Przybliżanie modeli nieliniowych cd.
Klasyczny Model Regresji Liniowej

Stanisław Cichocki

Natalia Nehrebecka

Wykład 9

Plan wykładu

- 1. Przybliżanie modeli nieliniowych:
  - Model schodkowy
  - Model krzywej łamanej
- 2. Założenia KMRL
- 3. Własności estymatora MNK w KMRL
  - Twierdzenie Gaussa-Markowa
- 4. Estymator wariancji błędu losowego

Plan wykładu

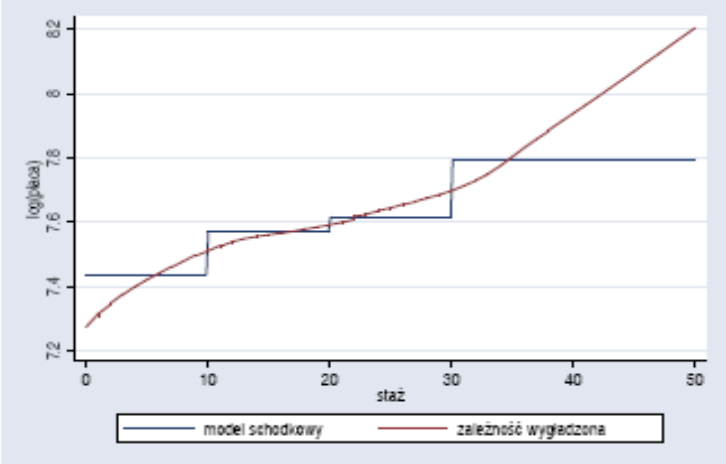
- 1. Przybliżanie modeli nieliniowych:
  - Model schodkowy
  - Model krzywej łamanej
- 2. Założenia KMRL
- 3. Własności estymatora MNK w KMRL
  - Twierdzenie Gaussa-Markowa
- 4. Estymator wariancji błędu losowego

Model schodkowy

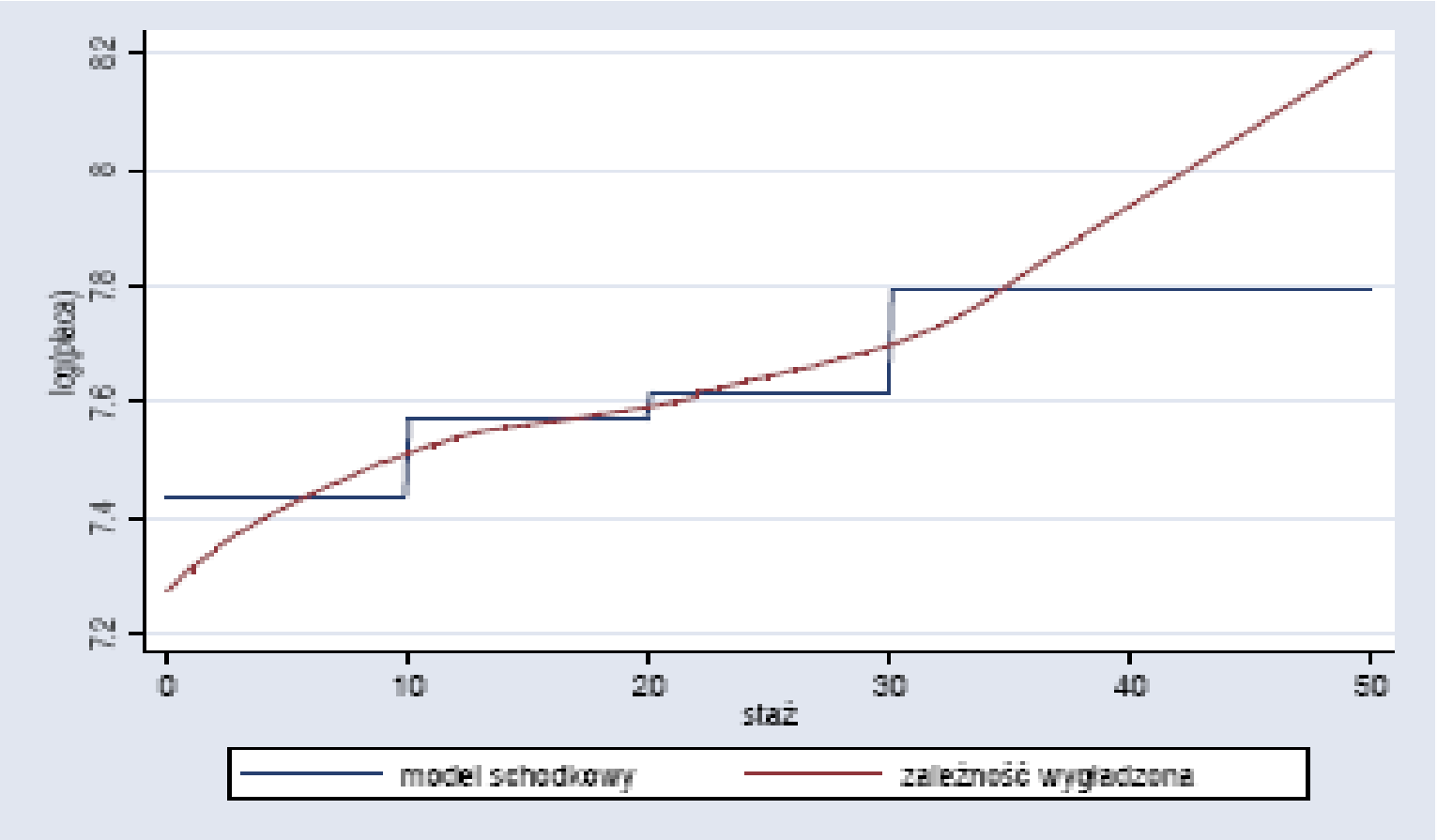
- Nieliniowa zależność między ***y*** a ***x*** można przybliżyć za pomocą modelu liniowego stosując model:
- **2. Model schodkowy**
- W tym przypadku definiujemy **zmienne zerojedynkowe**

