Testowanie autokorelacji Problemy z danymi

Stanisław Cichocki

Natalia Nehrebecka

Wykład 13

Plan wykładu

- 1. Testowanie autokorelacji
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
- 3.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
  - Obserwacje nietypowe i błędne
  - Współliniowość

Plan wykładu

- 1. Testowanie autokorelacji
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
- 3.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
  - Obserwacje nietypowe i błędne
  - Współliniowość

Testowanie autokorelacji

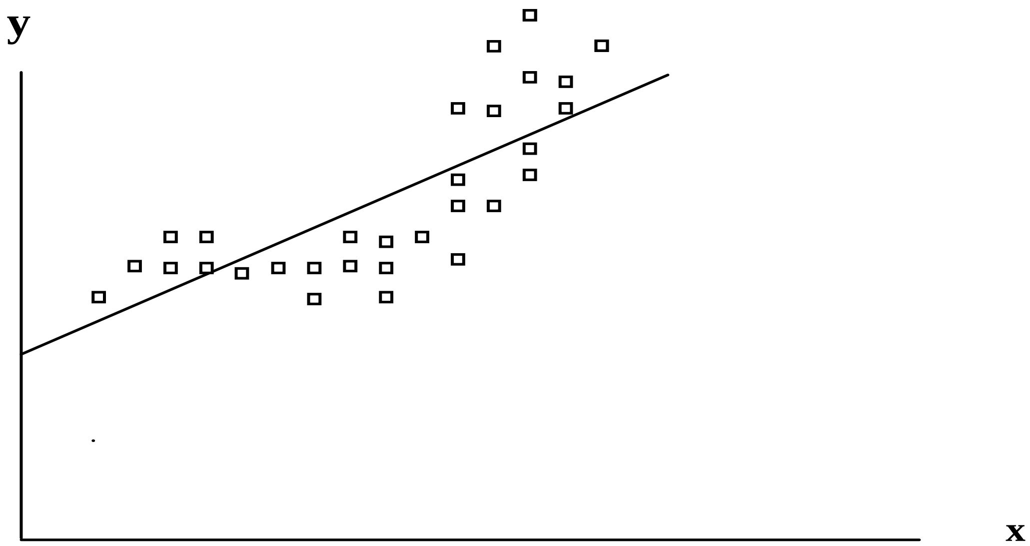
Przypomnienie: Co to znaczy, że w modelu występuje autokorelacja?

-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**

Testowanie autokorelacji

- Test Durbina-Watsona (Test DW):

$$H_0 : Cov(\varepsilon_t, \varepsilon_{t-1}) = 0$$ **- brak autokorelacji**

$$H_1 : Cov(\varepsilon_t, \varepsilon_{t-1}) \neq 0$$ **- autokorelacja**

gdzie $t = 1, \ldots, T$

Testowanie autokorelacji

- **Test Durbina-Watsona (Test DW):**

- specjalne tablice z wartościami krytycznymi: d_l, d_u

1. Statystyka DW<2

a) DW < d_l - odrzucamy hipotezę zerową o braku autokorelacji i przyjmujemy hipotezę o dodatniej autokorelacji

b) d_l < DW < d_u - brak konkluzji

c) DW > d_u - nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji

Testowanie autokorelacji

- Test Durbina-Watsona (Test DW):

