# Problemy z danymi

**Stanisław Cichocki**

**Natalia Nehrebecka**

Zajęcia 9

# Plan zajęć

1. Zmienne pominięte

2. Zmienne nieistotne

3. Obserwacje nietypowe i błędne

4. Współliniowość

# 1. Zmienne pominięte

»

# 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 $y$ ⟶ 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)

- Załóżmy, 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

- 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

- 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 | Korelacja między zmienną pominiętą a zmienną niezależną | Znak obciążenia |
|---|---|---|---|
| I | + | + | + (przeszacowanie) |
| II | - | - | + |
| III | + | - | - (niedoszacowanie) |
| IV | - | + | - |

# 2. Zmienne nieistotne

»

# 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 $\beta_1$ - 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

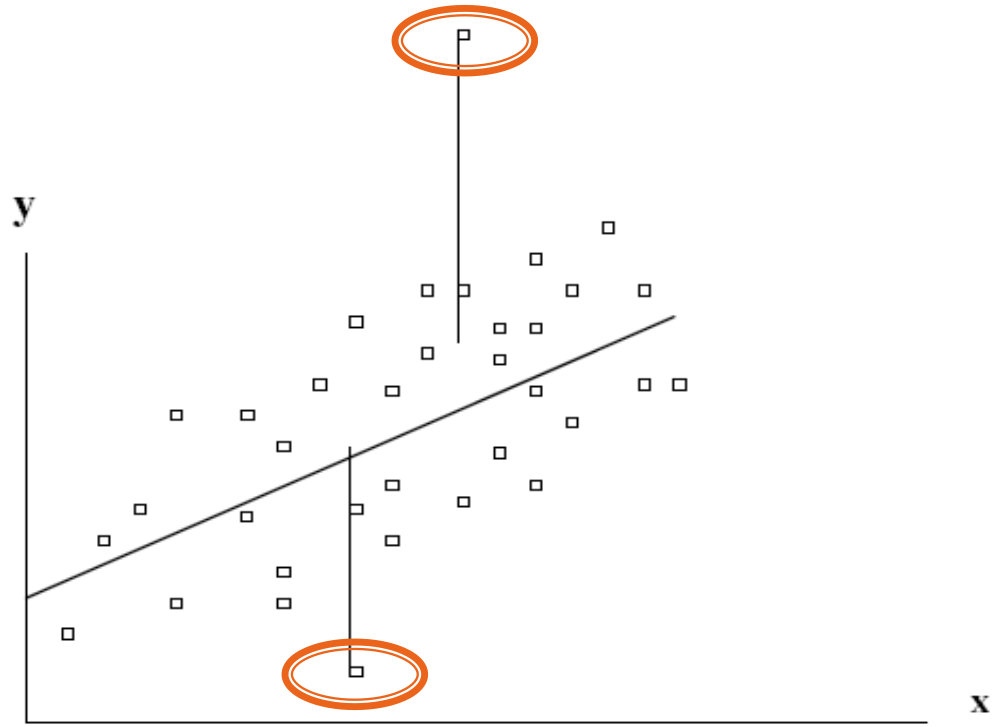
# 3. Obserwacje nietypowe i błędne

»