$$D_{s,i} = \mathbb{I}(x_s^* < x_i \leq x_{s+1}^*)$$

- związane z przedziałami x_i i
- przeprowadzamy regresję na tych zmiennych zamiast na x_i. Wyestymowany model można zilustrować rysunkiem:



Model schodkowy



Model schodkowy

```
generate wiek_2 = (wiek > 25 & wiek <= 35)
generate wiek_3 = (wiek > 35 & wiek <= 45)
generate wiek_4 = (wiek > 45 & wiek <= 55)
generate wiek_5 = (wiek > 55)
```

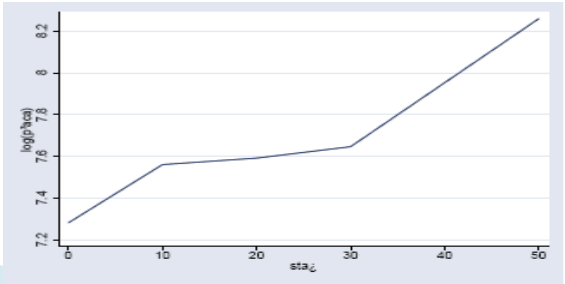
```
regress dochod wiek_?
      Source |       SS       df       MS              Number of obs =    1083
-------------+------------------------------           F(  4,  1078) =    2.33
       Model |  6403953.56     4  1600988.39            Prob > F      =  0.0544
    Residual |   741077182  1078  687455.642            R-squared     =  0.0086
-------------+------------------------------           Adj R-squared =  0.0049
       Total |   747481135  1082  690832.842            Root MSE      =  829.13

------------------------------------------------------------------------------
      dochod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      wiek_2 |   126.6784   88.27104     1.44   0.152    -46.52407     299.881
      wiek_3 |   239.7376   84.81751     2.83   0.005     73.31151    406.1637
      wiek_4 |    206.697   91.38316     2.26   0.024       27.388     386.006
      wiek_5 |   175.5193   141.5618     1.24   0.215    -102.2486    453.2873
       _cons |   639.0551   73.57334     8.69   0.000     494.6919    783.4183
------------------------------------------------------------------------------
```

Model krzywej łamanej

- Nieliniowa zależność między ***y*** a ***x*** można przybliżyć za pomocą modelu liniowego stosując model:
- **3. Model krzywej łamanej**

$$y = \begin{cases} \alpha + \beta_1 x + \varepsilon & dla \quad x_i \le x_1^* \\ \alpha + \beta_1 x + \beta_2 (x - x_1^*) + \varepsilon & dla \quad x_1^* < x_i \le x_2^* \\ \vdots & \\ \alpha + \beta_1 x + \sum_{j=2}^{s-1} \beta_j (x_j^* - x_{j-1}^*) + \beta_s (x - x_s^*) + \varepsilon & dla \quad x_i > x_s^* \end{cases}$$

