Problemy z danymi (cz. II)

Natalia Nehrebecka
Stanisław Cichocki

Wykład 13

Plan zajęć

1. Zmienne pominięte

2. Zmienne nieistotne

3. Obserwacje nietypowe i błędne

4. Współliniowość

Plan zajęć

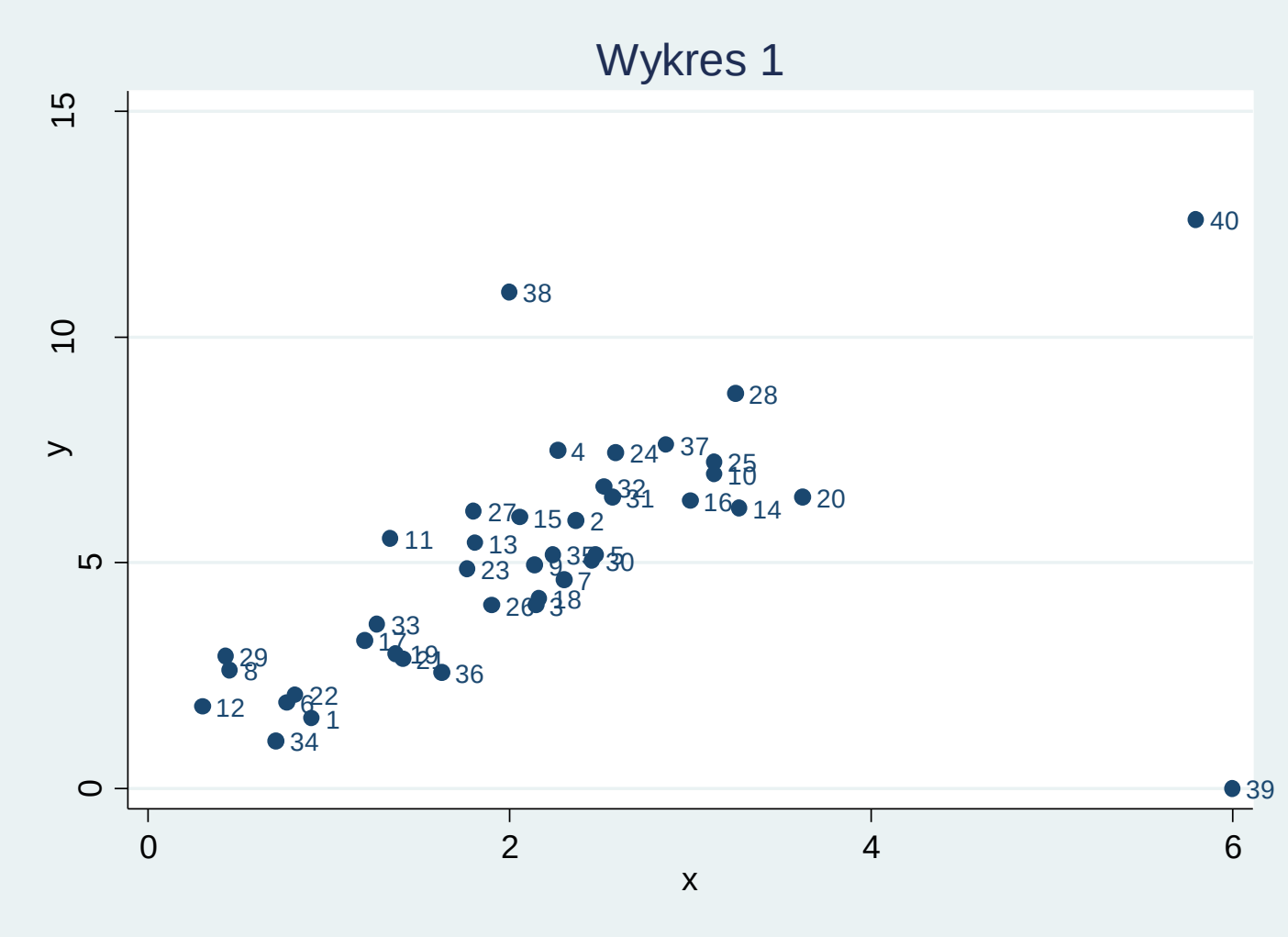
1. Zmienne pominięte

2. Zmienne nieistotne

3. Obserwacje nietypowe i błędne

4. Współliniowość

Obserwacje nietypowe i błędne



Obserwacje nietypowe i błędne

- **Obserwacja nietypowa** charakteryzuje się nietypowymi na tle pozostałych obserwacji cechami