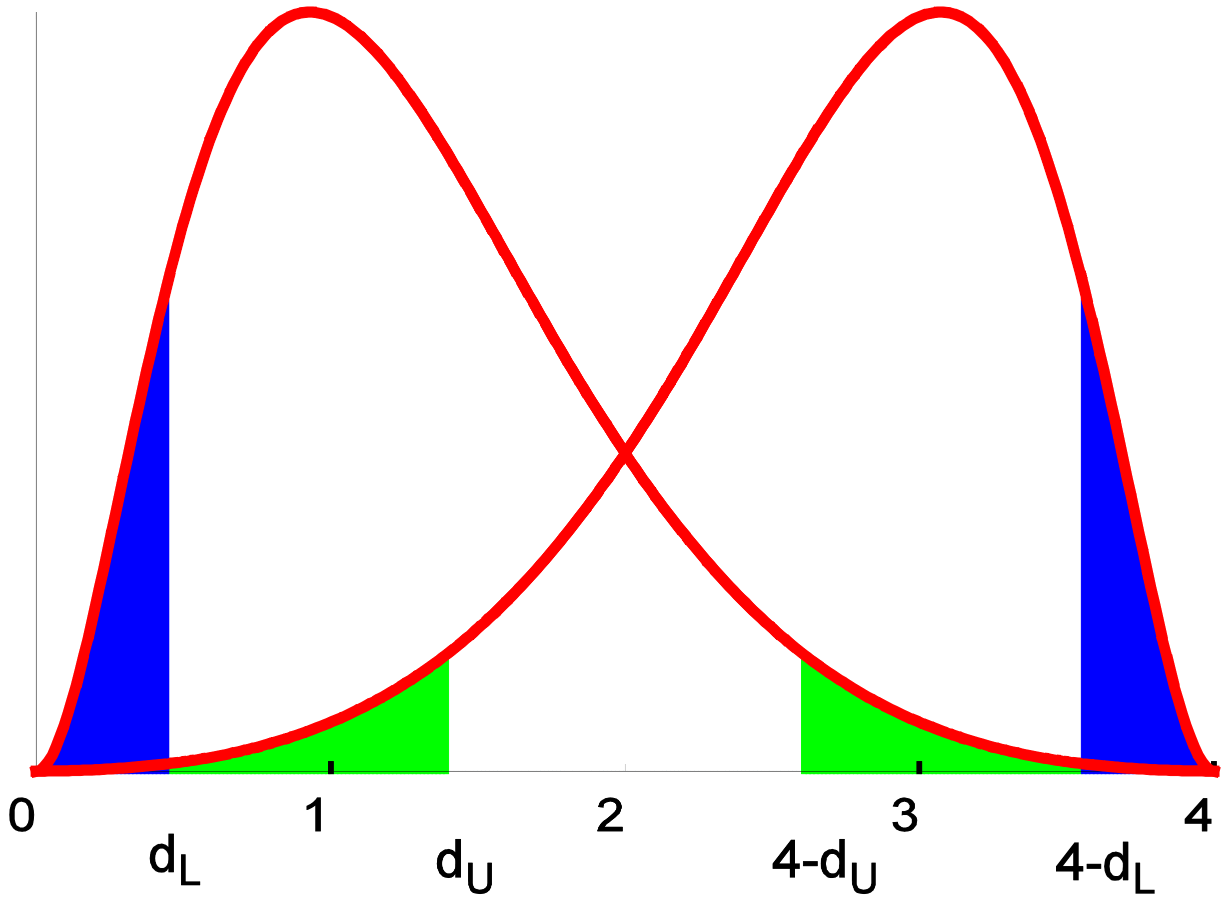
2. Statystyka DW >2

a) DW > $4-d_l$ - odrzucamy hipotezę zerową o braku autokorelacji i przyjmujemy hipotezę o ujemnej autokorelacji

b) $4-d_u$ < DW < $4-d_l$ - brak konkluzji

c) DW < $4-d_u$ - nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji

Testowanie autokorelacji



Testowanie autokorelacji

- Test Durbina-Watsona (Test BW):

- Do badania autokorelacji I rzędu (między $\varepsilon_t, \varepsilon_{t-1}$)

- Rozkład statystki testowej wyprowadzony dla małych prób

- Nie można go stosować w modelach gdzie jedną ze zmiennych objaśniających jest opóźniona zmienna zależna

- Wada: niestandardowy rozkład i możliwość wystąpienia braku konkluzji

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG):

- Do badania autokorelacji wyższego rzędu

- Można go stosować w modelach gdzie występują opóźnione zmienne zależne

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG):

$$H_0 : Cov(\varepsilon_{t,} \varepsilon_{t-i}) = 0$$ **gdzie** $$i = 1, \dots, s$$

$$H_1 : \varepsilon_t = \gamma_1 \varepsilon_{t-1} + \dots + \gamma_s \varepsilon_{t-s} + u_t$$ **gdzie** $$Var(u) = \sigma_u^2 I$$