- **Zależność nieliniowa przybliżona jest w tym przypadku krzywą, którą można zilustrować rysunkiem:**

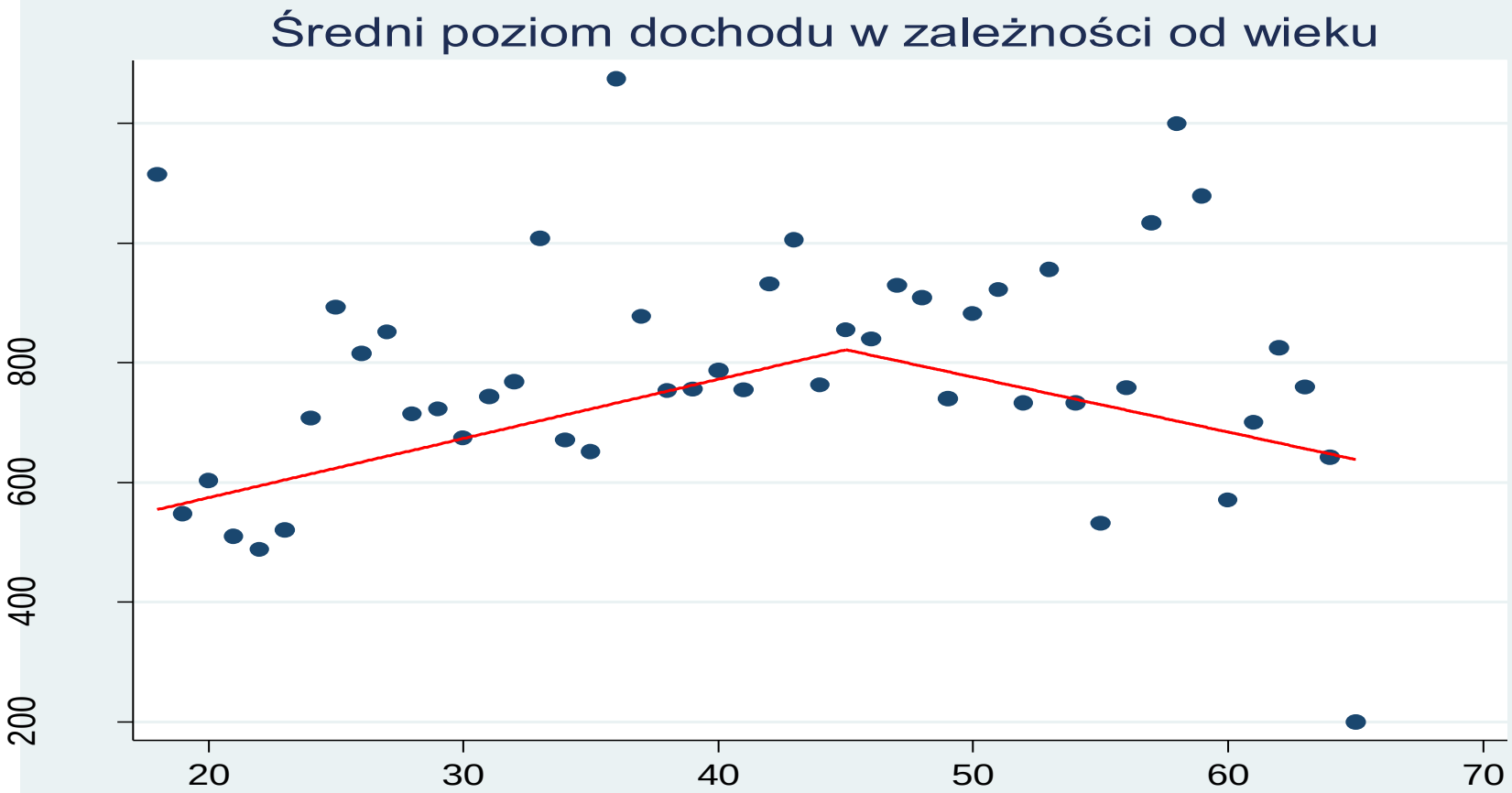


Model krzywej łamanej - DWIE LINIOWE FUNKCJE SKLEJONE W 45

```
regress dochod wiek wiek_45 plec srednie wyzsze
```

```
      Source |       SS       df       MS              Number of obs =    1083
-------------+------------------------------           F(  5,  1077) =   22.92
       Model |  71889880.6     5  14377976.1           Prob > F      =  0.0000
    Residual |   675591255  1077   627289.93           R-squared     =  0.0962
-------------+------------------------------           Adj R-squared =  0.0920
       Total |   747481135  1082  690832.842           Root MSE      =  792.02

------------------------------------------------------------------------------
      dochod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |    9.892845   3.449602     2.87   0.004     3.124143    16.66155
     wiek_45 |   -19.06609   9.716528    -1.96   0.050    -38.13156   -.0006177
        plec |   -338.9919   48.27437    -7.02   0.000    -433.7144   -244.2694
     srednie |    211.058    77.6635     2.72   0.007     58.66912     363.447
      wyzsze |    712.6863   99.4661     7.17   0.000      517.517    907.8556
       _cons |    376.4752   145.4995     2.59   0.010     90.98058    661.9698
------------------------------------------------------------------------------
