Problemy z danymi (cz. I)

Natalia Nehrebecka
Stanisław Cichocki

Wykład 12

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

- **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

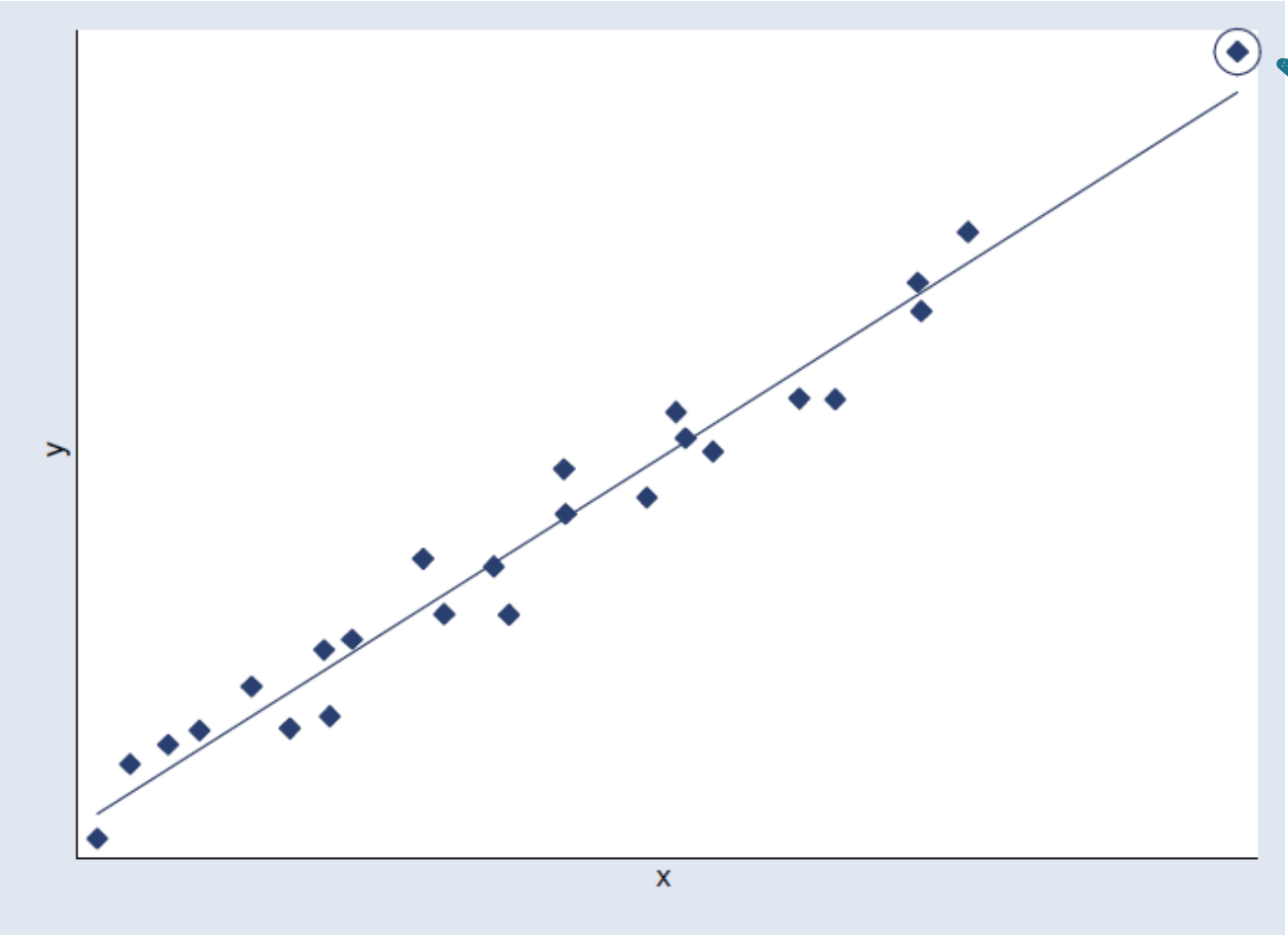
- Przykład:
  - Estymujemy **krzywą popytu na żywność dla różnych państw na świecie**.
  - W próbie występują państwa, w których obowiązuje reglamentacja żywności.
  - Obserwacje takie traktujemy jako obserwacje błędne – teoria opisująca krzywą popytu nie znajduje zastosowania w momencie nierynkowego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