- Hipoteza zerowa: brak autokorelacji

- Hipoteza alternatywna: autokorelacja

Testowanie autokorelacji

- Test Breuscha-Godfreya (Test BG) – sposób przeprowadzenia testu:

1. przeprowadzamy regresję y_i na x_i i uzyskujemy reszty

2. przeprowadzamy regresję pomocniczą:

$$e_t = x_t \hat{\mu} + \hat{\gamma}_1 e_{t-1} + \ldots + \hat{\gamma}_s e_{t-s} + \hat{u}_t$$

i testujemy H0: $\gamma_1 = \ldots. = \gamma_s = 0$

Testowanie autokorelacji

- Statystyka testowa:

$$LM = TR^2 \xrightarrow{D} \chi_p^2$$

lub statystyka F

Jakie założenie KMRL nie jest spełnione przy odrzuceniu H0?

- Brak autokorelacji błędu losowego – kowariancja dwóch różnych błędów losowych jest zerowa:

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad dla \quad i \neq j$$

Plan wykładu

- 1. Testowanie autokorelacji
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
- 3.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
  - Obserwacje nietypowe i błędne
  - Współliniowość

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** $\Longrightarrow$ 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

Plan wykładu

- 1. Testowanie autokorelacji
- 2. Heteroskedastyczność i autokorelacja
  - Konsekwencje heteroskedastyczności i autokorelacji
- 3.Problemy z danymi
  - Zmienne pominięte
  - Zmienne nieistotne
  - Obserwacje nietypowe i błędne
  - Współliniowość

Zmienne pominięte

- Mamy 2 modele:

$$y = X_1\beta_1 + u \qquad (1)$$

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \qquad (2)$$

- Potencjalnie każdy z tych modeli może prawidłowo opisywać zmienną $y \Longrightarrow$ problemy gdy przy liczeniu estymatorów zastosujemy niewłaściwy model

- Załóżmy, ze estymujemy model (1) a prawdziwy jest model (2)

Zmienne pominięte

- Zakładamy, że $\beta_2 = 0$ gdy w rzeczywistości $\beta_2 \neq 0$

- Przypadek ten nazywamy problemem **zmiennych pominiętych** (ommitted variables)

Zmienne pominięte

- $\hat{\beta}_1$ - estymator MNK wektora parametrów w modelu (1)

- Założmy, że prawdziwy jest model (2)

$$\hat{\beta}_1 = (X_1^{'} X_1)^{-1} X_1^{'} y = (X_1^{'} X_1)^{-1} X_1^{'} (X_1 \beta_1 + X_2 \beta_2 + \varepsilon)$$

$$= \beta_1 + (X_1^{'} X_1)^{-1} X_1^{'} X_2 \beta_2 + (X_1^{'} X_1)^{-1} X_1^{'} \varepsilon$$

Zmienne pominięte

- $$E(\hat{\beta}_1) = \beta_1 + (X_1^{'}X_1)^{-1}X_1^{'}X_2\beta_2 + (X_1^{'}X_1)^{-1}X_1^{'}E(\varepsilon)$$
$$= \beta_1 + (X_1^{'}X_1)^{-1}X_1^{'}X_2\beta_2$$

- Jeśli więc pominiemy istotne zmienne estymator nie jest estymatorem nieobciążonym

- Obciążenie: $$E(\hat{\beta}_1) - \beta_1 = (X_1^{'}X_1)^{-1}X_1^{'}X_2\beta_2$$

Zmienne pominięte

- Dwa przypadki, dla których pominięcie zmiennej nie powoduje obciążenia estymatora

a) $\beta_2 = 0$

b) $X_1^{'} X_2 = 0$ - zmienne pominięte nie są skorelowane ze zmiennymi objaśniającymi, które zostały uwzględnione w modelu

Zmienne pominięte

- Pominięcie istotnych zmiennych jest prawdopodobnie najczęstszym powodem błędów w oszacowaniach

- W praktyce nigdy nie dysponujemy danymi odnośnie wszystkich zmiennych mogących wpływać na zmienną zależną

- W takim przypadku warto umieć określić kierunek ewentualnego obciążenia (trudne w ogólnym przypadku)