```

Model krzywej łamanej



Plan wykładu

- 1. Przybliżanie modeli nieliniowych:
  - Model schodkowy
  - Model krzywej łamanej
- 2. Założenia KMRL
- 3. Własności estymatora MNK w KMRL
  - Twierdzenie Gaussa-Markowa
- 4. Estymator wariancji błędu losowego

Klasyczny model regresji liniowej

- Na poprzednich wykładach pokazaliśmy, iż estymator MNK daje oszacowania parametrów, które są najlepiej dopasowane do danych
- Obecnie zajmiemy się własnościami statystycznymi tego estymatora i w tym celu przyjmujemy pewne dodatkowe założenia
- Najprostszym i najpopularniejszym układem założeń jest KMRL

Założenia klasycznego modelu regresji liniowej

- 1. Związek pomiędzy zmienną zależną a zmiennymi niezależnymi opisany jest równaniem:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \qquad i = 1,2,3 \cdots n$$

- 2. Zmienne objaśniające $x_{2i}, x_{3i}, \ldots, x_{Ki}$ są nielosowe dla $i = 1,2,3 \cdots n$
- 3. Wartość oczekiwana błędu losowego jest równa zeru:

$$E(\varepsilon) = \mathbf{0}$$

- 4. Zaburzenia losowe ε są **sferyczne.** Oznacza to, że warunkowa macierz wariancji-kowariancji wektora zaburzeń przy danej macierzy X ma postać:

$$Var(\varepsilon) = \sigma^2 \mathbf{I}$$

- gdzie **I** oznacza macierz jednostkową.

Założenia klasycznego modelu regresji liniowej

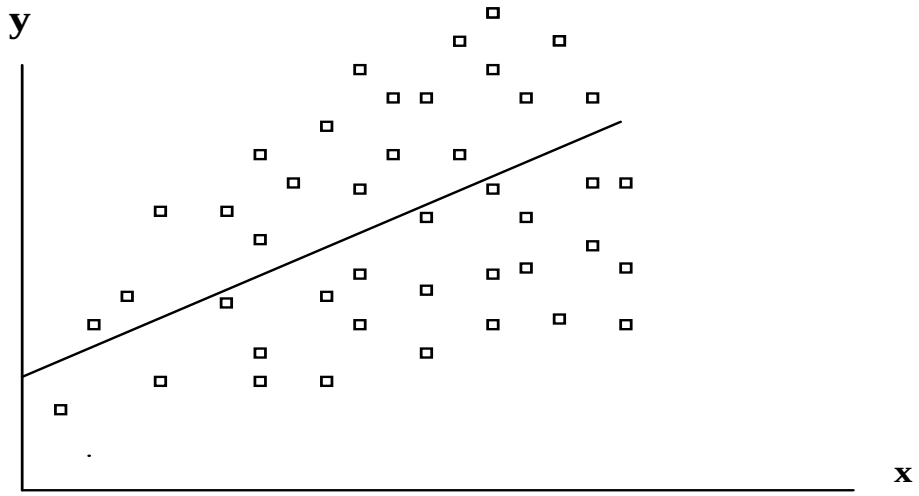
$$Var(\varepsilon) = \sigma^2 \mathbf{I}$$