# Wykrywanie obserwacji nietypowych

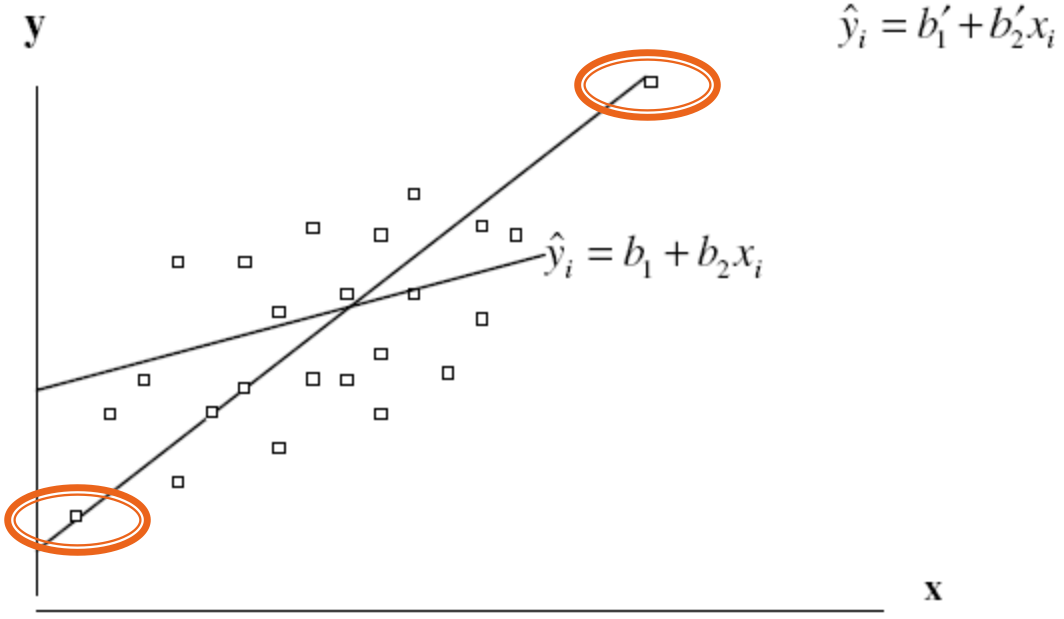
Można wyróżnić następujące rodzaje obserwacji nietypowych:

- 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

- 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

- 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

- 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

# 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 dźwigni).

- 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.

# Obserwacje nietypowe i błędne

- Do stwierdzenia, czy zmienna jest nietypowa na tle pozostałych zmiennych używamy następujących statystyk:

a) Dźwignia (leverage):

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

gdzie

$$\delta_i = [0, \ldots, 0, 1, 0, \ldots, 0]'$$

# 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 posiada duże reszty, aby się o tym przekonać musimy przyjrzeć się **standaryzowanym resztom**

# Obserwacje nietypowe i błędne

b) Standaryzowane reszty:

$$\hat{e}_i = \frac{e_i}{s\sqrt{1-h_i}}$$

- Uznaje się, że dla nietypowej obserwacji: $\left|\hat{e}_i\right| > 2$

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

# Obserwacje nietypowe i błędne

- 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:

$$CD_i = \frac{\hat{e}_i^2}{K} \frac{h_i}{1-h_i}$$

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

# Obserwacje nietypowe i błędne

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

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

# Obserwacje nietypowe i błędne

- Test wpływowych obserwacji **DFITS** :

$$DFITS_i = \hat{e}_i \sqrt{\frac{h_i}{1-h_i}}$$

# Obserwacje nietypowe i błędne

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

$$DFITS_i > 2 * \sqrt{\frac{K}{N}}$$

# Przykład

- R - stopa przestępczości; liczba przestępstw zanotowanych przez policje na 1 000 000 osób
- Age - liczba mężczyzn w wieku 14-24 lata na 1 000 mieszkańców
- Ed - indeks mierżący wykształcenie, średnia liczba lat nauki dla osób powyżej 25 roku przemnożona przez 10
- Exo - wydatki na policję
- U2 - stopa bezrobocia mężczyzn w wieku 35-39 na 1 000 mieszkańców
- W - "bogactwo" wyrażone jako mediana zasobów majątkowych rodzin
- X - liczba rodzin na 1 000, których zarobki są poniżej 1/4 mediany dochodów

# Przykład

gretl: model 15

Plik Edycja Testy Zapisz Wykresy Analiza LaTeX

```
Model 15: Estymacja KMNK z wykorzystaniem 47 obserwacji 1-47
Zmienna zależna: r

            współczynnik  błąd standardowy  t-Student  wartość p
  ------------------------------------------------------------------
  const      -618,503         108,246          -5,714    1,19E-06  ***
  w             0,159565        0,0939003        1,699    0,0970    *
  age           1,12518         0,350864         3,207    0,0026    ***
  ed            1,81786         0,480267         3,785    0,0005    ***
  ex0           1,05069         0,175224         5,996    4,78E-07  ***
  u2            0,828169        0,427399         1,938    0,0597    *
  x             0,823571        0,181490         4,538    5,10E-05  ***

  Średnia  arytmetyczna  zmiennej zależnej = 90,5085
  Odchylenie standardowe zmiennej zależnej = 38,6763
  Suma  kwadratów  reszt = 17351,1
  Błąd standardowy reszt = 20,8273
  Wsp. determinacji R-kwadrat = 0,74784
  Skorygowany  wsp. R-kwadrat = 0,71001
  Statystyka F (6, 40) = 19,7714 (wartość p < 0,00001)
  Logarytm wiarygodności = -205,605
  Kryterium informacyjne Akaike'a (AIC) = 425,21
  Kryterium bayesowskie Schwarza  (BIC) = 438,161
  Kryterium infor. Hannana-Quinna (HQC) = 430,083
```