Zmienne pominięte

- Kierunek obciążenia dla najprostszego przypadku (model ze stałą i jedną zmienną objaśniającą, pominięta jedna dodatkowa zmienna objaśniająca):

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \frac{s_{x_2}}{s_{x_1}} \rho_{x_1 x_2}$$

gdzie:

s_{x_1}, s_{x_2} - wariancja empiryczna x_1, x_2

$\rho_{x_1 x_2}$ - wsp. korelacji miedzy x_1 a x_2

Zmienne pominięte

- Kierunek obciążenia dla najprostszego przypadku (model ze stałą i jedną zmienną objaśniającą, pominięta jedna dodatkowa zmienna objaśniająca):

Przypadek	Wpływ zmiennej pominiętej na zmienną zależną (β_2)	Korelacja między zmienną pominiętą a zmienną niezależną (ρ)	Znak obciążenia
I	**+**	**+**	**+ (przeszacowanie)**
II	**-**	**-**	**+**
III	**+**	**-**	**- (niedoszacowanie)**
IV	**-**	**+**	**-**

Zmienne pominięte

- Przykład:

Dla pewnej badanej grupy osób przeprowadzono regresję logarytmu wynagrodzenia na latach nauki (zmienna *latanauki)*. Jaki będzie prawdopodobny kierunek obciążenia parametru przy zmiennej *latanauki* wynikający z pominięcia:

a) wielkości miejscowości, w której zamieszkuje badana osoba;

b) liczby dzieci badanej osoby?

Zmienne pominięte

- Obciążenie może prowadzić do:

a) Uznania za zmienną istotną zmiennej, która nie ma żadnego wpływu na zmienna zależną → najgorszy przypadek

b) Przeszacowania/niedoszacowania wpływu zmiennej objaśniającej na zmienna objaśnianą

Zmienne pominięte

- Przykład

reg wydg dochg

Source	SS	df	MS
Model	2.3577e+10	1	2.3577e+10
Residual	3.4367e+10	31677	1084914.37
Total	5.7944e+10	31678	1829163.02

Number of obs	= 31679
F(1, 31677)	=21732.03
Prob > F	= 0.0000
R-squared	= 0.4069
Adj R-squared	= 0.4069
Root MSE	= 1041.6

wydg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dochg	.5879668	.0039884	147.42	0.000	.5801493	.5957843
_cons	712.8104	10.01991	71.14	0.000	693.171	732.4498

Zmienne pominięte

- Przykład

reg wydg dochg los

```
      Source |       SS       df       MS              Number of obs =   31679
-------------+------------------------------           F(  2, 31676) =11107.42
       Model |  2.3886e+10     2  1.1943e+10           Prob > F      =  0.0000
    Residual |  3.4059e+10 31676  1075214.71           R-squared     =  0.4122
-------------+------------------------------           Adj R-squared =  0.4122
       Total |  5.7944e+10 31678  1829163.02           Root MSE      =  1036.9

------------------------------------------------------------------------------
        wydg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dochg |   .5688205   .0041284   137.78   0.000     .5607287    .5769123
         los |   65.35337   3.859286    16.93   0.000     57.78902    72.91772
       _cons |   548.4807   13.91655    39.41   0.000     521.2037    575.7577
------------------------------------------------------------------------------
```

Zmienne nieistotne

- Mamy 2 modele:

$$y = X_1\beta_1 + u \qquad (1)$$

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \qquad (2)$$

- Załóżmy, ze estymujemy model (2) a prawdziwy jest model (1)

- Zakładamy, że $\beta_2 \neq 0$ gdy w rzeczywistości $\beta_2 = 0$

- Przypadek ten nazywamy problemem zmiennych nieistotnych

Zmienne nieistotne

- Estymator β_1- nieobciążony, ale będzie miał większą wariancję niż estymator uzyskany na podstawie modelu (1)

- Inaczej mówiąc, w modelu w którym występują zmienne nieistotne estymator MNK ma wyższą wariancję niż w modelu, z którego usunięto zmienne nieistotne

