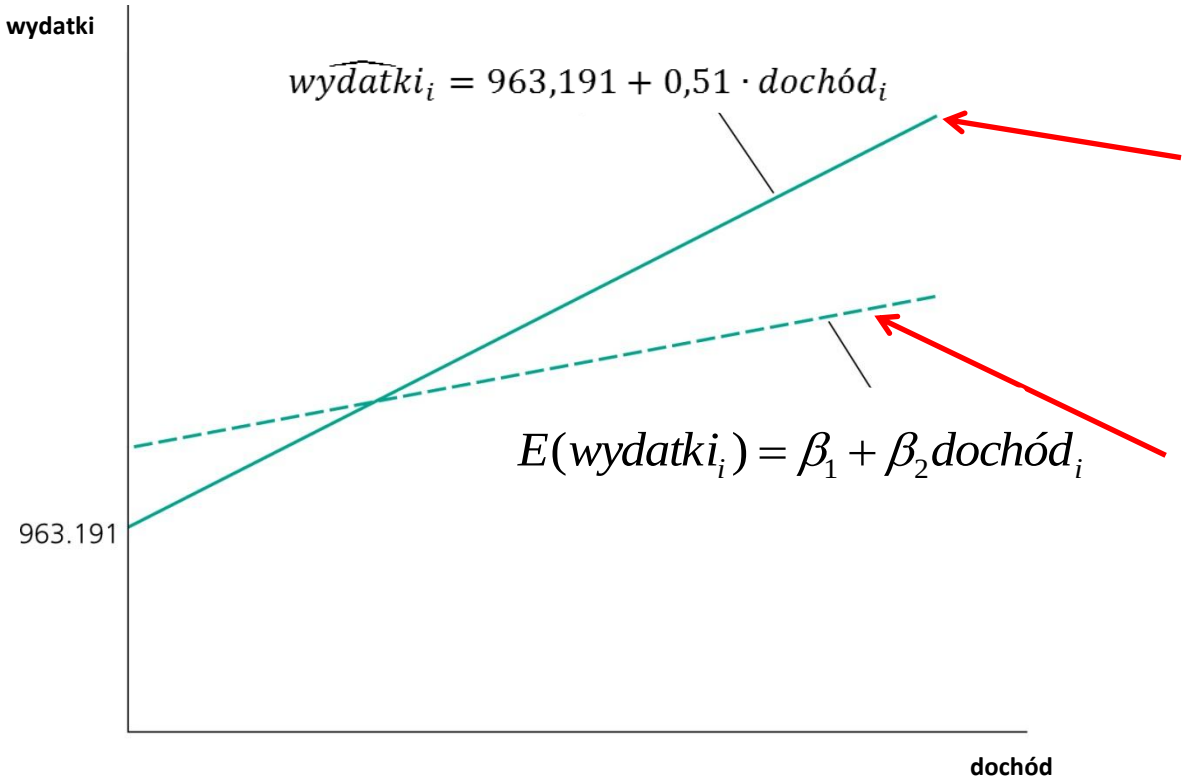


Prosta regresji (populacja)