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_2, \varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & Cov(\varepsilon_n, \varepsilon_1) & \cdots & Var(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

- Założenie **sferyczności zaburzeń** oznacza:
- po pierwsze, że **wariacje** kolejnych zaburzeń (elementy na diagonalnej) są takie same dla wszystkich obserwacji i równe σ^2, gdzie σ^2 jest nieznaną dodatnią stałą;
- po drugie, że elementy pozadiagonalne, które są **kowariancjami** zaburzeń dla różnych obserwacji są równe zero, a więc zaburzenia dla różnych obserwacji są ze sobą nieskorelowane.

Założenia klasycznego modelu regresji liniowej

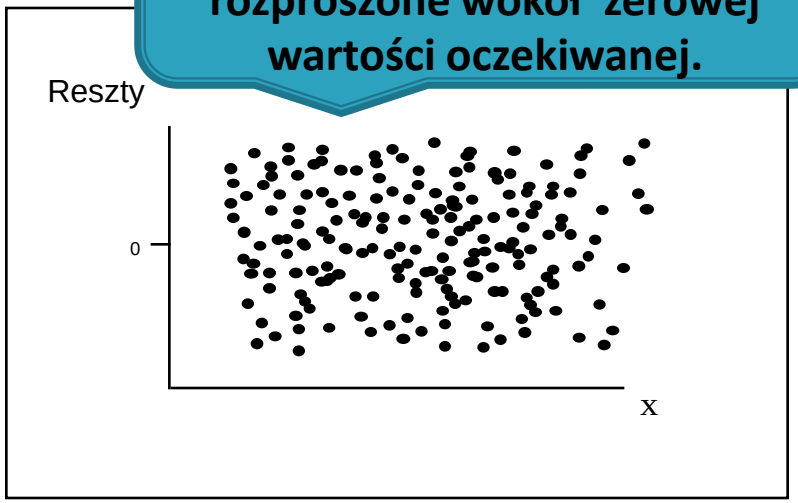
- Stałość wariancji zaburzeń nazywamy homoskedastycznością zaburzeń. **Oznacza to, że zaburzenia losowe są jednakowo rozproszone wokół zerowej wartości oczekiwanej.** Jeśli wariancje nie byłyby jednakowe, to sytuację taką nazywamy heteroskedastycznością.

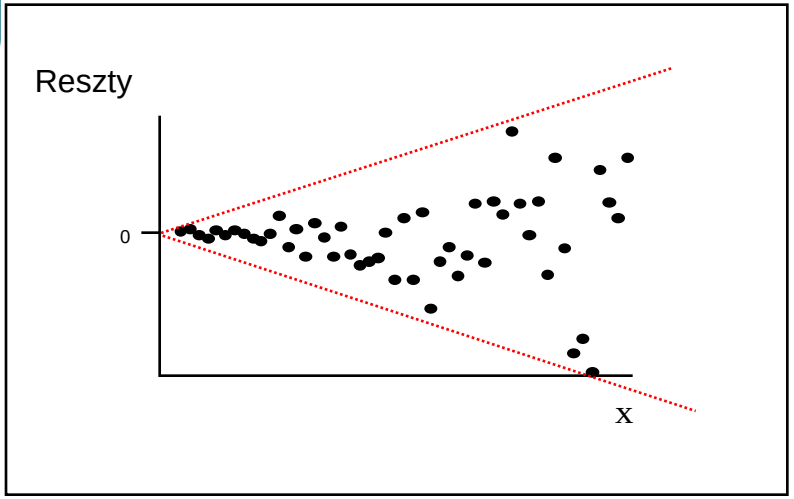


Rys.1. **Heteroskedastyczność**

Założenia klasycznego modelu regresji liniowej

Oznacza to, że zaburzenia losowe są jednakowo rozproszone wokół zerowej wartości oczekiwanej.



