Problemy z danymi
Heteroskedastyczność i autokorelacja

Stanisław Cichocki

Natalia Nehrebecka

Wykład 13

Plan wykładu

- 1. Problemy z danymi
  - Obserwacje nietypowe i błędne
  - Współliniowość
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji

Plan wykładu

- 1. Problemy z danymi
  - Obserwacje nietypowe i błędne
  - Współliniowość
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji

Obserwacje nietypowe i błędne

- **Obserwacja nietypowa** charakteryzuje się nietypowymi na tle pozostałych obserwacji cechami
- Mechanizm, który w przypadku tej zmiennej generuje zmienną zależną jest mechanizmem opisywanym przez model
- **Obserwacja błędna** jest obserwacją, której powstania nie da się wytłumaczyć w ramach teoretycznego modelu ekonomicznego stanowiącego podstawę estymowanego modelu
- Obserwacje błędne często pojawiają się w wyniku pomyłek przy wpisywaniu obserwacji do bazy danych

Obserwacje nietypowe i błędne

- Niekiedy jednak obserwacje błędne są rzeczywistymi obserwacjami, związanymi z pewnymi nietypowymi zdarzeniami, które nie mogą być wyjaśnione za pomocą naszego modelu

- Przykład:

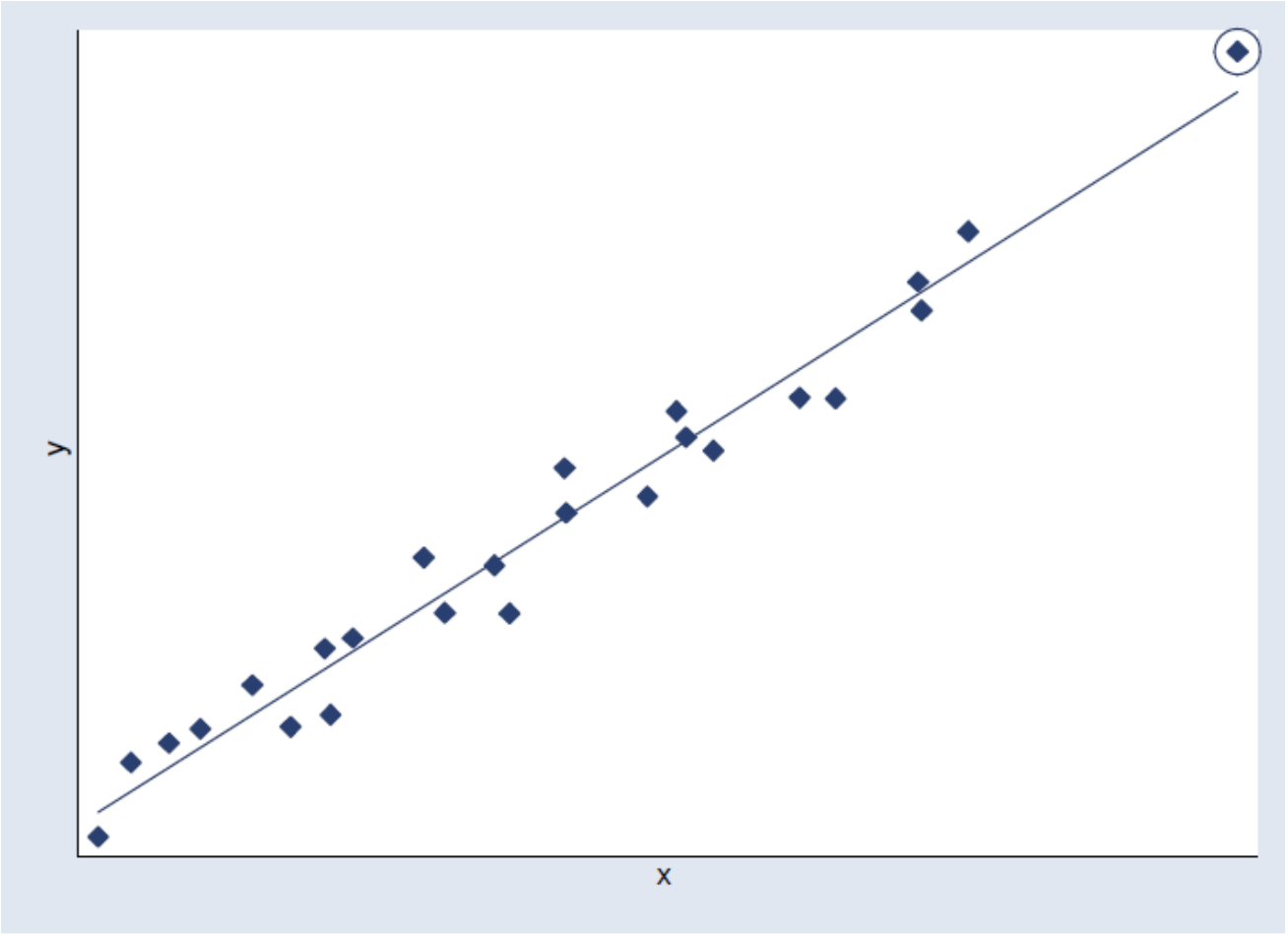
Estymujemy krzywą popytu na żywność dla różnych państw na świecie. W próbie występują państwa, w których obowiązuje reglamentacja żywności. Obserwacje takie traktujemy jako obserwacje błędne – teoria opisująca krzywą popytu nie znajduje zastosowania w momencie nierynkowego podziału dóbr.