$$E(y_i) = \beta_1 + x_{2i}\beta_2$$

Populacja vs próba

$$wydatki_i = \beta_1 + \beta_2 dochód_i + \varepsilon_i$$



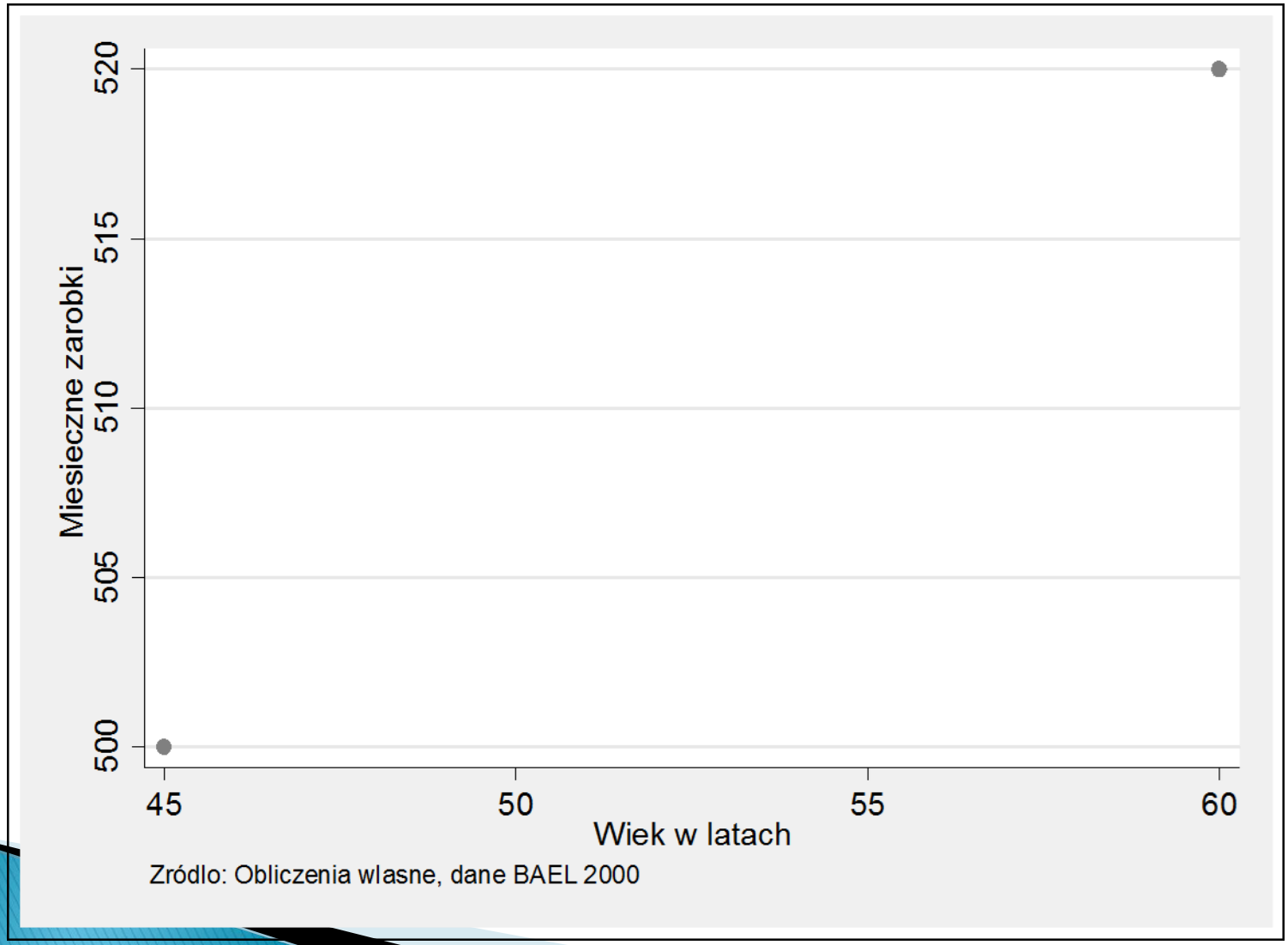
Estymacja - wstęp

- Teoria zwykle nie dostarcza informacji nt. wielkości parametrów modelu ($\beta_1, \ldots, \beta_K$).

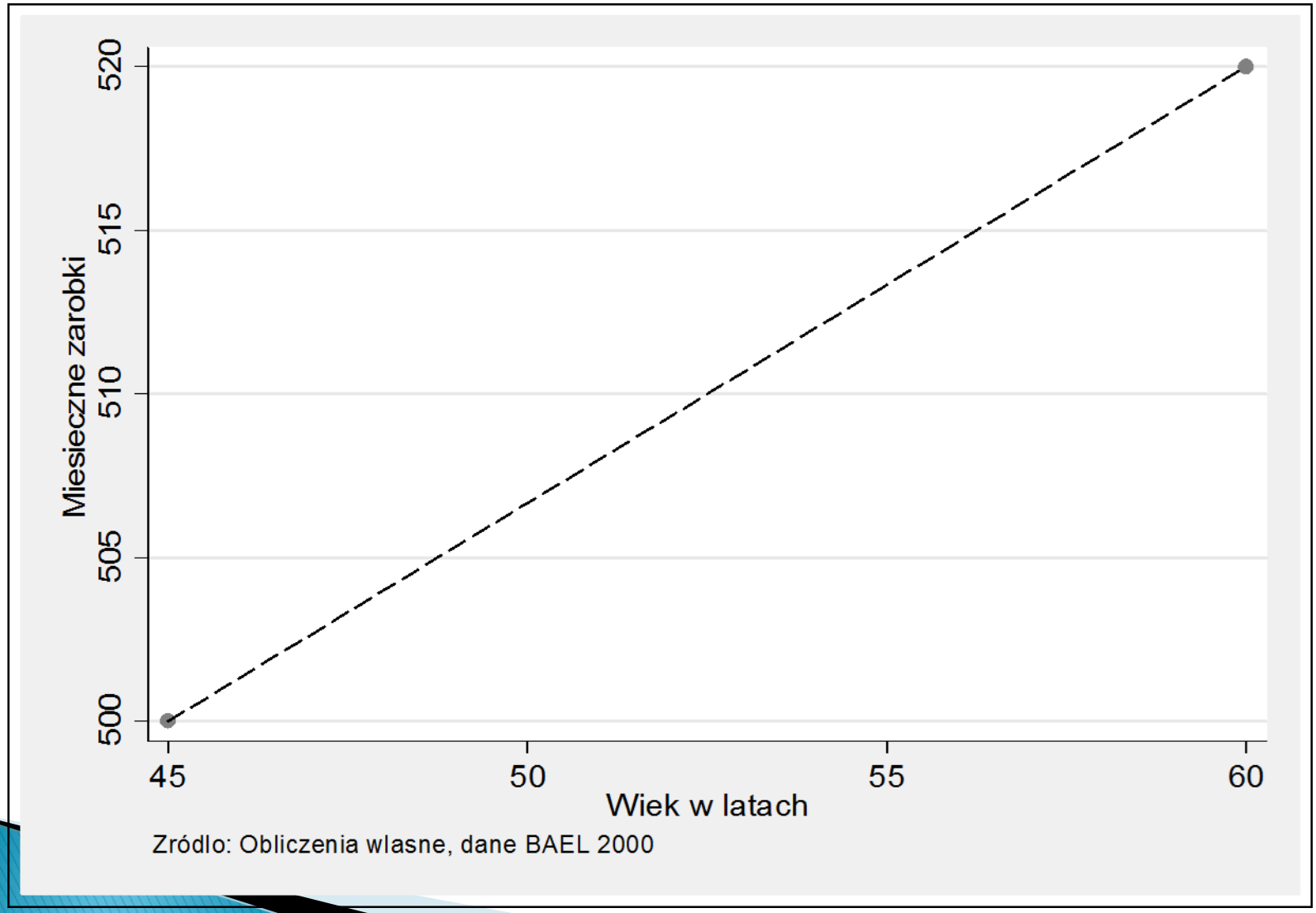
- Wielkość nieznanych parametrów należy oszacować (*estymować*) na podstawie danych empirycznych (*próby*).

- Oszacowane wielkości parametrów (*estymatory*) ($b_1,\ldots,b_K$) są niedokładne (*losowe*), zależą od próby.