Obserwacje nietypowe i błędne

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

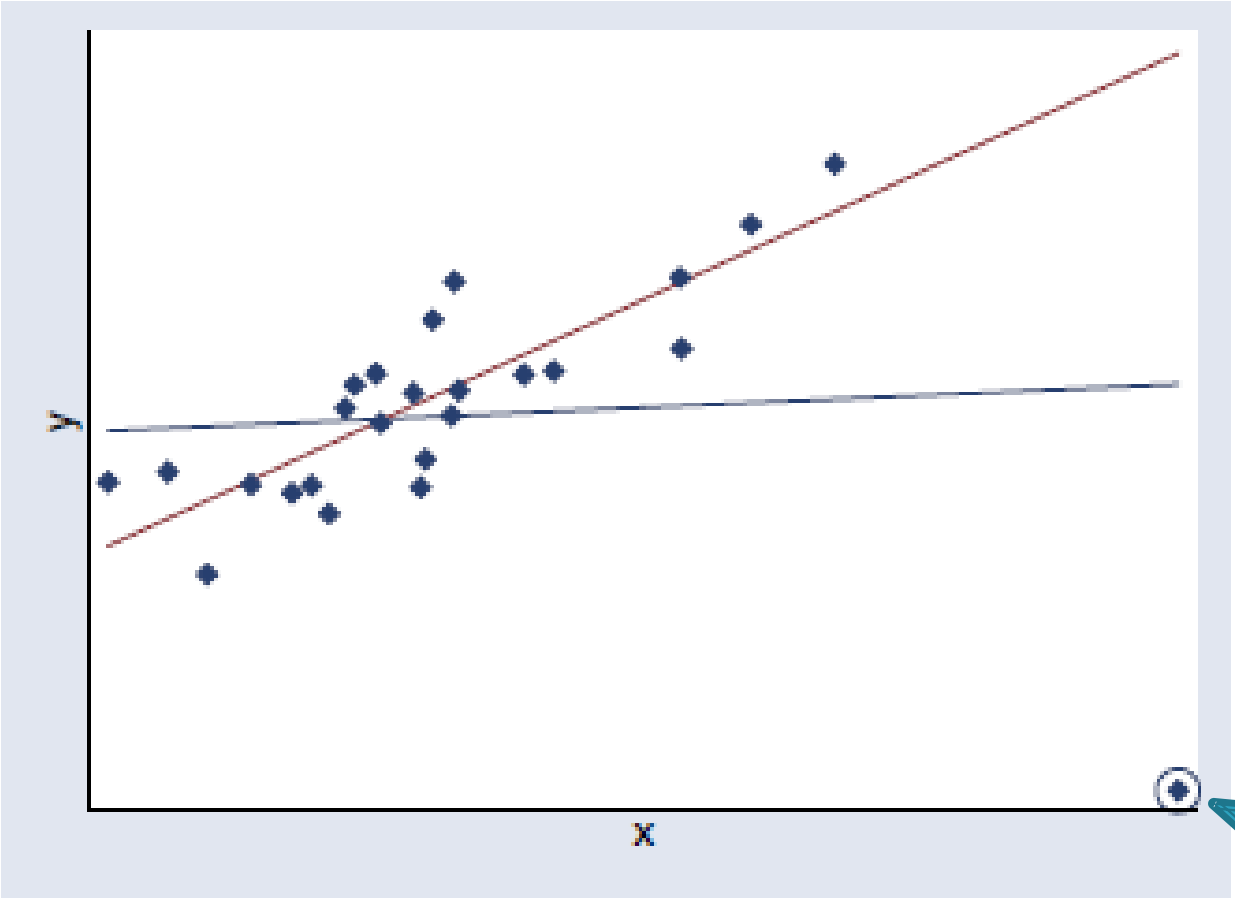
- Najbardziej niepokojąca jest sytuacja gdy obserwacja ma **nietypowe wartości dla zmiennych niezależnych i słabo pasuje do prostej regresji**

Obserwacje nietypowe i błędne



Obserwacja nietypowa pasująca do linii regresji

Obserwacje nietypowe i błędne



Obserwacja nietypowa niepasująca do linii regresji

Obserwacje nietypowe i błędne

- 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

- **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

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:

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

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 $\boldsymbol{X'X}$ 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

- 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

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