Obserwacje nietypowe i błędne

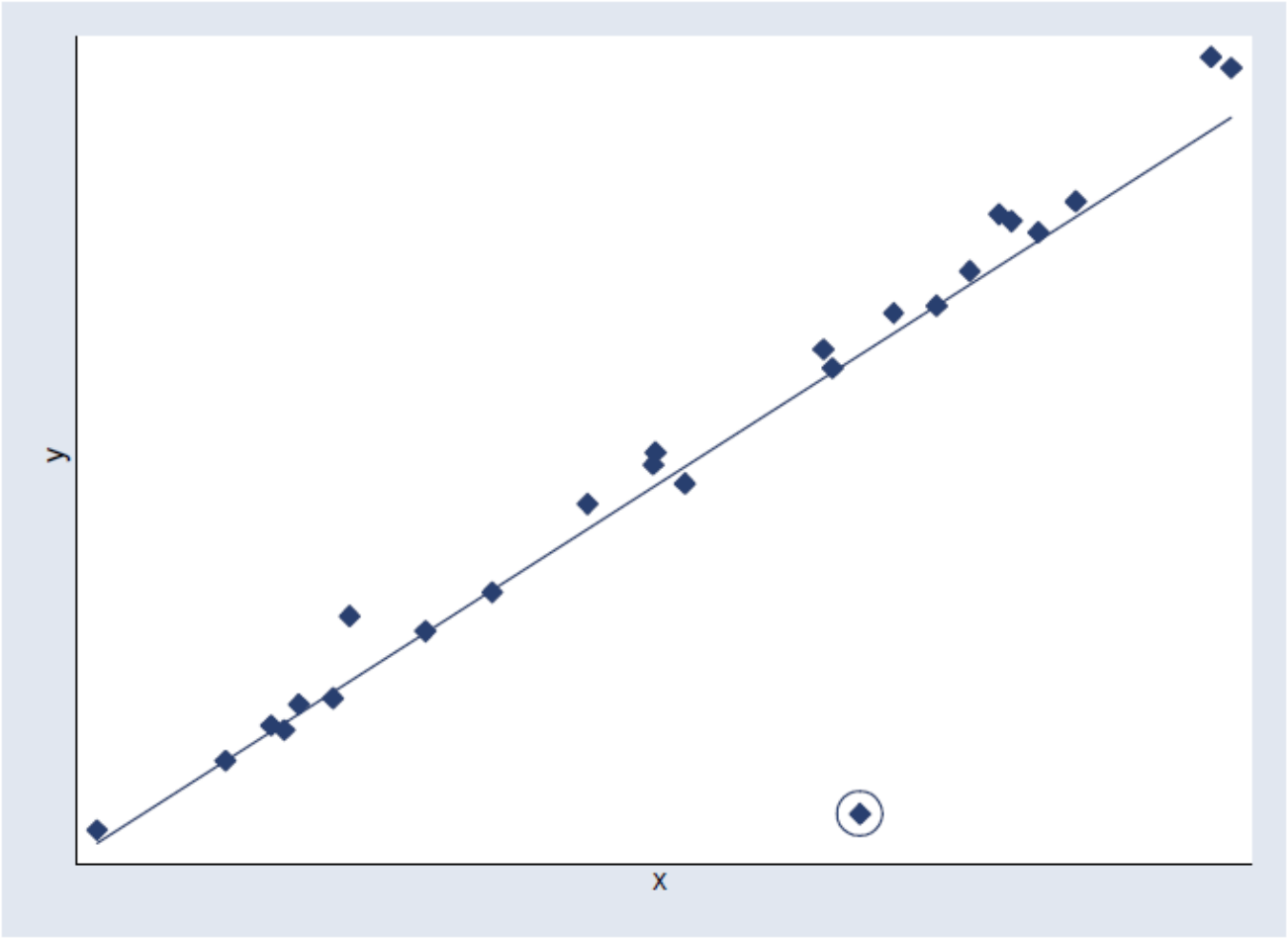
- Wpływ obserwacji nietypowej/błędnej na wynik regresji zależy od tego na ile ta obserwacja pasuje do prostej regresji

- Najbardziej niepokojąca jest sytuacja gdy obserwacja ma nietypowe wartości dla zmiennych niezależnych i słabo pasuje do prostej regresji

Obserwacje nietypowe i błędne



Obserwacje nietypowe i błędne



Obserwacje nietypowe i błędne

- Na podstawie samego modelu nie da się ustalić, które obserwacje są błędne ⟶ fakt, że obserwacja nie pasuje do modelu nie może być powodem do jej usunięcia ⟶ tak postępując zawsze udawałoby się nam uzyskać dobrze dopasowany model (usuwając obserwacje, które nie pasują do modelu)