Estymacja - przykład

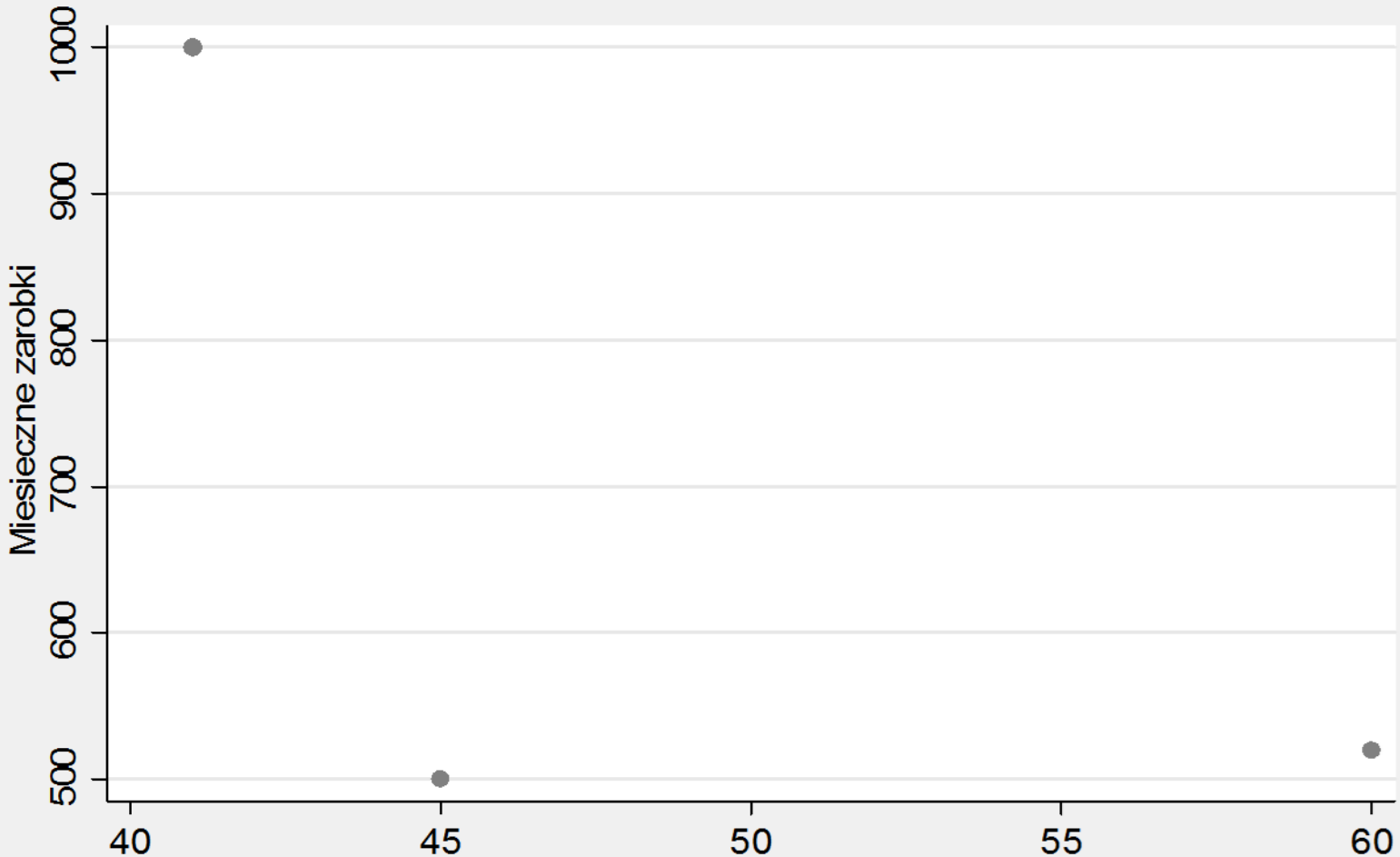


Estymacja - przykład



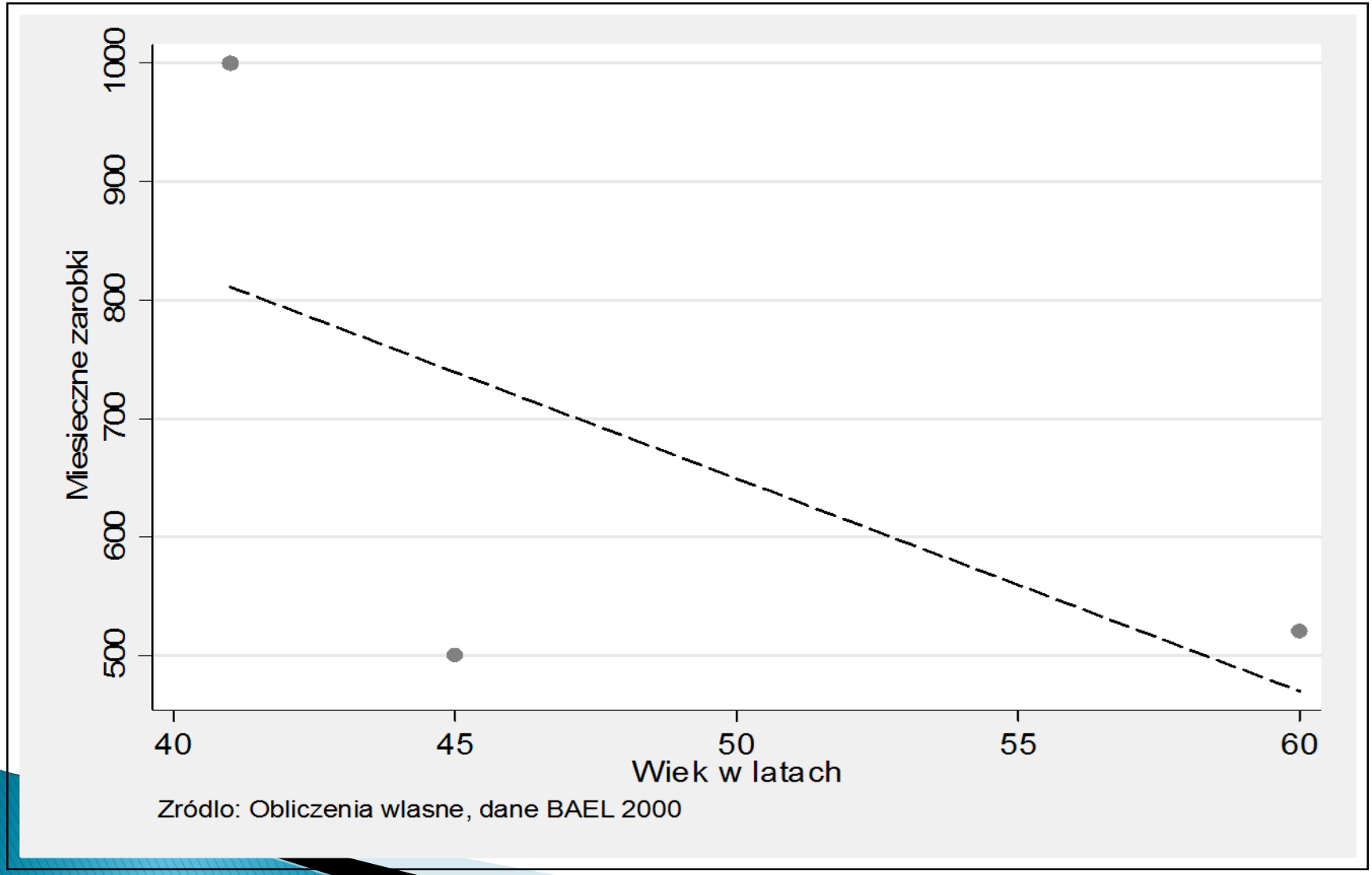
Zródlo: Obliczenia wlasne, dane BAEL 2000