- Mechanizm, który w przypadku tej obserwacji 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

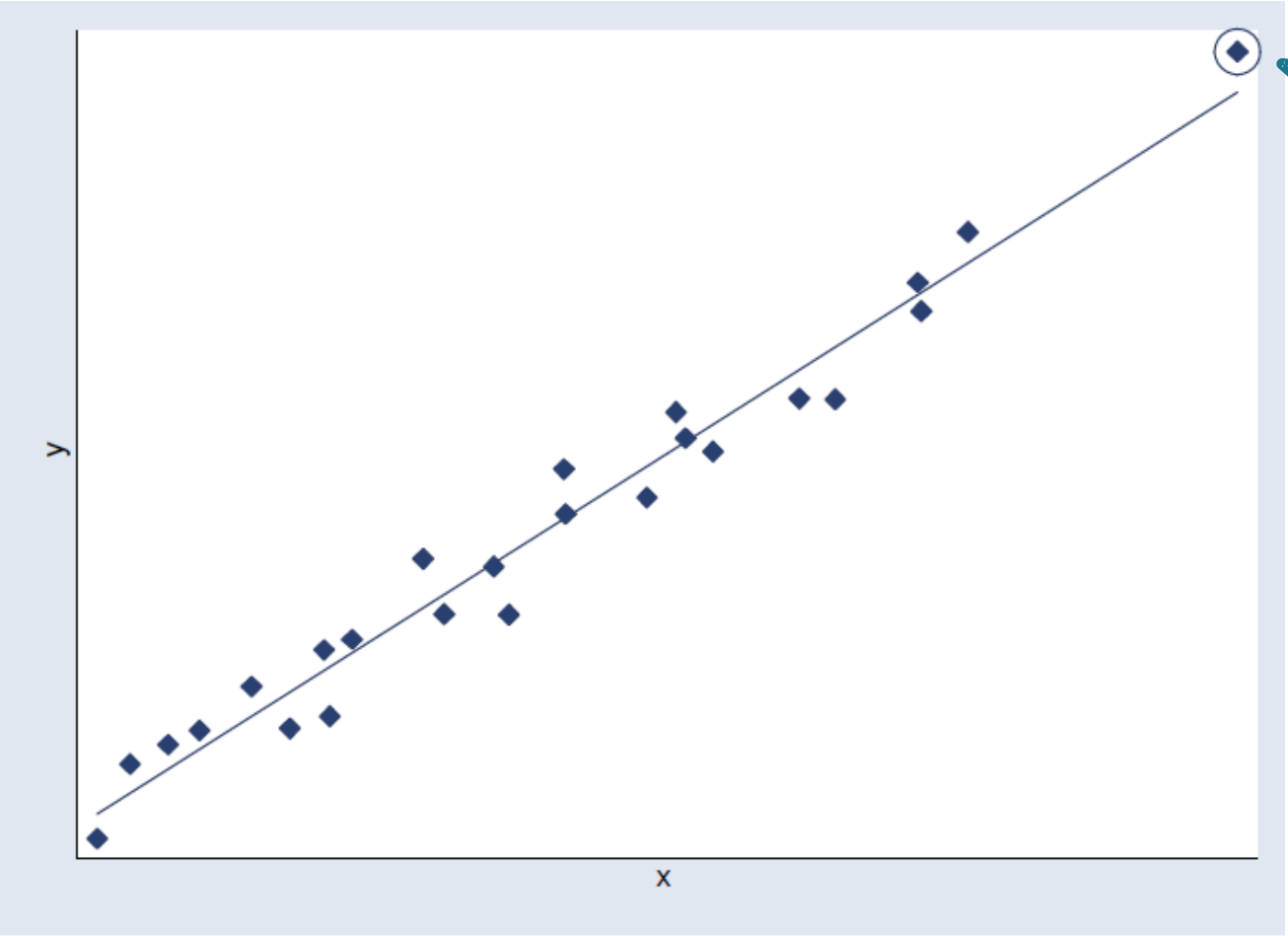
- Przykład:
  - Estymujemy **krzywą popytu na żywność dla różnych państw na świecie**.
  - W próbie występują państwa, w których obowiązuje reglamentacja żywności.
  - Obserwacje takie traktujemy jako obserwacje błędne – teoria opisująca krzywą popytu nie znajduje zastosowania w momencie nierynkowego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