- Część obserwacji możemy uznać za błędne na podstawie teorii
  np. zmienna wiek przyjmuje dla pewnych obserwacji wartości ujemne ⟶ wiemy, że wiek musi przyjmować wartości dodatnie więc obserwacja błędna

Obserwacje nietypowe i błędne

- Przykład:

  Badamy wynagrodzenia dla próby osób przebadanych w 2007 przez CASE pod kątem wykonywania pracy nierejestrowanej.

```
sum wynagrodzenia

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     zarobki |       5773    13392.31    32264.34         0      99997


  count if wynagrodzenia==99997
  703
```

Obserwacje nietypowe i błędne

- Uwzględnienie obserwacji nietypowej pozytywnie wpływa na:
  a) precyzję oszacowań
  b) dopasowanie modelu

- Uwzględnienie obserwacji błędnej negatywnie wpływa na:
  a) precyzję oszacowań
  b) dopasowanie modelu

Obserwacje nietypowe i błędne

Przykład:

Porównujemy rentownośc dwóch kontraktów: A i B. Dysponujemy 10 obserwacjami dotyczącymi stóp zwrotu (IRR – internal rate of return) dla tych dwóch kontraktów

kontrakt	stopa zwrotu									
A	10	8	8	9	11	10	8	9	11	10
B	16	15	18	17	16	-80	17	16	16	17

Obserwacje nietypowe i błędne

Regresja z pominięciem jednej obserwacji:

IRR	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_IB_1	7.155556	.4808912	14.88	0.000	6.140964	8.170147
_cons	9.4	.330972	28.40	0.000	8.70171	10.09829

Obserwacje nietypowe i błędne

Regresja ze wszystkimi obserwacjami:

IRR	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_IB_1	-3.5	10.66526	-0.33	0.747	-25.90688	18.90688
_cons	9.4	7.541478	1.25	0.229	-6.444057	25.24406

Obserwacje nietypowe i błędne

- Statystki służące do wykrycia obserwacji nietypowych, słabo pasujących do prostej regresji, silnie wpływających na wynik regresji:

  a) dźwignia

  b) standaryzowane reszty

  c) odległość Cooka'a

Obserwacje nietypowe i błędne

- Dźwignia – używana do stwierdzenie czy wektor zmiennych niezależnych x_i dla obserwacji i jest nietypowy na tle pozostałych x:

$$h_i = \delta_i' X (X'X)^{-1} X' \delta_i = \delta_i' P_X \delta_i = (P_X)_{ii}$$

$$= x_i (X'X)^{-1} x'_i$$

gdzie:

$$\delta_i = [0,\ldots,0,\ 1,\ 0,\ldots,0]' \qquad P_X = X(X'X)^{-1}X'$$

Obserwacje nietypowe i błędne

- Dla każdego modelu:

$$0 \leq h_i \leq 1$$

- Dla modelu ze stałą:

$$\frac{1}{N} \leq h_i \leq 1$$

Obserwacje nietypowe i błędne

- Nieformalna reguła mówi, ze obserwacje można traktować jako nietypową gdy:

$$h_i \geq \frac{2K}{N}$$

-To, że obserwacja jest nietypowa nie oznacza, że nie pasuje do modelu

- Aby się o tym przekonać musimy przyjrzeć się **standaryzowanym resztom**

Obserwacje nietypowe i błędne

- Standaryzowane reszty:
- Przypomnienie: $e = M_x \varepsilon$
- Wobec tego:

$$Var(e) = Var(M_x \varepsilon) = M_x (I\sigma^2) M_x = \sigma^2 M_x$$

Obserwacje nietypowe i błędne

- Standaryzowane reszty:

$$Var(e_i) = Var(\delta'_i e) = \sigma^2 \delta'_i M_x \delta_i$$
$$= \sigma^2 [\delta'_i \delta_i - \delta'_i (X'X)^{-1} X' \delta_i]$$
$$= \sigma^2 (1 - \delta'_i P_X \delta_i) = \sigma^2 (1 - h_i)$$

Obserwacje nietypowe i błędne

- Standaryzowane reszty:

- Jeśli $\varepsilon \sim N(0,\sigma^2 I)$ to:

$$\tilde{e}_i = \frac{e_i}{\sigma\sqrt{1-h_i}} \sim N(0,1)$$

- Ponieważ σ jest nieznane stosujemy estymator **s:**

$$\hat{e}_i = \frac{e_i}{s\sqrt{1-h_i}} \sim t_{N-K}$$

Obserwacje nietypowe i błędne

- Dla nietypowej obserwacji:

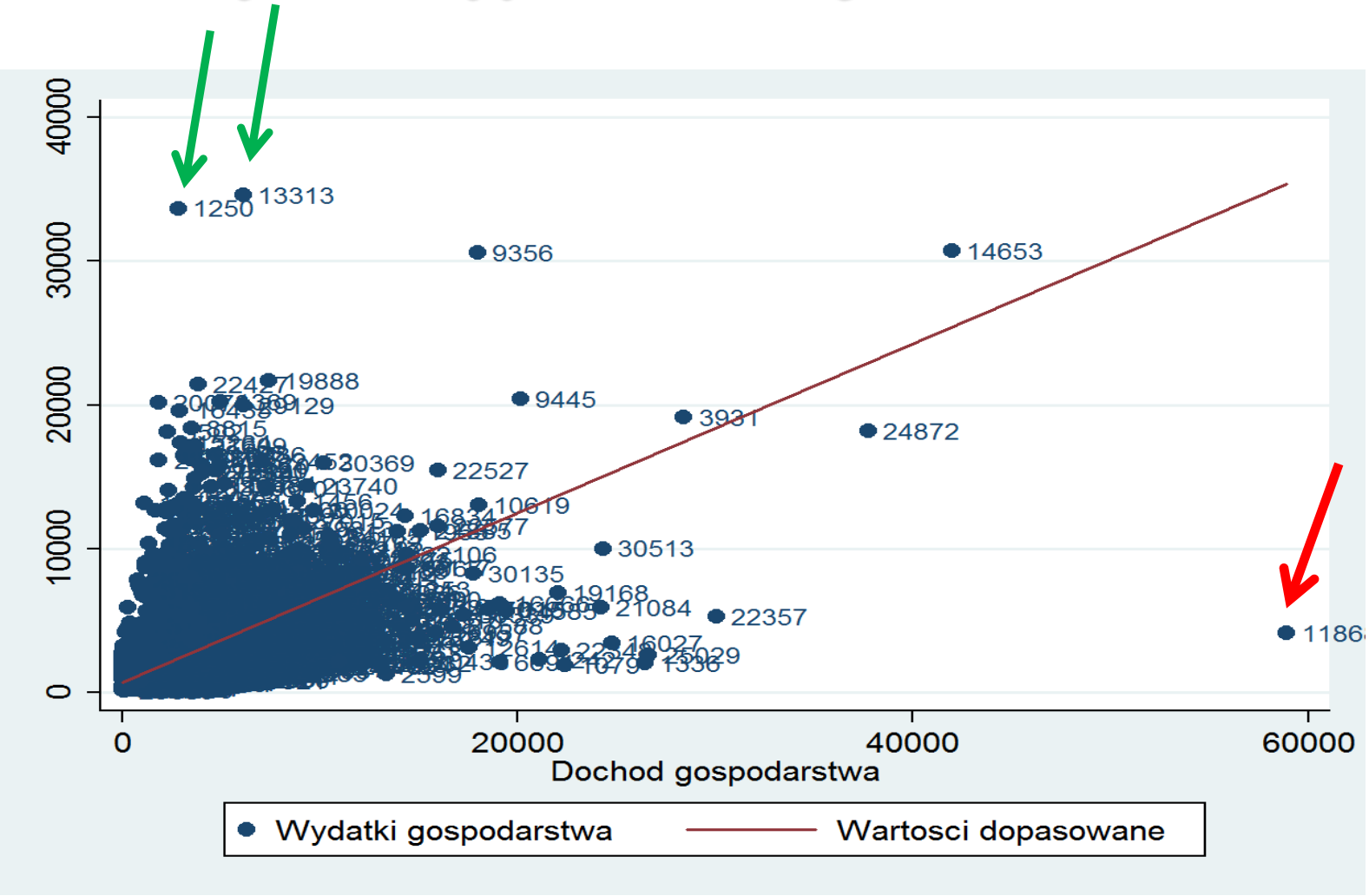
$$\left|\hat{e}_i\right| > 2$$

- Jednak (jeżeli błąd losowy ma rozkład normalny), to statystycznie dla ok. 5% obserwacji:

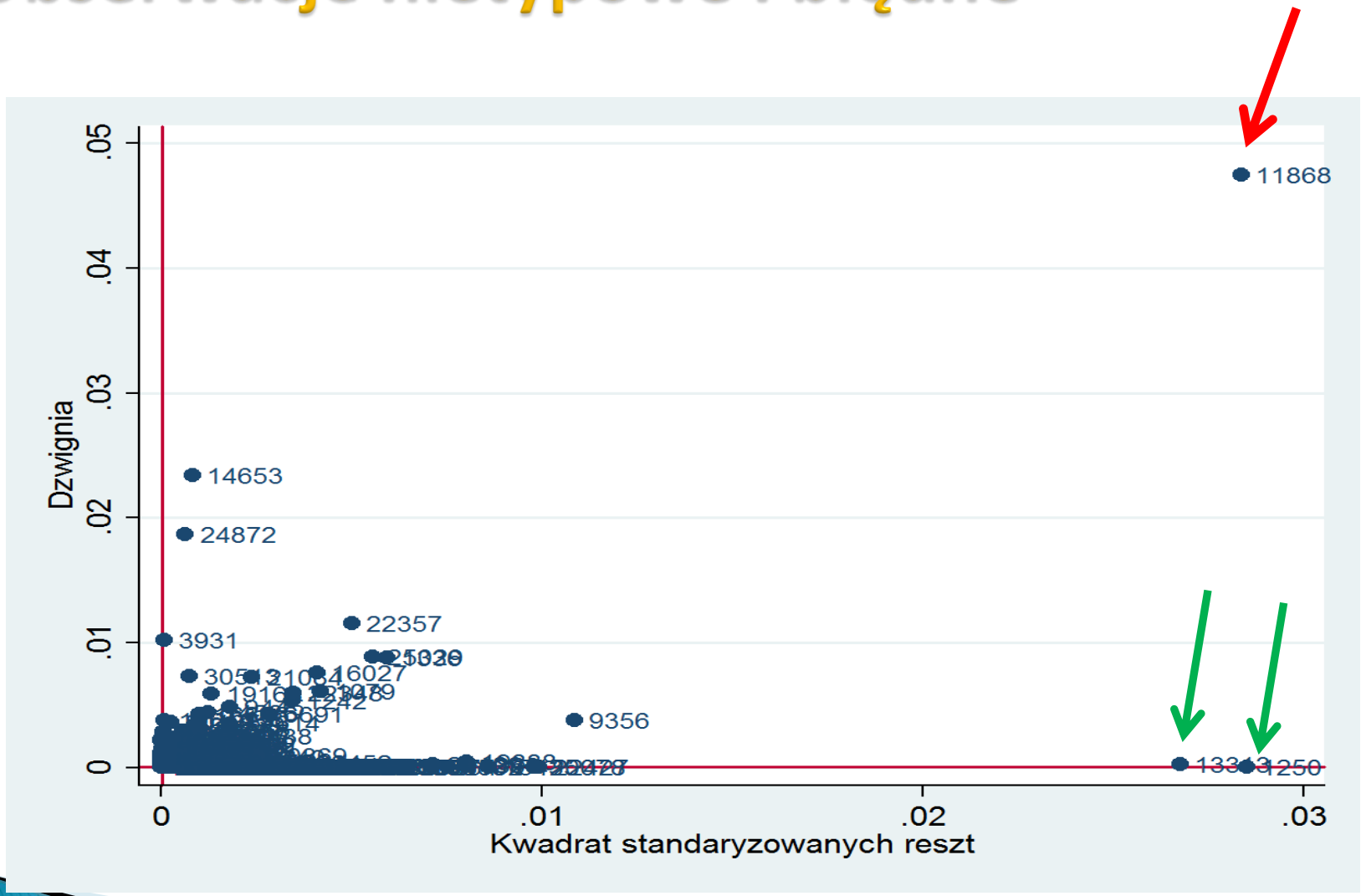
$$\left|\hat{e}_i\right| > 2$$

- Niepokojące jest nie tyle fakt występowania dużych reszt, ile raczej występowanie dużych wartości reszt dla obserwacji nietypowych (o dużych dźwigniach)

Obserwacje nietypowe i błędne



Obserwacje nietypowe i błędne



Obserwacje nietypowe i błędne

- Odległość Cook'a ⟶ mierzy wpływ pojedynczej obserwacji na wynik regresji:

$$CD_i = \frac{(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})}{Ks^2} = \frac{\hat{e}_i^2}{K} \frac{h_i}{1 - h_i}$$

gdzie:

$\hat{y}_{(i)} = X_{(i)} b_{(i)}$ - wartości dopasowne powstałe po usunięciu z próby i – tej obserwacji

$\hat{y} = X_{(i)} b$

Obserwacje nietypowe i błędne

- Odległość Cook'a:

- Najbardziej wpływowe są obserwacje, która maja równocześnie duże $\hat{e}_i^2$ i h_i

- Nieformalna zasada mówi, ze powinniśmy uważnie przyjrzeć się obserwacjom, dla których:

$$CD_i > \frac{4}{N}$$

Obserwacje nietypowe i błędne

	numer	dochg	wydg	reszty_st	dzwignia	cook_d~t
1.	11868	58935	4132	-30.72398	.0474962	23.53513
2.	1336	26453	2008	-13.74937	.0087709	.8363862
3.	25029	26645	2563	-13.32397	.0089089	.7979006
4.	22357	30069	5267	-12.67469	.0115515	.9387016
5.	1079	22392	1892	-11.54321	.0061053	.4092522

$$h_i \geq \frac{2K}{N} = \frac{2*2}{31679} \approx 0,00012$$

$$CD_i > \frac{4}{N} = \frac{4}{31679} \approx 0,00012$$

Obserwacje nietypowe i błędne

	numer	dochg	wydg	cook_d~t	reszty_st	dzwignia
1.	11868	58935	4132	23.53513	-30.72398	.0474962
2.	1336	26453	2008	.8363862	-13.74937	.0087709
3.	25029	26645	2563	.7979006	-13.32397	.0089089
4.	22357	30069	5267	.9387016	-12.67469	.0115515
5.	1079	22392	1892	.4092522	-11.54321	.0061053

Współliniowość

- O współliniowości mówimy w przypadku występowania silnej korelacji między zmiennymi objaśniającymi ⟶ utrudnia to zidentyfikowanie zmiennej, która jest przyczyną zmiennej zależnej

- Wyróżniamy dwa typy współliniowości:

  a) dokładną współliniowość

  b) niedokładną współliniowość

Współliniowość

- O dokładnej współliniowości mówimy, gdy kolumny macierzy obserwacji są współliniowe ⟶ jedna z kolumn macierzy jest kombinacją liniową pozostałych kolumn ⟶ macierz X'X jest osobliwa i wobec tego nieodwracalna