Estymacja - przykład



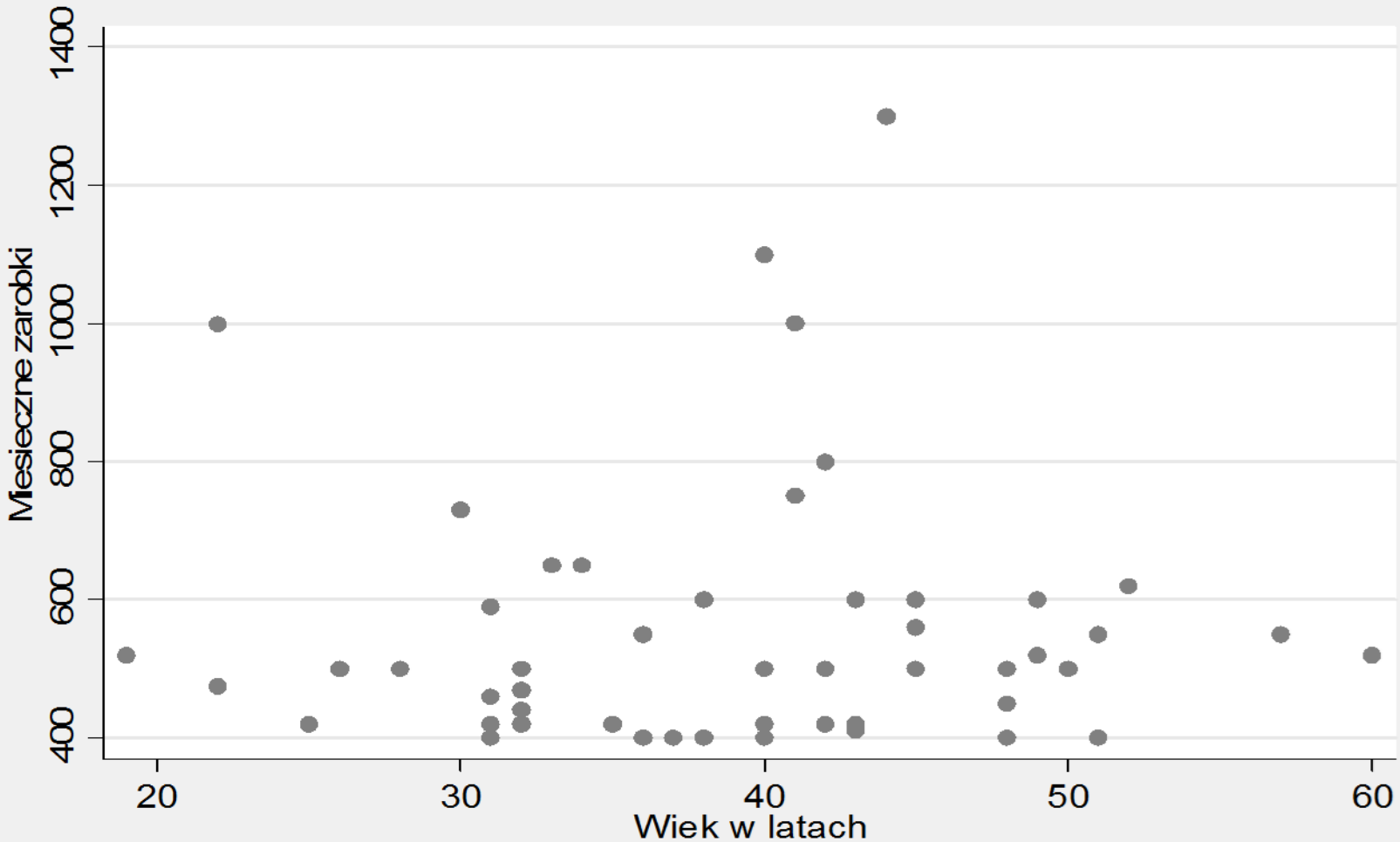
Zródlo: Obliczenia wlasne, dane BAEL 2000