# Przykład

$$DFFITS_i = 2 * \sqrt{K/N}$$

$$= 2 * \sqrt{7/47} = \pm 0{,}772$$

gretl: obserwacje dźwigniowe (leverage) i wpływowe (influence)

| | reszty u | leverage 0<=h<=1 | influence u*h/(1-h) | DFFITS |
|---|---|---|---|---|
| 1 | -10,44 | 0,099 | -1,1491 | -0,174 |
| 2 | 28,998 | 0,078 | 2,4668 | 0,429 |
| 3 | 23,492 | 0,210 | 6,2524 | 0,660 |
| 4 | 8,6422 | 0,214 | 2,3509 | 0,242 |
| 5 | -3,3273 | 0,139 | -0,5357 | -0,068 |
| 6 | -11,117 | 0,197 | -2,7254 | -0,293 |
| 7 | 15,205 | 0,109 | 1,8693 | 0,270 |
| 8 | 33,672 | 0,245 | 10,919 | 1,095 |
| 9 | 8,348 | 0,099 | 0,914 | 0,138 |
| 10 | -4,7377 | 0,093 | -0,48328 | -0,075 |
| 11 | 54,544 | 0,157 | 10,126 | 1,359 |
| 12 | 17,623 | 0,076 | 1,4528 | 0,252 |
| 13 | -21,494 | 0,137 | -3,403 | -0,443 |
| 14 | -8,0781 | 0,100 | -0,8984 | -0,135 |
| 15 | -8,4258 | 0,125 | -1,2029 | -0,162 |
| 16 | -2,2307 | 0,126 | -0,32118 | -0,043 |
| 17 | -1,2152 | 0,160 | -0,231 | -0,027 |
| 18 | -23,521 | 0,109 | -2,8639 | -0,420 |
| 19 | -37,518 | 0,113 | -4,7658 | -0,706 |
| 20 | 0,66408 | 0,217 | 0,18359 | 0,019 |
| 21 | 0,046354 | 0,098 | 0,0050633 | 0,001 |
| 22 | -26,841 | 0,195 | -6,4829 | -0,716 |
| 23 | 30,345 | 0,138 | 4,8637 | 0,640 |
| 24 | 9,1817 | 0,115 | 1,1915 | 0,167 |
| 25 | -8,8961 | 0,161 | -1,7027 | -0,202 |
| 26 | 15,648 | 0,215 | 4,2917 | 0,443 |
| 27 | 7,4682 | 0,212 | 2,012 | 0,207 |
| 28 | -6,6246 | 0,147 | -1,1379 | -0,141 |
| 29 | -38,306 | 0,307* | -16,999 | -1,551 |
| 30 | -1,3133 | 0,157 | -0,24485 | -0,029 |
| 31 | 1,3045 | 0,119 | 0,17665 | 0,024 |