- Oznacza to, że jedna ze zmiennych niezależnych jest kombinacją liniową pozostałych zmiennych niezależnych i nie wnosi żadnej dodatkowej informacji do modelu ⟶ powinniśmy usunąć ja z modelu

- Dokładna współliniowość jest wynikiem błędnej specyfikacji modelu

Obserwacje nietypowe i błędne

Przykład:

zmienne objaśniające w modelu:

a) ln(PKB),

b) ln(Liczba ludności)

c) ln(PKB per capita)

- Zmienna ln(PKB per capita) jest kombinacją zmiennej ln(PKB) i ln(Liczba ludności)

Współliniowość

- O niedokładnej współliniowości mówimy, gdy występuje silna korelacja między zmiennymi objaśniającymi

- W przypadku danych ekonometrycznych występowanie korelacji między zmiennymi objaśniającymi jest regułą ⟶ problemem jest nie samo występowanie korelacji lecz przypadek gdy jest ona bardzo silna ⟶ obniża to precyzję oszacowań

Współliniowość

- Statystyka służąca do wykrywania niedokładnej współliniowości nazywa się współczynnikiem inflacji wariancji:

$$VIF_k = \frac{1}{1 - R_k^2}$$

gdzie:

R_k^2 - R^2 w regresji x_k na pozostalych zmiennych objaśniających

Współliniowość

- Wysokie wartości VIF (>10) dla zmiennych objaśniających sygnalizują występowanie silnej niedokładnej współliniowości między zmiennymi
- Rozwiązaniem problemu silnej niedokładnej współliniowości jest usunięcie zmiennej o najwyższym VIF, co powinno poprawić precyzję oszacowań przy pozostałych zmiennych
- Należy jednak pamiętać, że jeśli usunięta zmienna była istotna w modelu to jej usunięcie może spowodować obciążenie estymatorów przy zmiennych, z którymi jest skorelowana
- Niedokładna współliniowość nie jest wynikiem błędnej specyfikacji modelu lecz wynika z własności konkretnego zbioru danych

Współliniowość

```
reg wydg dochg dochg2 dochg3

      Source |       SS       df       MS              Number of obs =   31679
-------------+------------------------------           F(  3, 31675) = 8591.16
       Model |  2.5996e+10     3  8.6653e+09           Prob > F      =  0.0000
    Residual |  3.1948e+10 31675  1008629.31           R-squared     =  0.4486
-------------+------------------------------           Adj R-squared =  0.4486
       Total |  5.7944e+10 31678  1829163.02           Root MSE      =  1004.3

------------------------------------------------------------------------------
        wydg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dochg |   .8616921   .0080228   107.41   0.000     .8459671     .877417
     dochg2 |  -.0000275   9.05e-07   -30.36   0.000    -.0000292   -.0000257
     dochg3 |   2.76e-10   1.58e-11    17.46   0.000     2.45e-10    3.07e-10
      _cons |   316.3596   13.51433    23.41   0.000      289.871    342.8482
```

Współliniowość

```
  vif

    Variable |       VIF       1/VIF
-------------+----------------------
      dochg2 |     21.44    0.046644
      dochg3 |     13.40    0.074609
       dochg |      4.35    0.229769
-------------+----------------------
    Mean VIF |     13.06
```

Plan wykładu

- 1. Problemy z danymi
  - Obserwacje nietypowe i błędne
  - Współliniowość
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji

Heteroskedastyczność

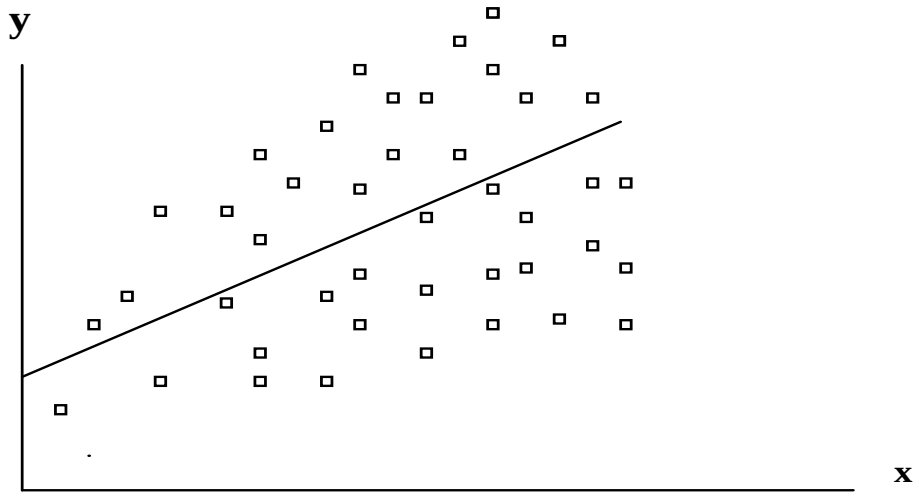
Przypomnienie: Co to znaczy, że w modelu występuje homoskedastyczność/heteroskedastyczność?