Zmienne nieistotne

- Usuwamy z modelu zmienne nieistotne bo:

a) Poprawia to precyzję oszacowań parametrów przy zmiennych istotnych (estymator MNK ma mniejszą wariancję)

b) Uzyskujemy uproszczenie modelu

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:

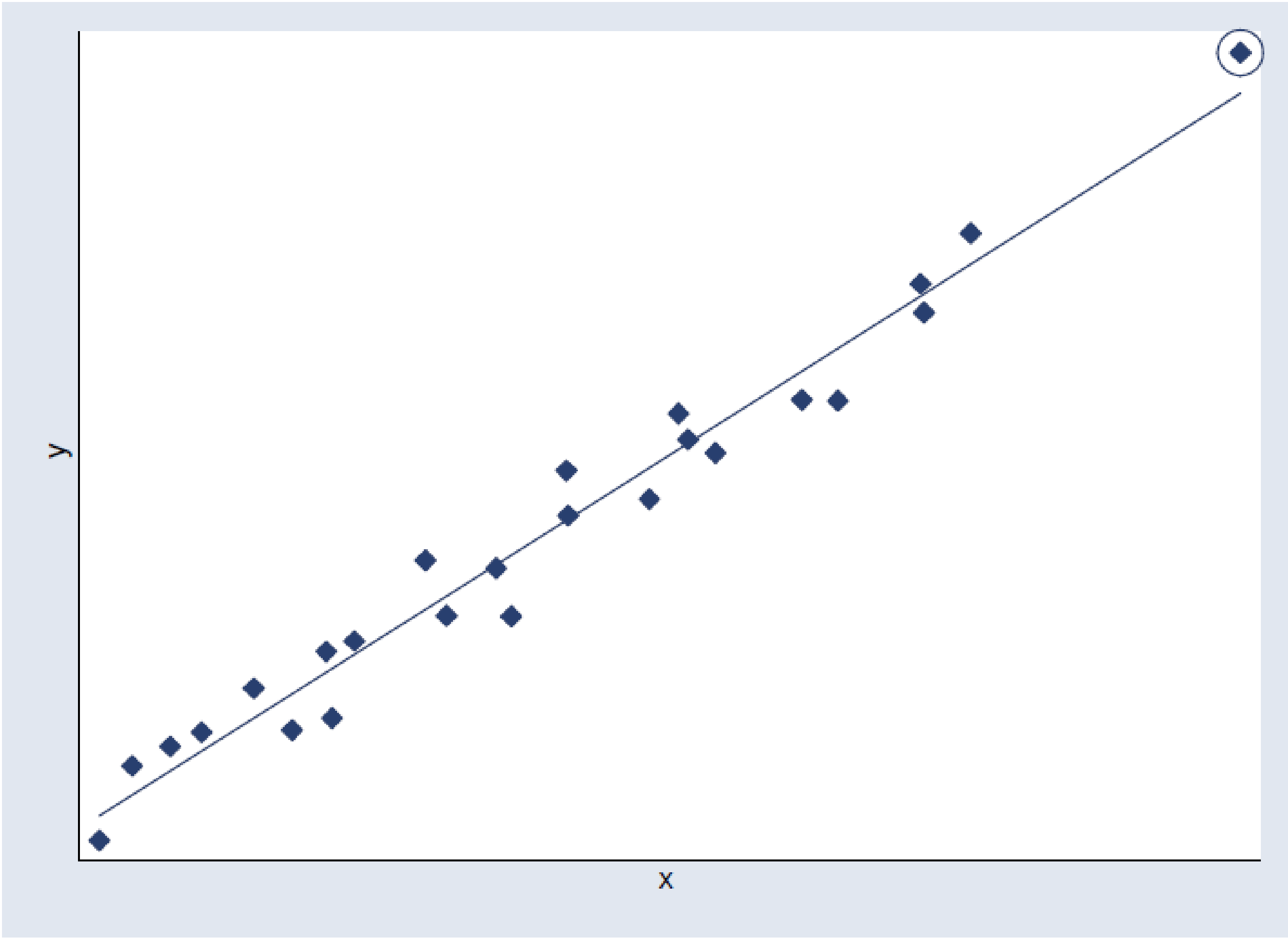
Estymujemy krzywą popytu na żywność dla różnych państw na świecie. W próbie występują państwa, w których obowiązuje reglamentacja żywności. Obserwacje takie traktujemy jako obserwacje błędne – teoria opisująca krzywą popytu nie znajduje zastosowania w momencie nierynkowego podziału dóbr.