Estymacja - przykład



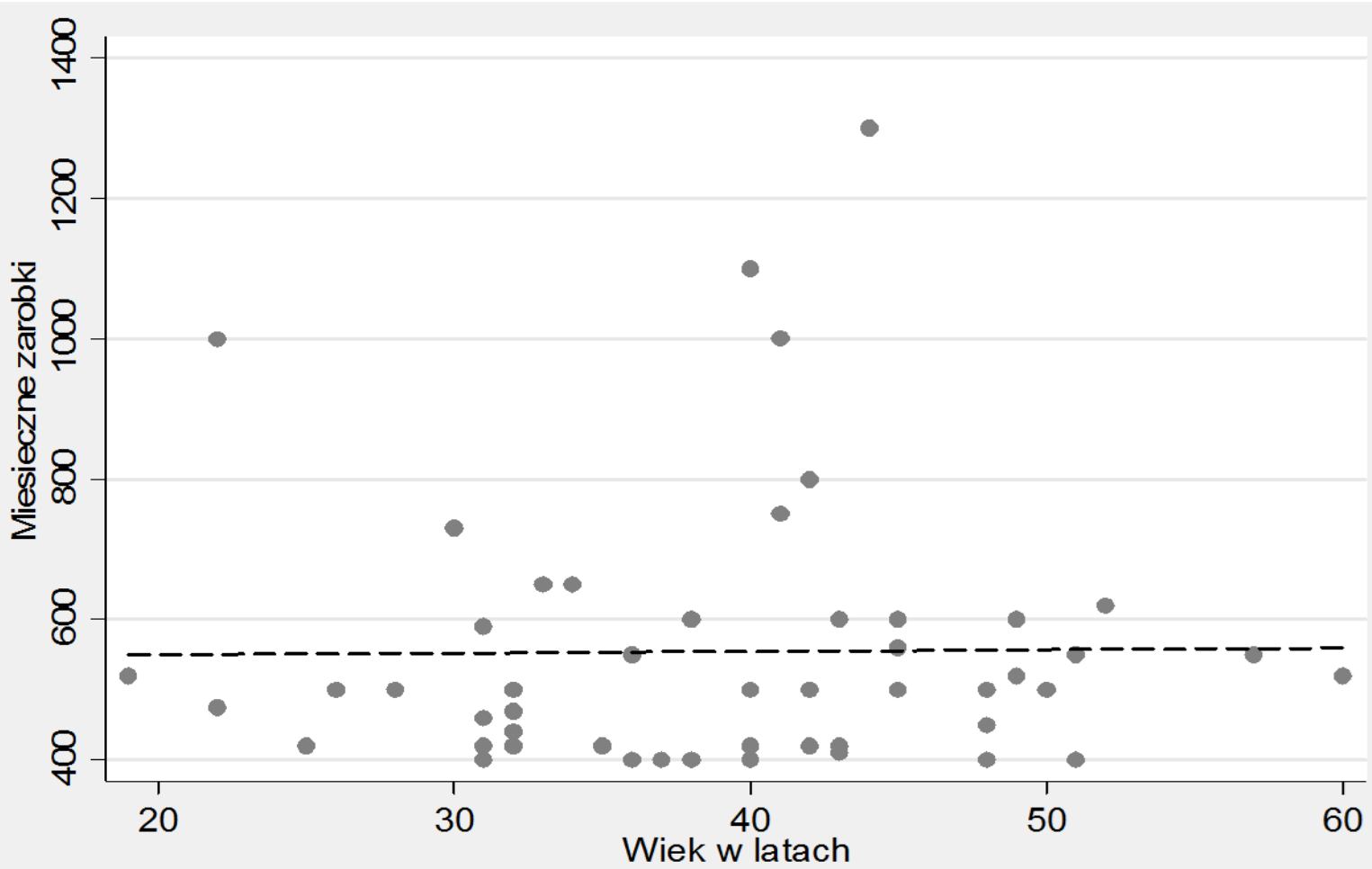
Zródlo: Obliczenia wlasne, dane BAEL 2000

Estymacja - przykład



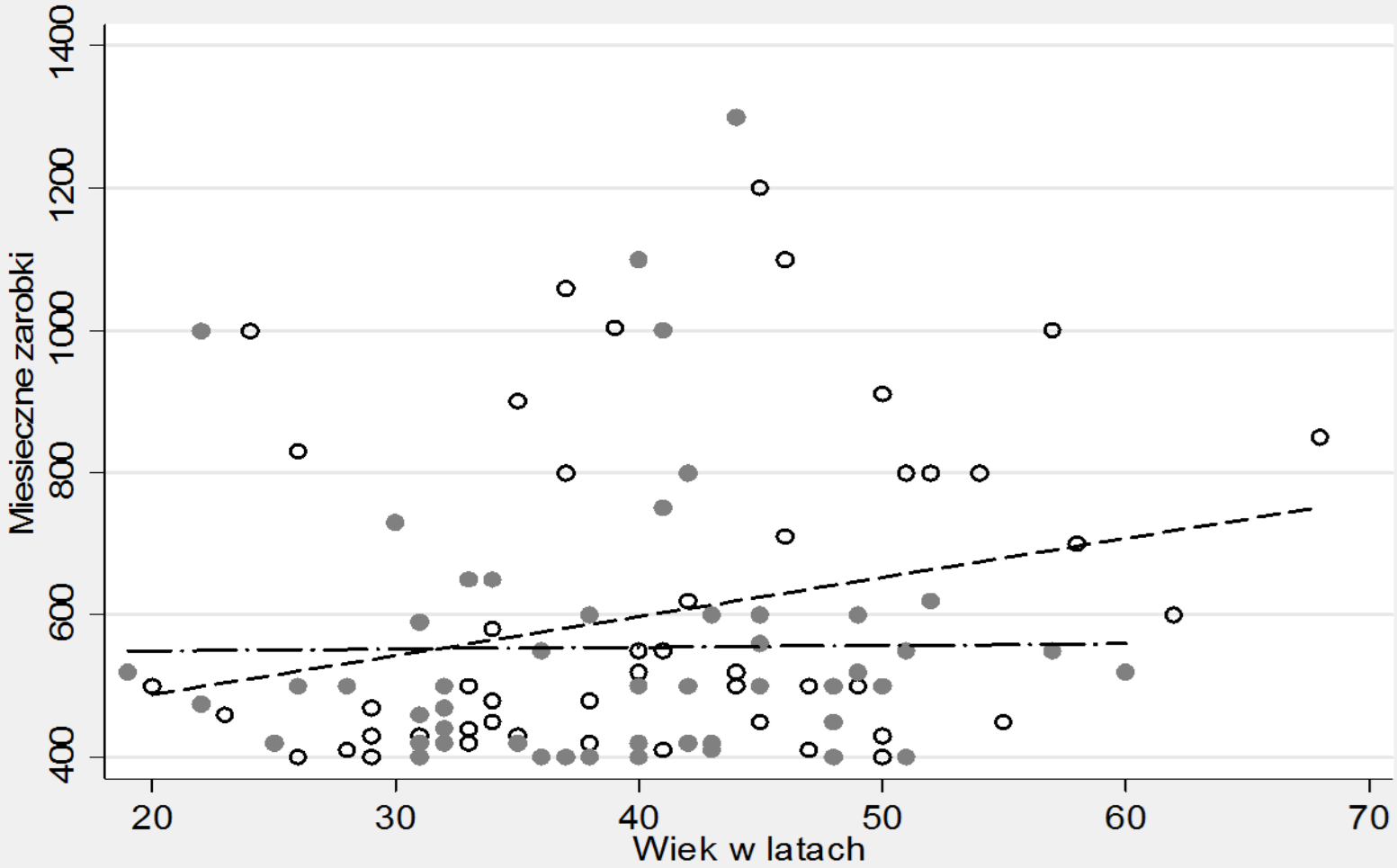
Zródlo: Obliczenia wlasne, dane BAEL 2000