Obserwacje nietypowe i błędne

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

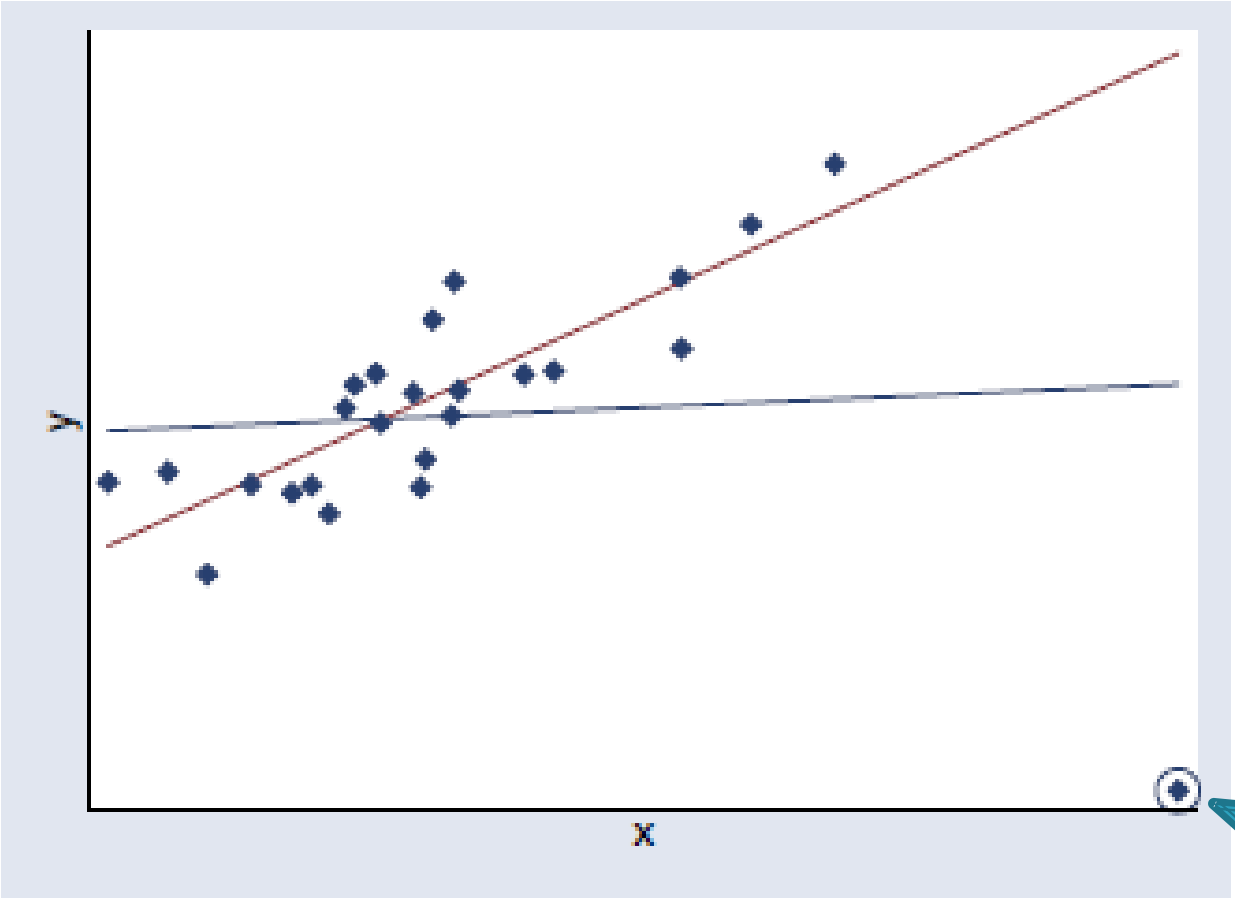
- Najbardziej niepokojąca jest sytuacja gdy obserwacja ma **nietypowe wartości dla zmiennych niezależnych i słabo pasuje do prostej regresji**

Obserwacje nietypowe i błędne



Obserwacja nietypowa pasująca do linii regresji

Obserwacje nietypowe i błędne



Obserwacja nietypowa niepasująca do linii regresji

Obserwacje nietypowe i błędne

- Przykład:

  Badamy wynagrodzenia dla próby osób przebadanej 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 rentowność 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

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

Wykrywanie obserwacji nietypowych

W zależności od wielkości reszty i dźwigni dla danej obserwacji, możemy wyróżnić trzy interesujące grupy obserwacji:

- **duża reszta**,
- leverage points (**duża dźwignia**),
- influential points (**duże wartości zarówno reszty, jak i leverage**).