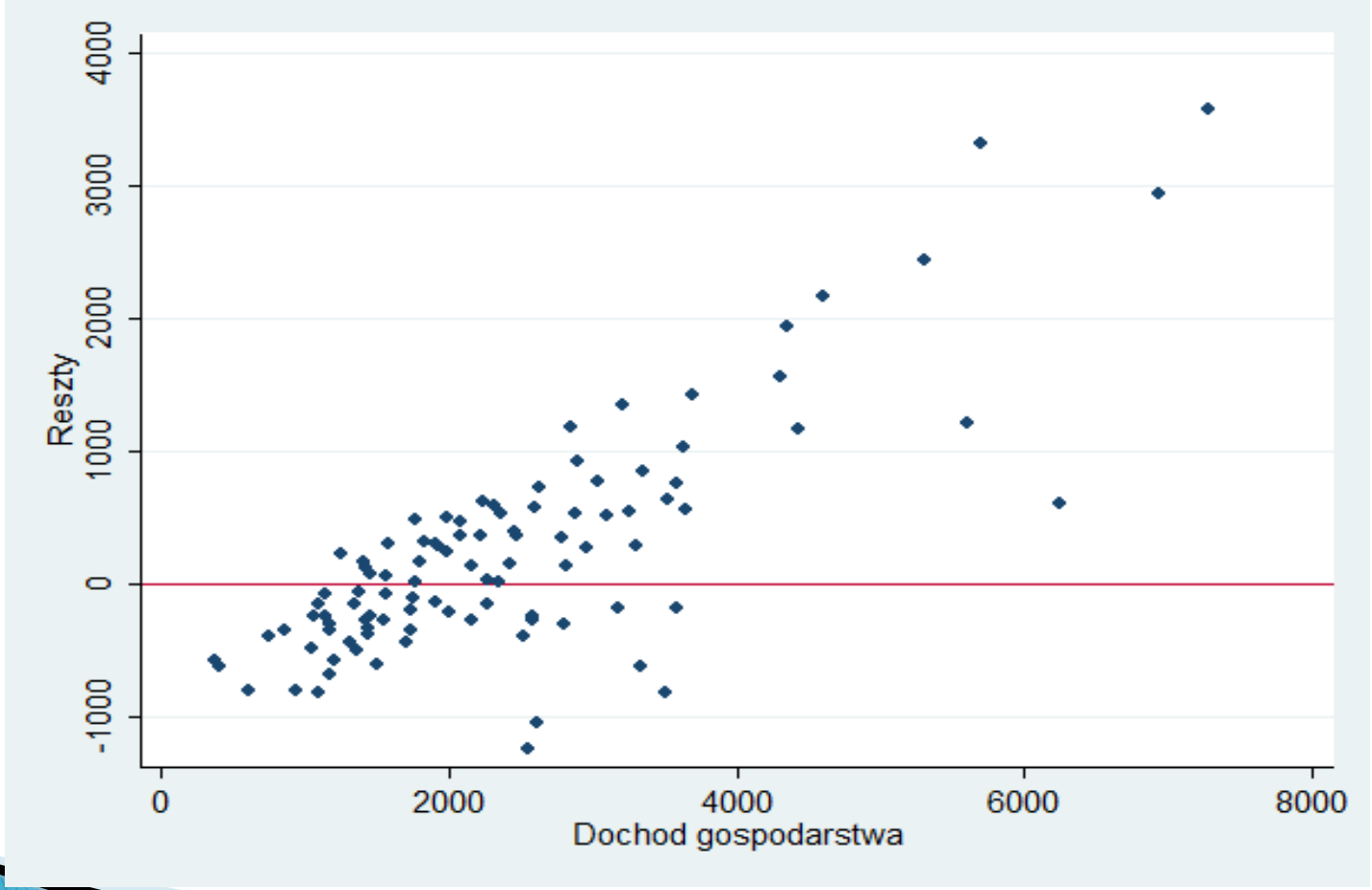


Homoskedastyczność: reszty zachowują się losowo.

Heteroskedastyczność: Wariancja reszt zmienia się wraz ze zmianą zmiennej niezależnej X.