Estymacja - przykład



Zródlo: Obliczenia wlasne, dane BAEL 2000

Estymacja - przykład



Zródlo: Obliczenia wlasne, dane BAEL 2000

Wartość teoretyczna (dopasowana)

- Wartości dopasowane: wartości zmiennej objaśnianej (y_i) przewidywane na podstawie oszacowanego modelu - regresji liniowej y_i na $x_{1i}, \ldots, x_{Ki}$:

$$\hat{y}_i = x_{1i}b_1 + x_{2i}b_2 + \ldots + x_{Ki}b_K$$

- Różnią się od wartości rzeczywistych, bo:
  - zamiast nieznanych prawdziwych wielkości parametrów ($\beta_1, \ldots, \beta_K$) używamy ich estymatorów ($b_1, \ldots, b_K$)
  - pomijamy błąd losowy (ε_i)

Reszty

- **Reszty:** różnica między wartością rzeczywistą a dopasowaną zmiennej objaśnianej, są to oszacowania (ε_i) :

$$e_i = y_i - \hat{y}_i = y_i - x_{1i}b_1 - x_{2i}b_2 - \ldots - x_{Ki}b_K$$

- Zależność między resztami, obserwacjami i oszacowaniami parametrów:

$$y_i = \hat{y}_i + e_i = x_{1i}b_1 + x_{2i}b_2 + \ldots + x_{Ki}b_K + e_i$$

Estymatory i reszty

- **Estymatory** ($b_1, \ldots, b_K$) są to oszacowania ($\beta_1, \ldots, \beta_K$), ale nie są im równe

- **Reszty** (e_i) są to oszacowania (ε_i), ale nie są im równe

Przykład

- Analizujemy wydatki na żywność w gospodarstwach małżeństw pracowniczych z dwójką dzieci w zależności od dochodu

$$wydatki_i = \beta_1 + \beta_2 dochód_i + \varepsilon_i$$

Przykład

- $$\widehat{wydatki}_i = 463 + 0{,}08 \cdot dochód_i$$

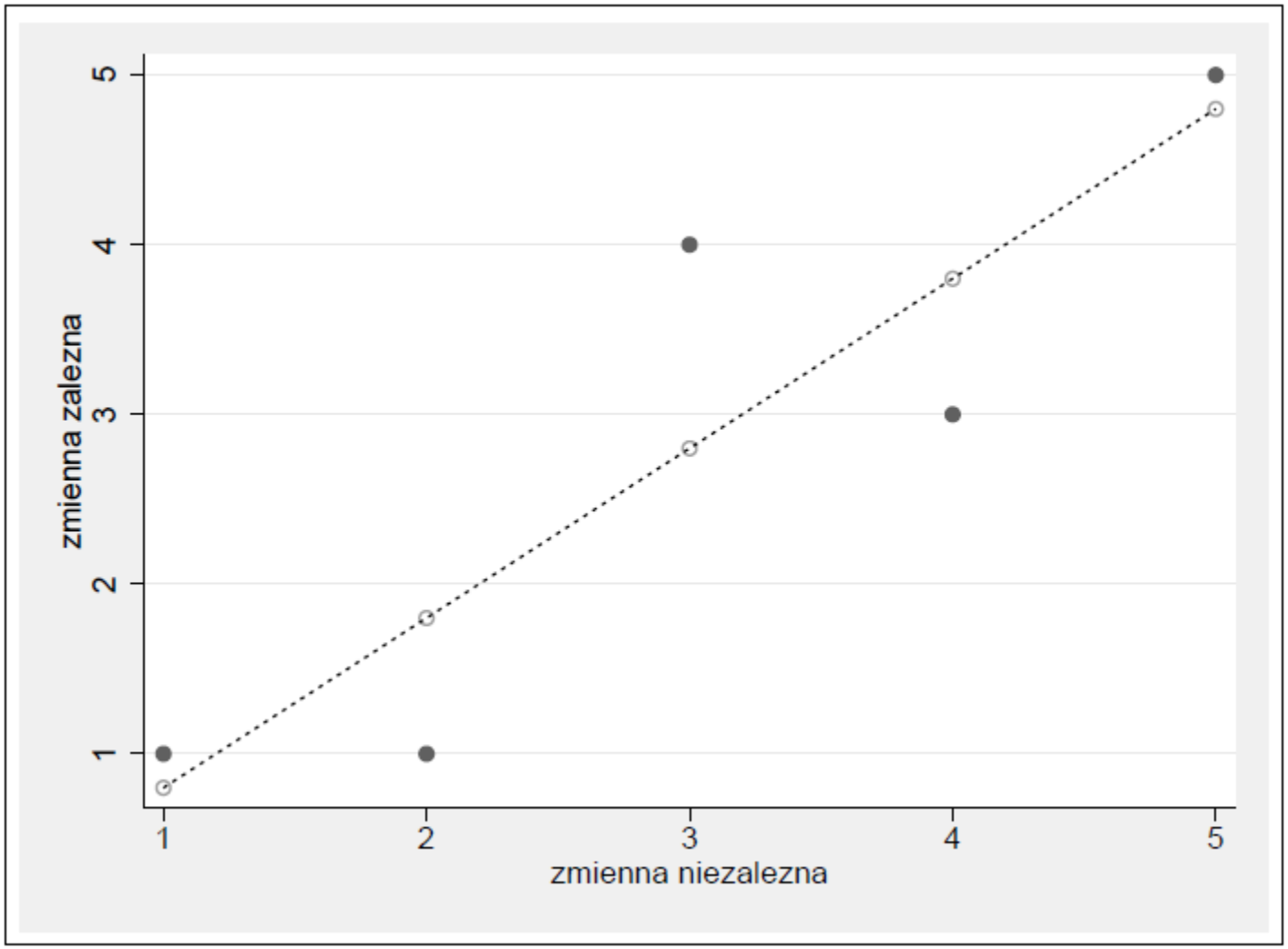
- Dla pierwszej jednostki w badaniu:

Id	Wydatki	Dochód
1	639,1	890,6

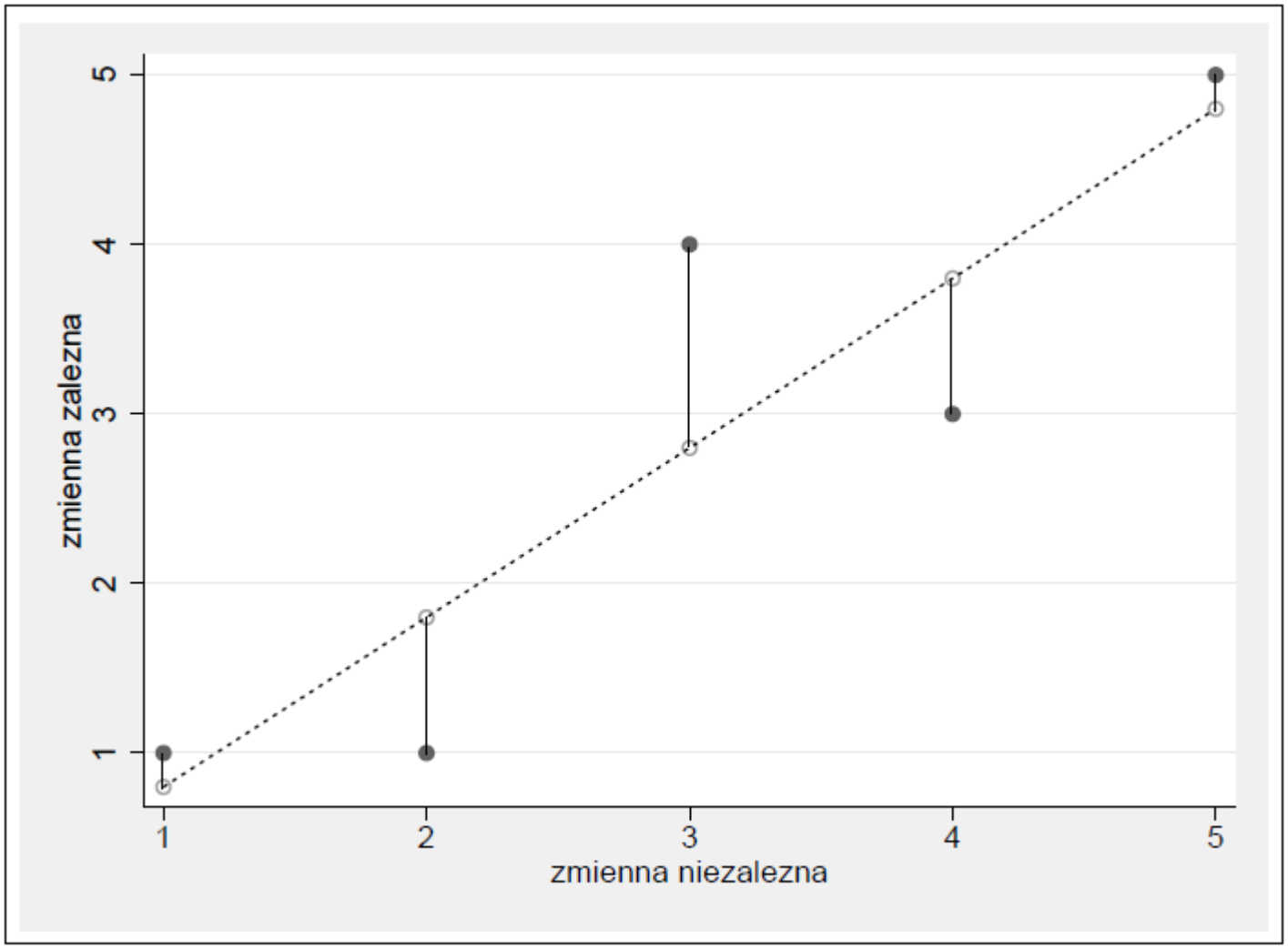
$$\hat{y}_1 = 463 + 0{,}08 \cdot 890{,}6 = 534{,}2$$

$$e_1 = 639{,}1 - 534{,}2 = 104{,}9$$

Reszty



Reszty



Metoda Najmniejszych Kwadratów

- Im mniejsza jest odległość wartości rzeczywistych od teoretycznych tym lepszy model ⟶
- estymatory parametrów modelu minimalizują sumę odległości y_i od $\hat{y}_i$:

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

Metoda Najmniejszych Kwadratów

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} e_i^2$$

- Funkcja ta jest **ciągła** i **różniczkowalna** dla wszystkich e_i, dzięki czemu można znaleźć jej minimum względem wielkości parametrów poprzez rozwiązanie standardowych warunków pierwszego rzędu.

Pytanie

- Jaką znasz inną funkcję odległości?
- Dlaczego trudno jest ją stosować w procesie estymacji?

Plan wykładu

- 1. Model liniowy
  - Postać modelu liniowego
  - Zapis macierzowy modelu liniowego
- 2. Estymacja modelu
  - Wstęp
  - Wartość teoretyczna (dopasowana)
  - Reszty
  - Metoda Najmniejszych Kwadratów
- 3. MNK – przypadek jednej zmiennej

Zadanie

- Zapisz model teoretyczny, wartości dopasowane oraz reszty dla modelu linowego zawierającego jedną zmienną objaśniającą i stałą

MNK dla modelu z jedną zmienną

- Model teoretyczny:

$$y_i = \beta_1 + x_{2i}\beta_2 + \varepsilon_i$$

- Wartość dopasowana (teoretyczna):

$$\hat{y}_i = b_1 + x_{2i}b_2$$

- Reszta:

$$e_i = y_i - \hat{y}_i = y_i - b_1 - x_{2i}b_2$$

MNK dla modelu z jedną zmienną

- Oszacowania b_1 *i* b_2 powinny być dobrane tak, by suma kwadratów reszt była jak najmniejsza.

$$S(b_1, b_2) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - b_1 - x_i b_2)^2 =$$

$$= \sum_{i=1}^{N} \left(y_i^2 - 2 y_i b_1 - 2 y_i b_2 x_i + 2 b_1 b_2 x_i + b_1^2 + b_2^2 x_i^2 \right)$$

MNK dla modelu z jedną zmienną

- Policz pochodne cząstkowe względem parametrów $\mathbf{b_1}$ i $\mathbf{b_2}$ powyższego równania i przyrównaj je do zera.