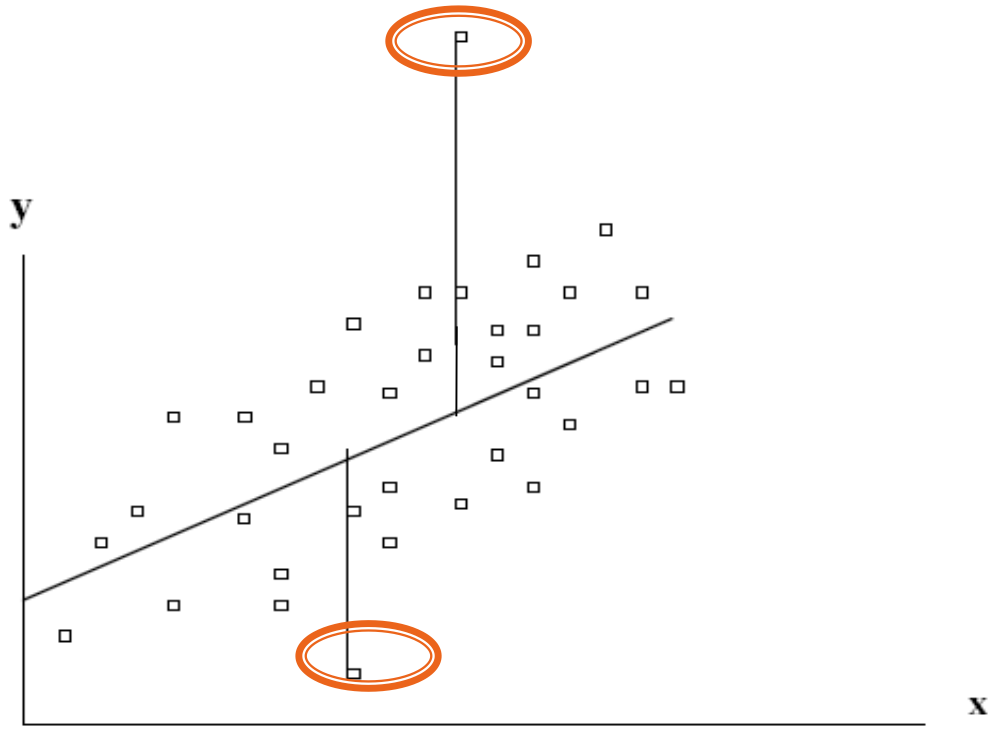
Wszystkie takie obserwacje powinny być dokładnie zbadane:

- mogą być rezultatem błędu przy wprowadzaniu danych, reprezentować dane spoza badanej populacji lub też zarejestrowane w nadzwyczajnych okolicznościach.
- mogą jednak również zawierać kluczowe dla nas informacje, zatem nie należy ich lekka ręką usuwać ze zbioru.

Wykrywanie obserwacji nietypowych

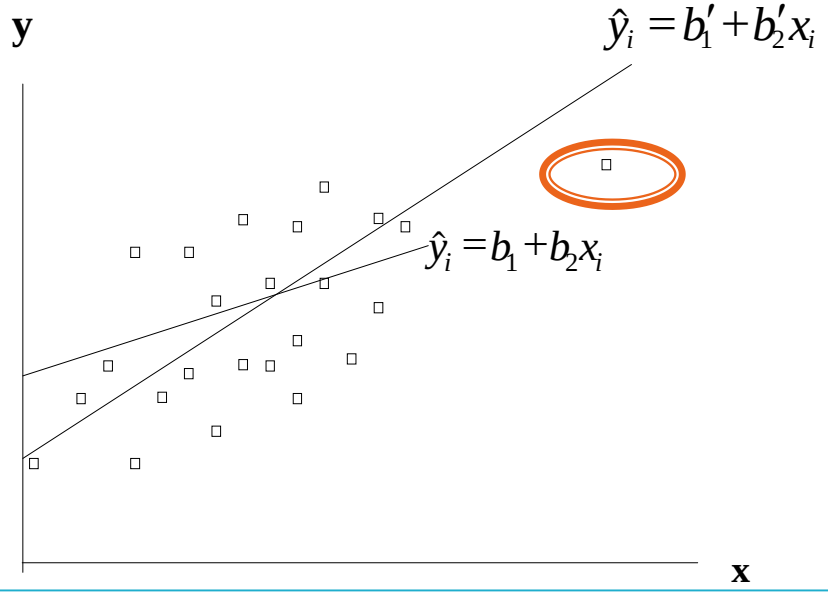
Można wyróżnić następujące rodzaje obserwacji nietypowych:

- Pierwszy ich rodzaj zwany **nietypowymi obserwacjami zmiennej objaśnianej** charakteryzuje się nieoczekiwanie **dużymi resztami**.



Wykrywanie obserwacji nietypowych

- Drugi rodzaj, to tak zwane **nietypowe obserwacje zmiennych objaśniających** lub punkty dźwigniowe (*leverage points*).
  - Cecha charakterystyczna punktów dźwigniowych jest ich znaczne oddalenie od środka zmienności zmiennych objaśniających, co istotnie wpływa na wyznaczone oceny parametrów przy jednocześnie małej wielkości reszty



Obserwacje nietypowe i błędne

- **Statystyki** 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:

Macierz projekcji (rzutu)

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

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

- Wariancja elementu *i* wektora reszt:
- $Var(e_i) = Var(\delta'_i e) = \delta'_i \underbrace{Var(e)}_{\sigma^2 M_x} \delta_i = \sigma^2 \delta'_i \underbrace{M_x}_{I-P_x} \delta_i =$

$$= \sigma^2 \left[\delta'_i (\mathrm{I} - P_x)\delta_i\right] = \sigma^2 \left[\delta'_i (\mathrm{I} - \mathrm{X}(X'X)^{-1}X')\delta_i\right]$$

$$= \sigma^2 \left[\delta'_i \delta_i - \delta'_i \mathrm{X}(X'X)^{-1}X'\delta_i\right] = \sigma^2 \left[1 - \underbrace{\delta'_i P_x \delta_i}_{h_i}\right] =$$