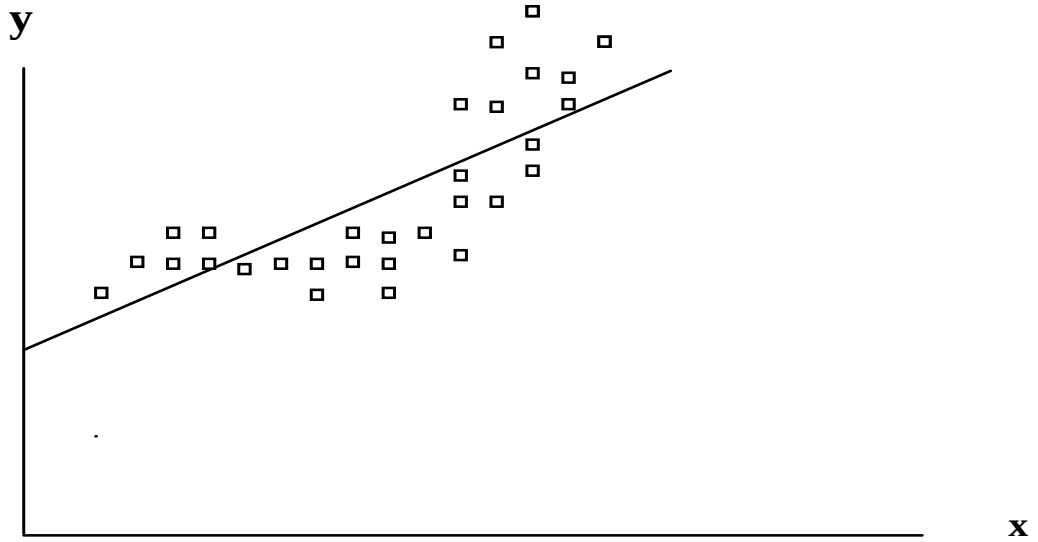
Założenia klasycznego modelu regresji liniowej

- Regresja wydatków na dochodzie



Założenia klasycznego modelu regresji liniowej

- Przypadek zerowych kowariancji dla różnych zaburzeń losowych ε_i oraz ε_j nazywamy brakiem autokorelacji zaburzeń. Oznacza to, że **zaburzenia losowe dla różnych obserwacji są niezależne**, a przez to nieskorelowane, a więc nie mają tendencji do gromadzenia się np. wokół dodatnich lub ujemnych (lub naprzemiennie dodatnich i ujemnych) wartości



Rys. 2. **Autokorelacja**

Plan wykładu

- 1. Przybliżanie modeli nieliniowych:
  - Model schodkowy
  - Model krzywej łamanej
- 2. Założenia KMRL
- 3. Własności estymatora MNK w KMRL
  - Twierdzenie Gaussa-Markowa
- 4. Estymator wariancji błędu losowego

Twierdzenie Gaussa-Markowa

W klasycznym modelu regresji liniowej najlepszym liniowym i nieobciążonym estymatorem wektora parametrów β jest b wyznaczone za pomocą MNK

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

o macierzy wariancji-kowariancji

$$Var(b) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$Var(b) = \begin{bmatrix} Var(b_1) & Cov(b_2,b_1) & \cdots & Cov(b_n,b_1) \\ Cov(b_1,b_2) & Var(b_2) & \cdots & Cov(b_n,b_2) \\ \vdots & \vdots & & \vdots \\ Cov(b_1,b_n) & Cov(b_2,b_n) & \cdots & Var(b_n) \end{bmatrix}$$

Twierdzenie Gaussa-Markowa

- 1. Estymator **b** jest estymatorem **liniowym**, gdyż jest liniową funkcją zmiennej losowej **y.**
- 2. **b** jest estymatorem **nieobciążonym**, to znaczy $E(\mathbf{b}) = \beta$.

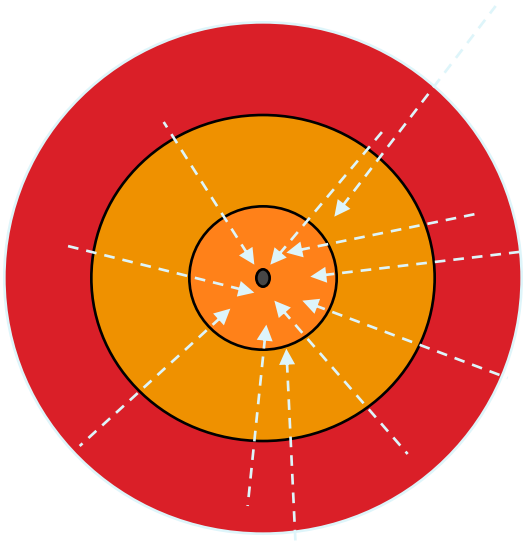
  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- i podstawiając za **y** $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
- otrzymamy:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