Obserwacje nietypowe i błędne

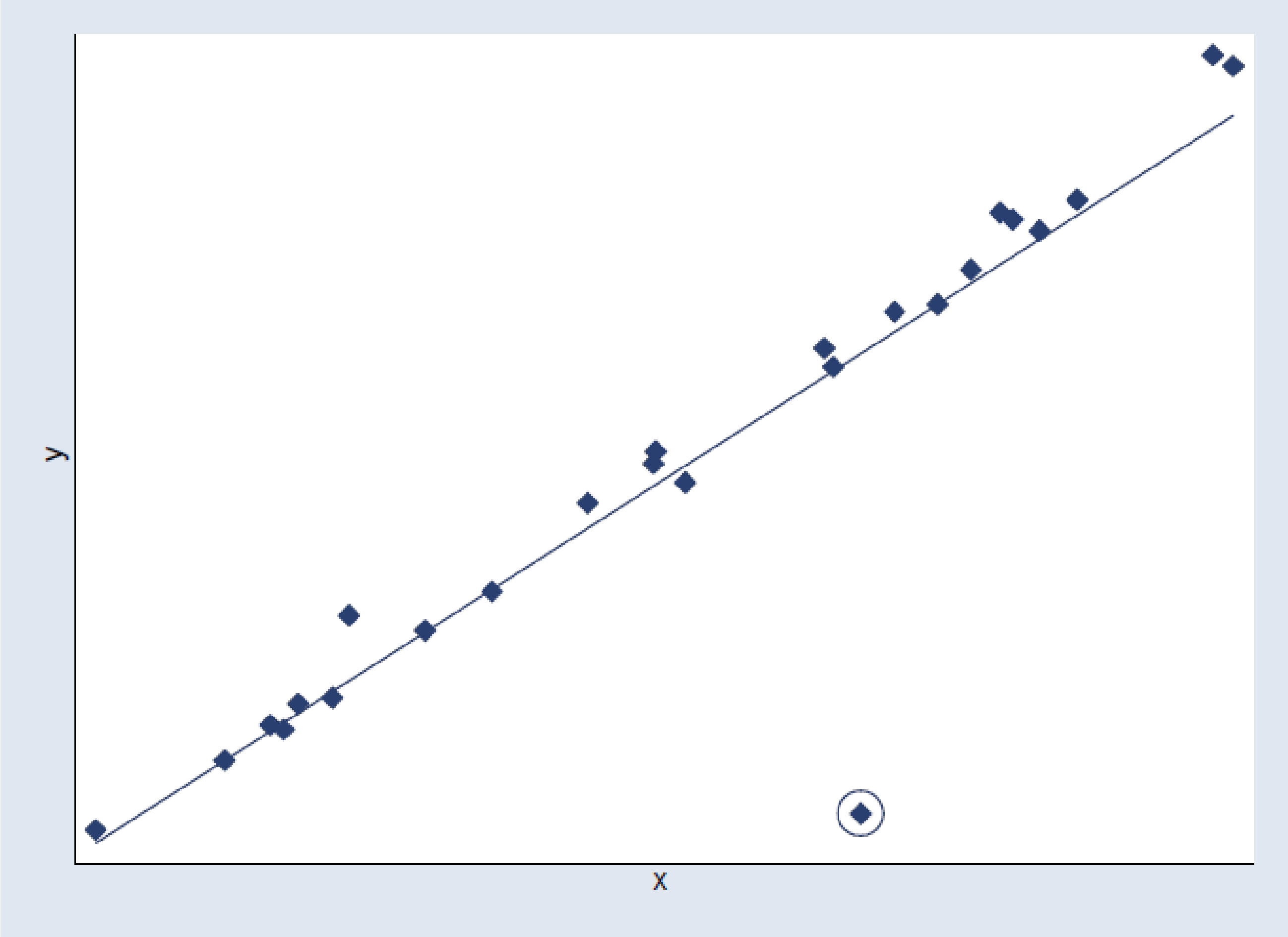
- Wpływ obserwacji nietypowej/błędnej na wynik regresji zależy od tego na ile ta obserwacja pasuje do prostej regresji