$$= \sigma^2 (1 - h_i)$$

  - gdzie: $\delta_i = [0, \dots, 0, 1, 0, \dots 0]'$

Obserwacje nietypowe i błędne

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

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

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

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

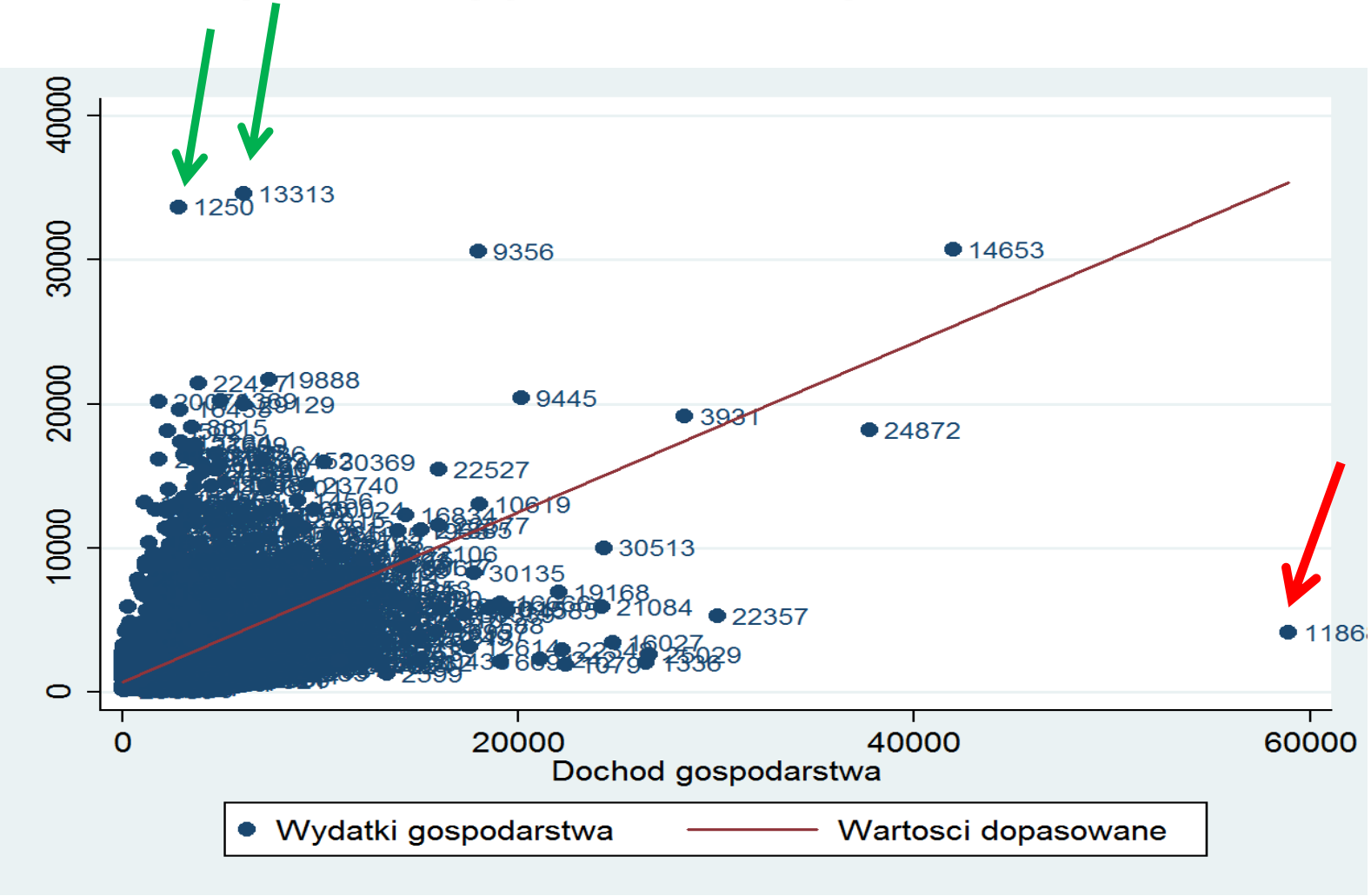
Obserwacje nietypowe i błędne

- Dla nietypowej obserwacji:
- $|\hat{e}_i| > 2$
- Jednak (*jeżeli błąd losowy ma rozkład normalny*), to statystycznie dla ok. 5% obserwacji:

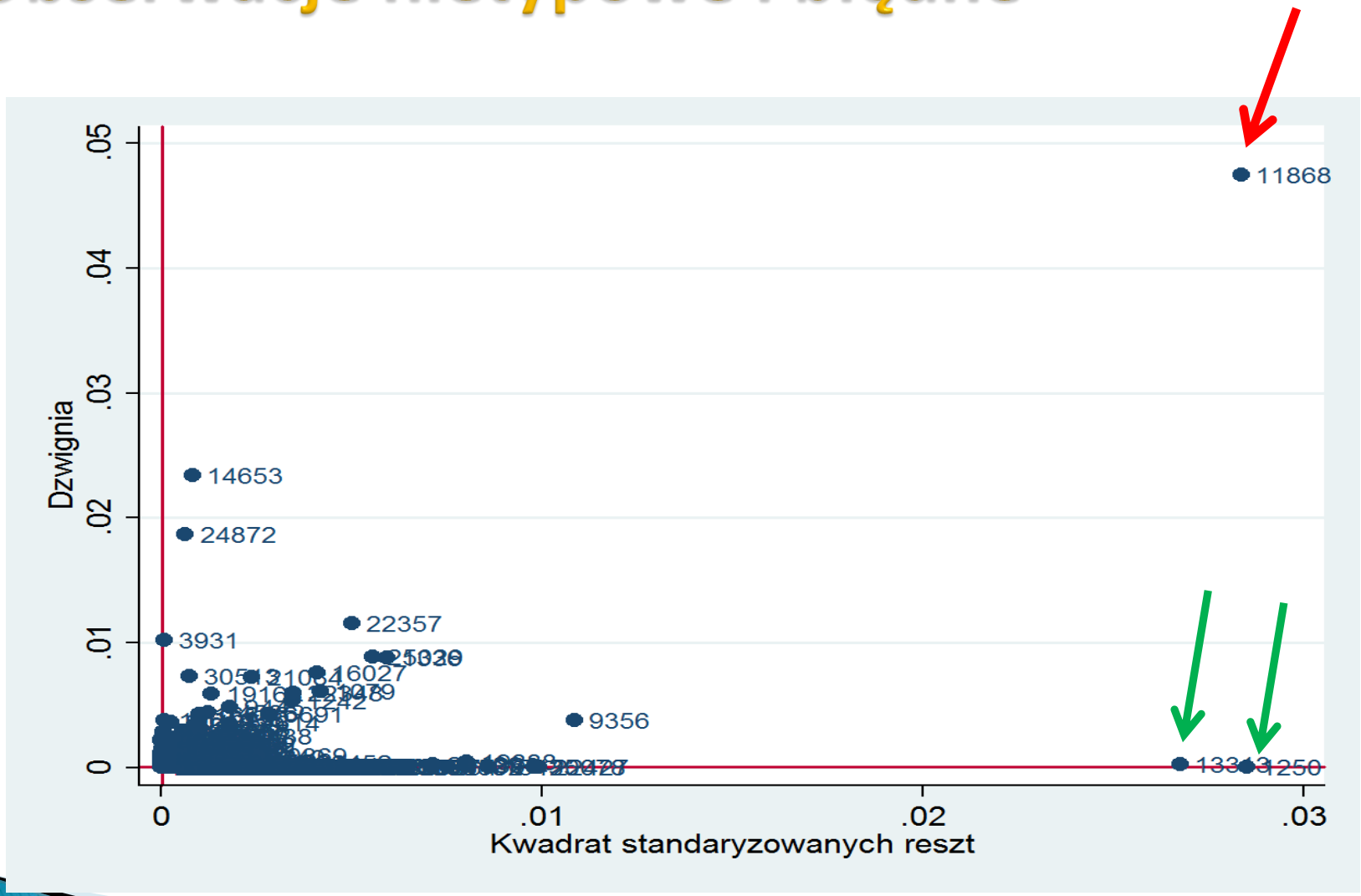
$$|\hat{e}_i| > 2$$

- 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 Cooka

- 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 dopasowane powstałe po usunięciu z próby *i* –tej obserwacji