- heteroskedastyczność

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & \cdots & Cov(\varepsilon_1,\varepsilon_n) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2,\varepsilon_n) \\ \vdots & \vdots & & \vdots \\ Cov(\varepsilon_n,\varepsilon_1) & Cov(\varepsilon_n,\varepsilon_1) & \cdots & Var(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Heteroscedastyczność

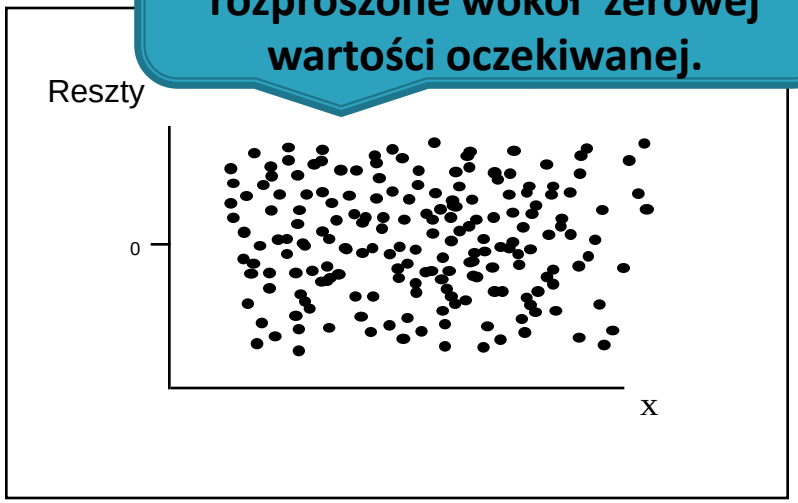
- Stałość wariancji zaburzeń nazywamy homoskedastycznością zaburzeń. **Oznacza to, że zaburzenia losowe są jednakowo rozproszone wokół zerowej wartości oczekiwanej.** Jeśli wariancje nie byłyby jednakowe, to sytuację taką nazywamy heteroskedastycznością.