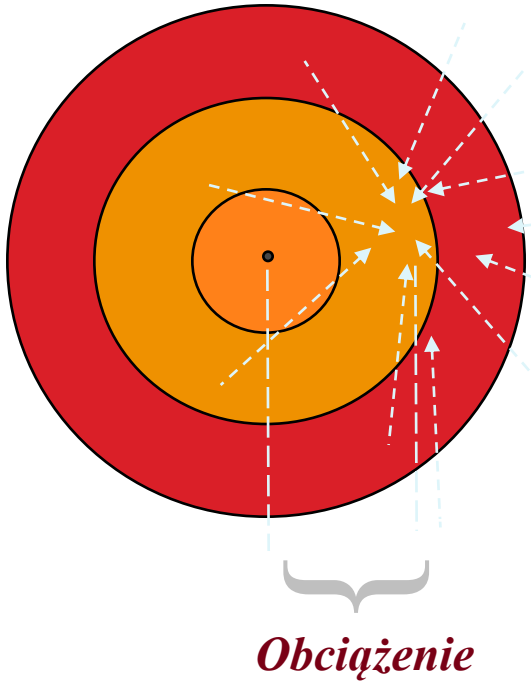
$$E(\mathbf{b}) = \beta + E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\right) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0 = \beta$$

- 3. Estymator **b** jest estymatorem **najlepszym** w tym sensie, że każdy inny estymator liniowy i nieobciążony ma macierz wariancji-kowariancji większą od tej dla **b**. Estymator taki nazywamy estymatorem **efektywnym**.

Obciążony i nieobciążony estymator



Estymator nieobciążony



Estymator obciążony

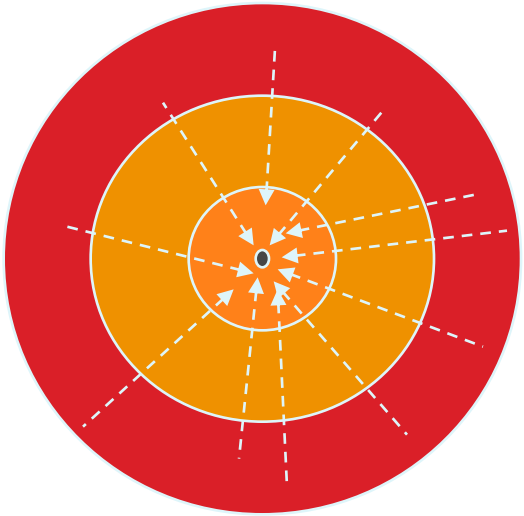
Twierdzenie Gaussa-Markowa

- Wariancja estymatora **b**

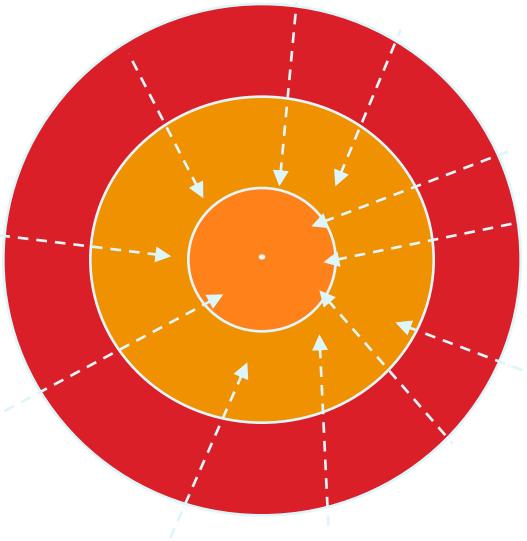
$$Var(b) = Var(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\varepsilon)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$
$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sum$$

Efektywność

Estymator jest efektywnym, jeśli ma najniższą wariancję i odchylenie standardowe.



Estymator efektywny



Estymator nieefektywny

Plan wykładu

- 1. Przybliżanie modeli nieliniowych:
  - Model schodkowy
  - Model krzywej łamanej
- 2. Założenia KMRL
- 3. Własności estymatora MNK w KMRL
  - Twierdzenie Gaussa-Markowa
- 4. Estymator wariancji błędu losowego

Dziękuję za uwagę