$$\begin{cases} \dfrac{\partial S(b_1, b_2)}{\partial b_1} = 0 \\ \dfrac{\partial S(b_1, b_2)}{\partial b_2} = 0 \end{cases}$$

Warunki pierwszego rzędu

MNK dla modelu z jedną zmienną

- Licząc pochodne dla poszczególnych równań uzyskujemy układ równań zwany **układem równań normalnych**.

$$\begin{cases} \sum_{i=1}^{N} \left[-2y_i + 2b_1 + 2b_2 x_i \right] = 0 \\ \sum_{i=1}^{N} \left[-2y_i x_i + 2b_1 x_i + 2b_2 x_i^2 \right] = 0 \end{cases}$$

MNK dla modelu z jedną zmienną

$$b_1 = \bar{y} - b_2 \bar{x}$$

$$b_2 = \frac{\frac{\sum_{i=1}^{N} y_i x_i}{N} - \overline{yx}}{\frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2}$$

MNK dla modelu z jedną zmienną

- Przypomnij wzór na wariancję (s^2_x) i kowariancję (s_{xy}) empiryczną.

$$b_1 = \bar{y} - b_2 \bar{x}$$

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

Przykład

- Estymacja modelu wyjaśniającym wielkość wydatków na żywność dochodami gospodarstwa

$$wydatki_i = \beta_1 + \beta_2 dochód_i + \varepsilon_i$$

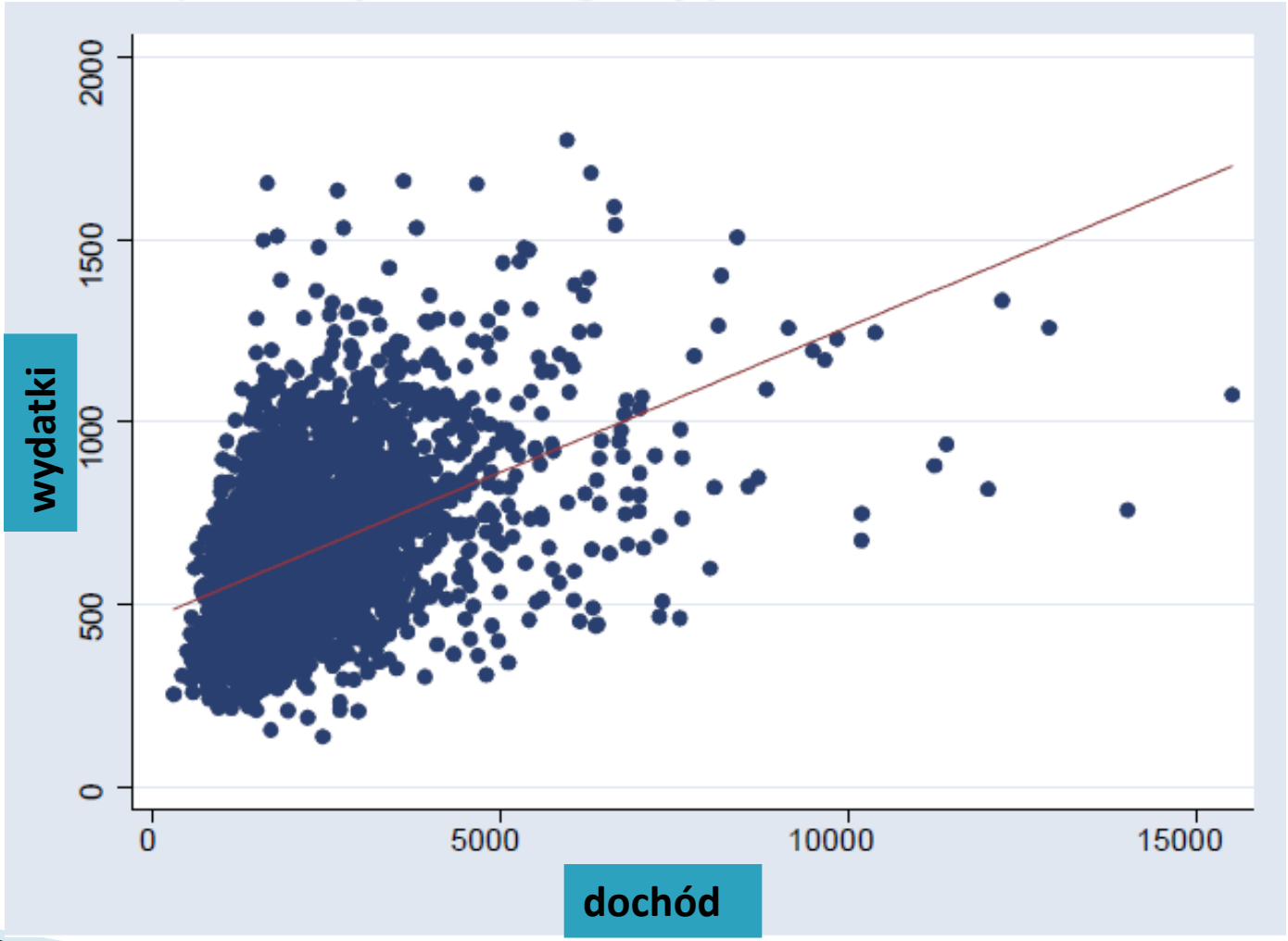
Zmienna	Średnia	Wariancja
Wydatki	644,18	46737
Dochód	2262,34	1584300

- Kowariancja empiryczna między zmiennymi jest równa 126211.

$$b_1 = \frac{126211}{1584300} = 0{,}079664$$

$$b_2 = 644{,}18 - 0{,}079664 \cdot 2262{,}34 = 463{,}95$$

Zależność wydatków na żywność od dochodu (dane i prosta regresji)



Pytania teoretyczne

1. Zapisać model liniowy liniowy. Podać interpretację poszczególnych elementów tego modelu.
2. Podać wzajemne relacje między wartościami obserwowanymi zmiennej zależnej, oszacowaniami parametrów, wartościami dopasowanymi i resztami.
3. Wyjaśnij różnicę między parametrami i oszacowaniami parametrów oraz między odchyleniami losowymi i resztami.
4. Skąd bierze się nazwa Metoda Najmniejszych Kwadratów?
5. Wyprowadzić estymator MNK dla modelu ze stałą i jedną zmienną objaśniającą.

Dziękuję za uwagę