- 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:

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

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:

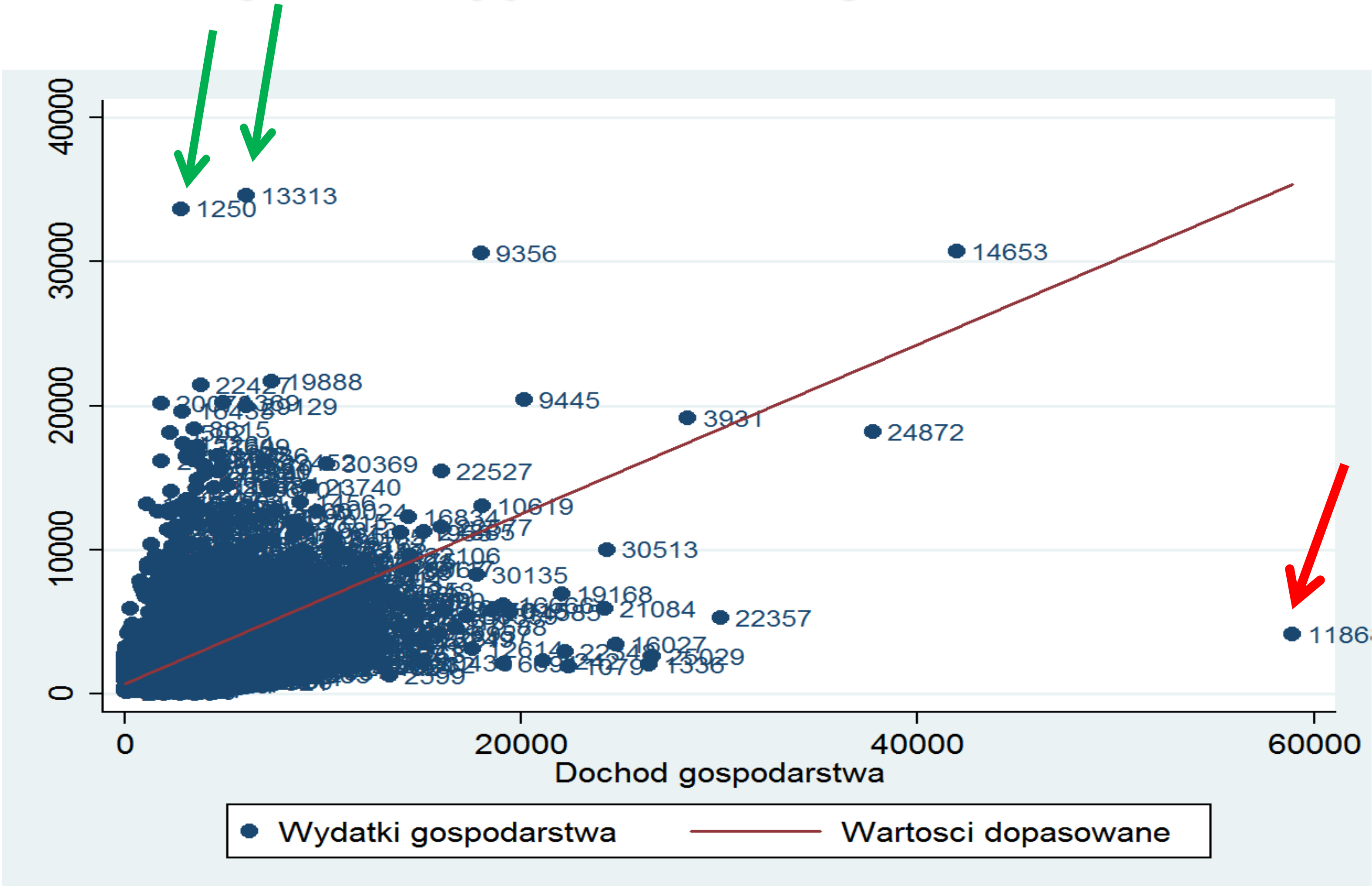
$$\left|\hat{e}_i\right| > 2$$