Obserwacje nietypowe i błędne

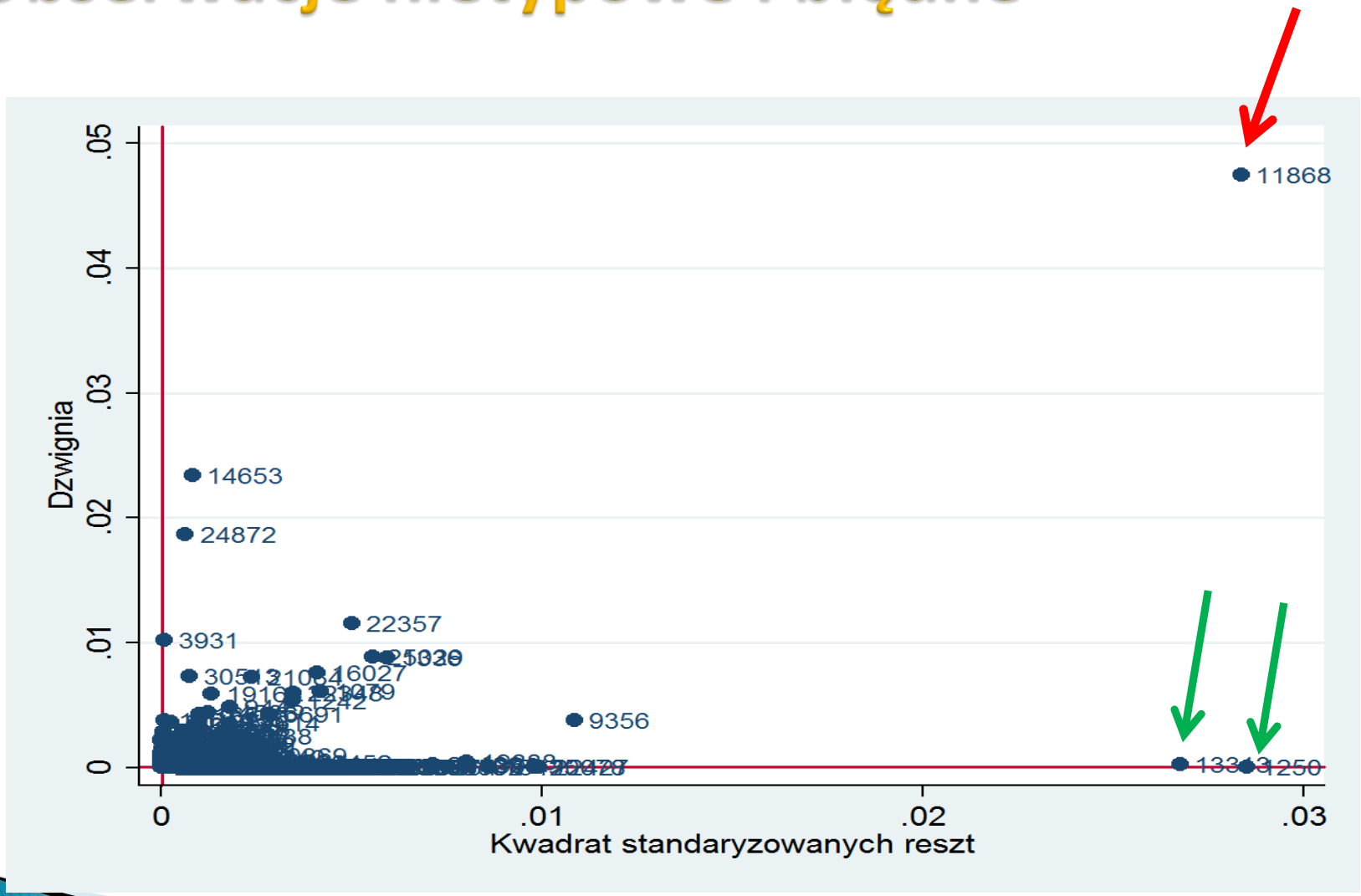
- **Odległość Cooka:**

- 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



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

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

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

Plan zajęć

1. Zmienne pominięte

2. Zmienne nieistotne

3. Obserwacje nietypowe i błędne

4. Współliniowość

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 ją z modelu

- Dokładna współliniowość jest wynikiem **błędnej specyfikacji modelu**

Współliniowość

- Przykład:

zmienne objaśniające w modelu na logarytmach:

a) *PKB*,

b) *Liczba ludności*

c) *PKB per capita*

- Zmienna *PKB per capita* jest kombinacją zmiennej *PKB* i *liczby ludności*

Współliniowość

- Przykład:

- dla wyjaśnienia mechanizmu zakupu dóbr trwałych w gospodarstwie domowym, zgodnie z hipoteza dochodów permanentnych Miltona Friedmana, za regresory wstawimy trzy wielkości:

  - 1. dochody,
  - 2. dochody permanentne (dochody trwale uzyskiwane) i
  - 3. dochody tranzytywne (przechodnie, okazjonalne),

- z definicji suma dochodów permanentnych i tranzytywnych jest równa kategorii dochodów, co spowoduje, że kolumny obserwacji na trzech kategoriach dochodów są dokładnie liniowo zależne.

Współliniowość

- O **niedokładnej współliniowości** mówimy, gdy występuje silna korelacja między zmiennymi objaśniającymi
  - Na przykład przy szacowaniu **płacy** jako funkcji ***wykształcenia, płci, wieku, stażu pracy*** możemy oczekiwać, że wiek badanej osoby i jej staż pracy wykażą silna dodatnia korelacje.

- 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

Przykład

Variable	VIF	1/VIF
wiek_2	41.82	0.023911
wiek	41.81	0.023916
miasto_sre~e	2.15	0.464424
miasto_male	1.88	0.532160
miasto_duze	1.29	0.777298
plec	1.01	0.987011
Mean VIF	14.99	

Pytania teoretyczne

1. Co to jest obserwacja nietypowa? Kiedy obserwację nietypową można uznać za błędną?
2. W jakim przypadku obserwacja nietypowa będzie miała znaczący wpływ na wynik regresji?
3. Jakich statystyk używamy do wykrywania obserwacji nietypowych i błędnych?
4. Kiedy mówimy, że zmienne w modelu są dokładnie współliniowe? Jak można rozwiązać ten problem?
5. Jakie są konsekwencje niedokładnej współliniowości? Za pomocą jakiej statystyki można wykryć niedokładną współliniowość w modelu?

Dziękuję za uwagę