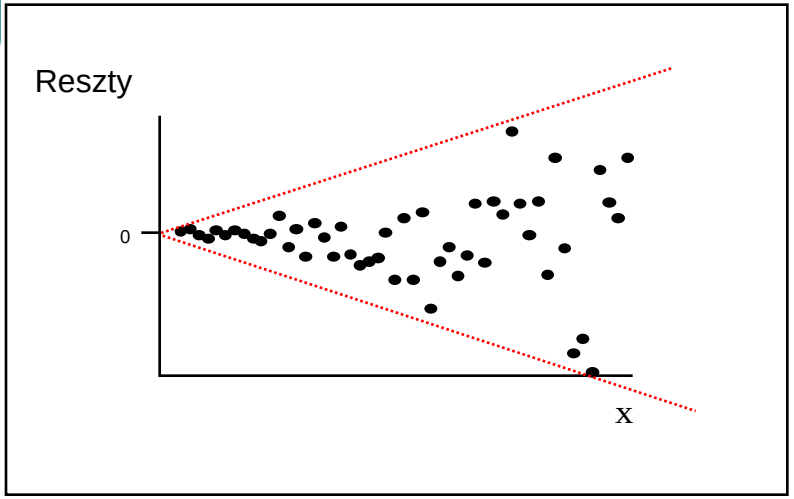


Rys.1. **Heteroskedastyczność**

Heteroskedastyczność

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



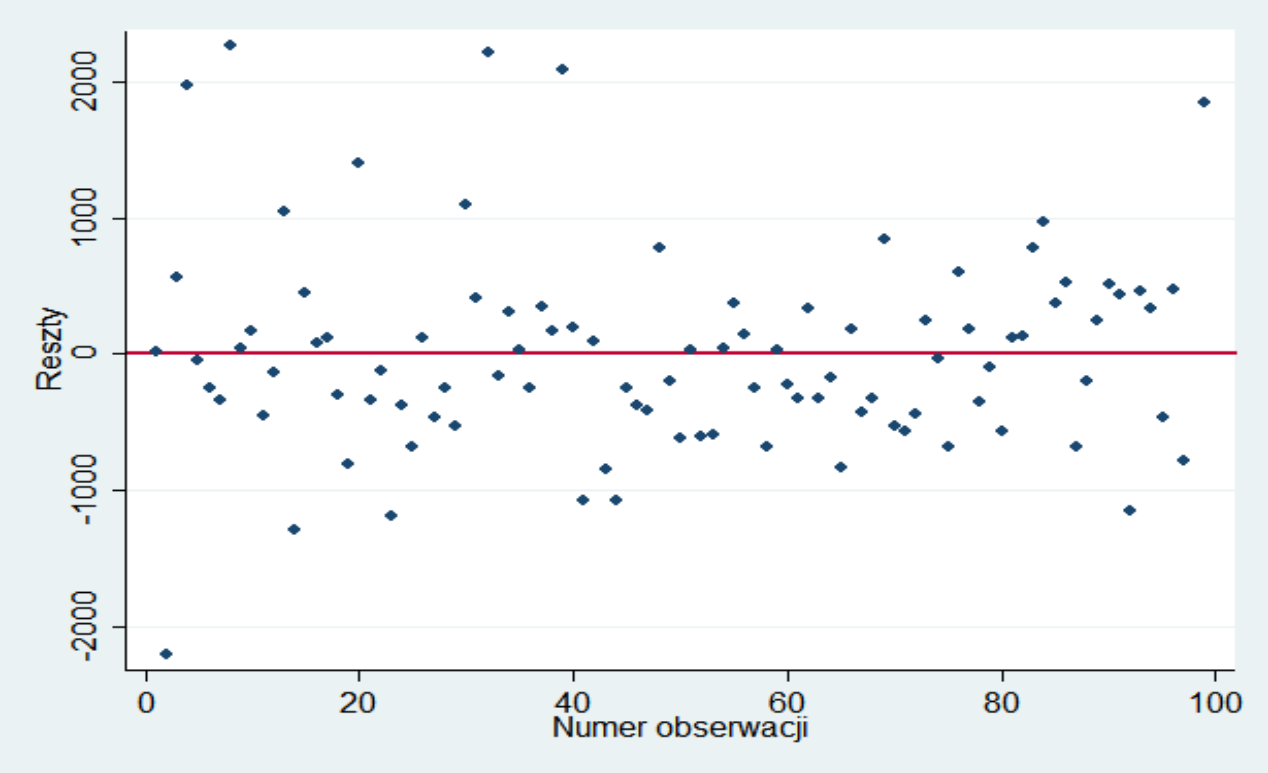


Homoskedastyczność: reszty zachowują się losowo.

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

Heteroskedastyczność

- Regresja wydatków na dochodzie



Autokorelacja

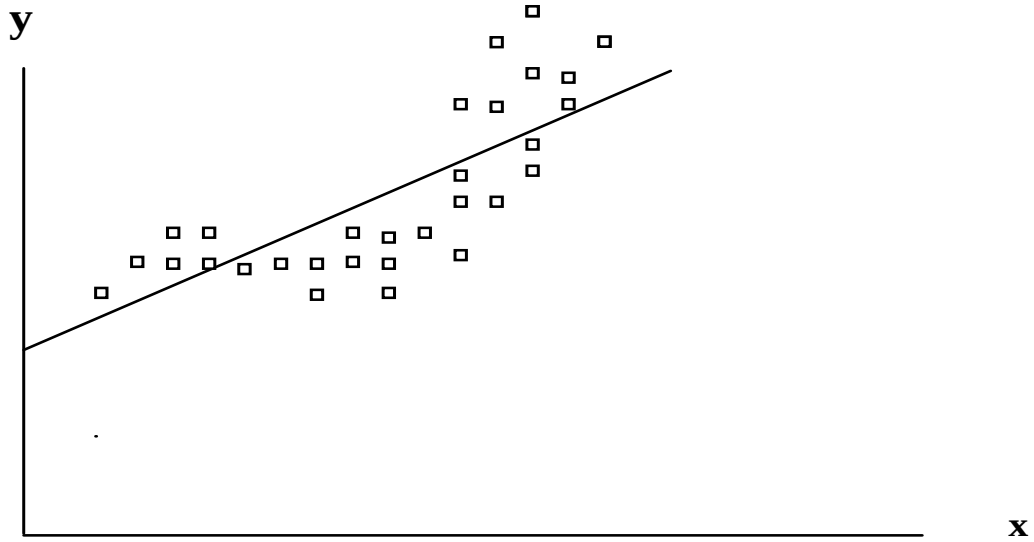
Przypomnienie: Co to znaczy, że w modelu występuje autokorelacja?

-Brak autokorelacji

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

Autokorelacja

- 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**

Autokorelacja

$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) > 0$ dla $i \neq j$ - dodatnia autokorelacja

$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) < 0$ dla $i \neq j$ - ujemna autokorelacja

Sferyczność błędów losowych

- Jeżeli założenie o homoskedastyczności i autokorelacji jest spełnione to błędy losowe są **sferyczne**

- Jeżeli, któreś z tych założeń nie jest spełnione to błędy losowe są **niesferyczne** a macierz wariancji i kowariancji ma postać dowolnej macierzy symetrycznej i dodatnio półokreślonej:

$$Var(\varepsilon) = \Omega = \sigma^2 V$$

Konsekwencje heteroskedastyczności i autokorelacji

- Estymator b jest nadal **nieobciążony**:

$$E(b) = E\left[(X'X)^{-1}X'y\right] =$$

$$E\left[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\right] =$$

$$\beta + (X'X)^{-1}X'E(\varepsilon) = \beta$$

- Nie będzie on jednak **efektywny** ⟶ można znaleźć estymator o mniejszej wariancji

Konsekwencje heteroskedastyczności i autokorelacji

- Macierz wariancji i kowariancji b:

$$Var(b) = E\left((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right) =$$

$$(X'X)^{-1}X'\Omega X(X'X)^{-1} =$$

$$\sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$$

- Wzór ten różni się znacznie od prawidłowego wzoru na wariancję MNK:

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

Konsekwencje heteroskedastyczności i autokorelacji

- W rezultacie estymator macierzy wariancji i kowariancji b, którym posługiwaliśmy się do tej pory, nie będzie dobrym oszacowaniem macierzy wariancji i kowariancji b

Dziękuję za uwagę