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

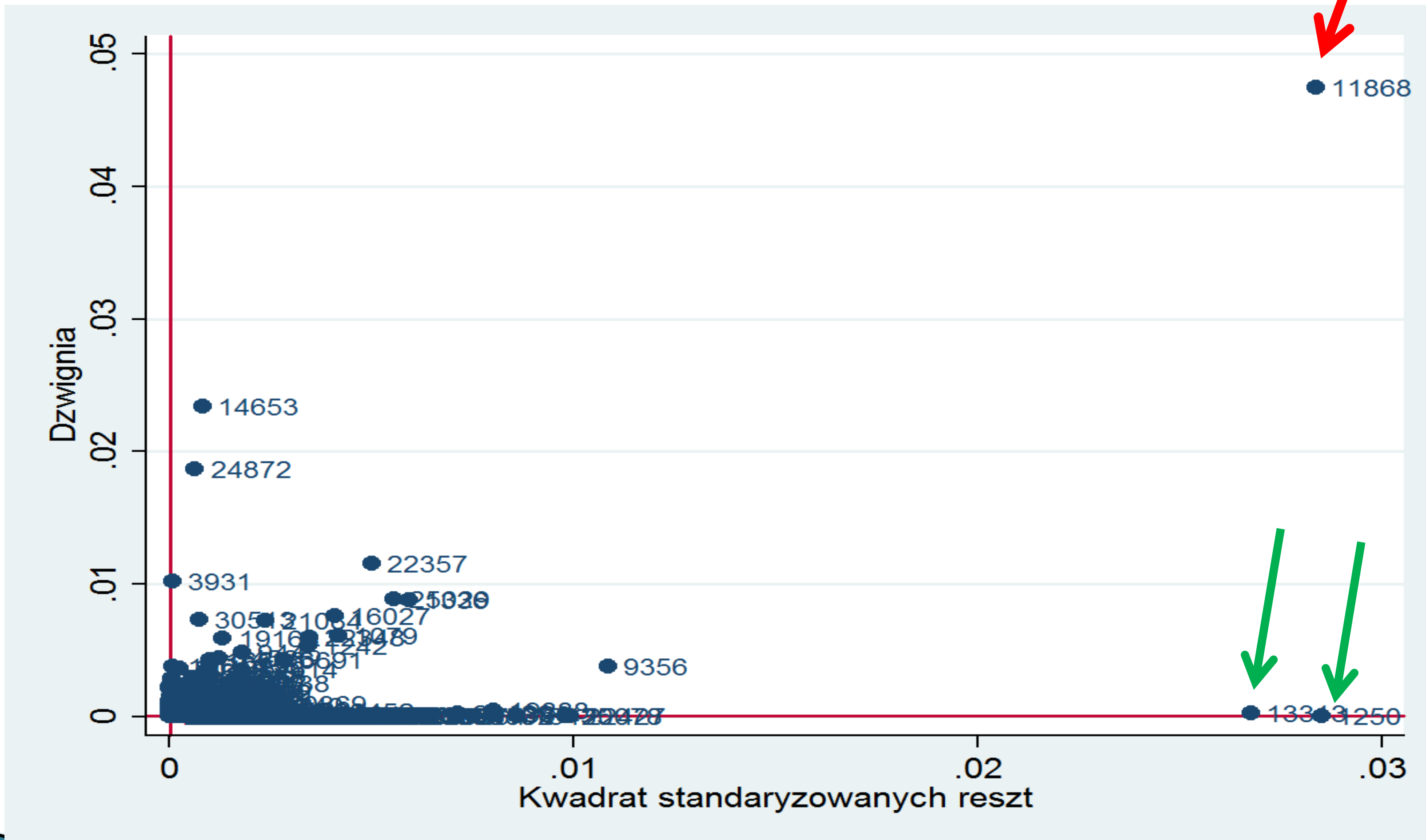
$$\left|\hat{e}_i\right| > 2$$

- 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 $\longrightarrow$ 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

Dziękuję za uwagę