# 4. Współliniowość

# Współliniwość

- 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

# Współliniwość

- 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

# Współliniwość

- 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)*

# Współliniwość

- 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

# Współliniwość

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

# Współliniwość

- 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

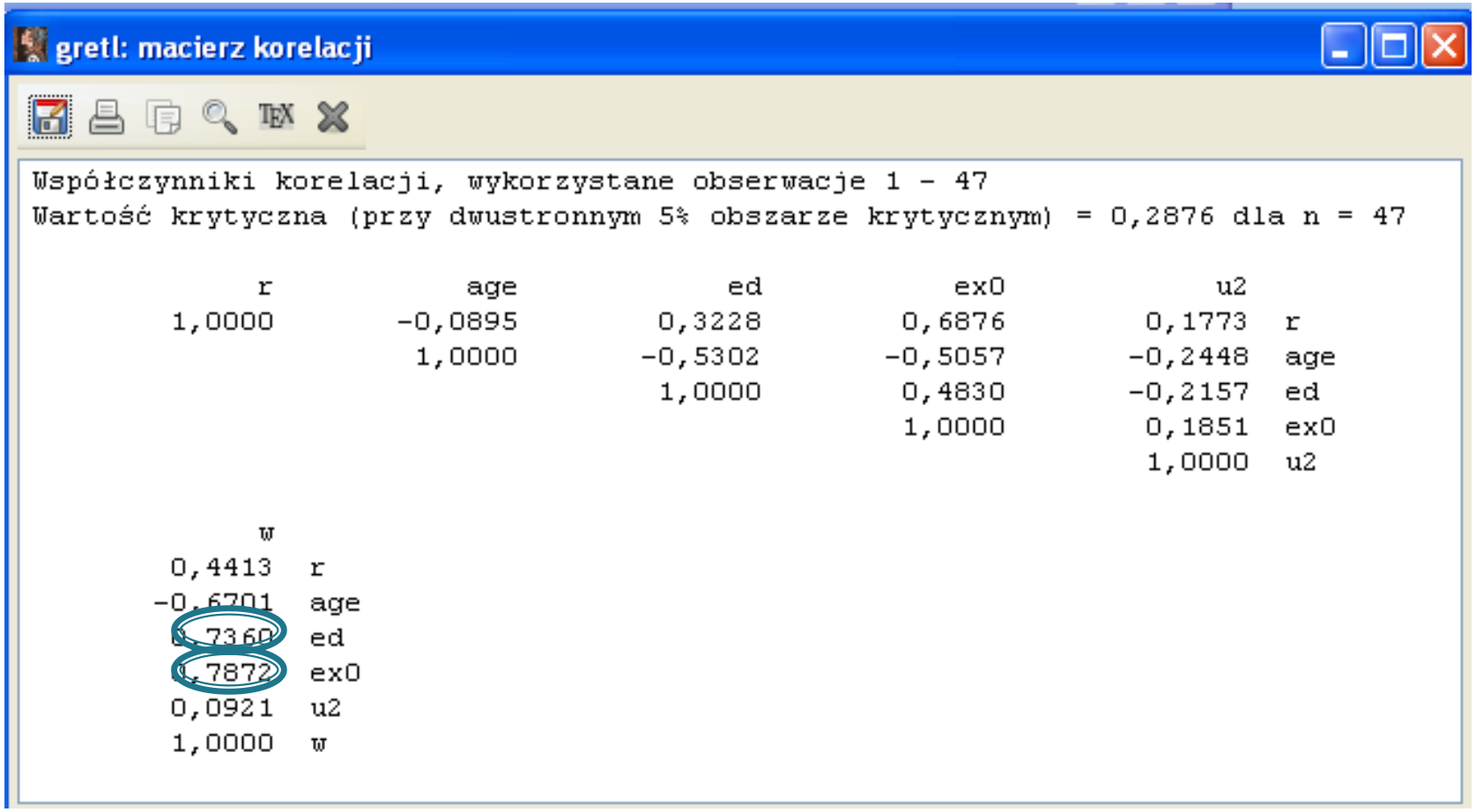
- Niedokładna współliniowość nie jest wynikiem błędnej specyfikacji modelu lecz wynika z własności konkretnego zbioru danych

# Przykład

gretl: ocena współliniowości VIF (Variance Inflation Factor)

```
Ocena współliniowości VIF - czynnika powiększania wariancji

 Minimalna możliwa wartość = 1.0
 Wartości > 10.0 mogą wskazywać na problem współliniowości-rozdęcia wariancji

  13)                w    8,706
   2)              age    2,062
   4)               ed    3,061
   5)              ex0    2,876
  12)               u2    1,382
  14)                x    5,560

VIF(j) = 1/(1 - R(j)^2), gdzie R(j) jest współczynnikiem korelacji wielorakiej
pomiędzy zmienną 'j' a pozostałymi zmiennymi niezależnymi modelu.

Własności macierzy X'X:

 1-norm = 27134910
 Wyznacznik = 9,7361297e+025
 Wskaźnik uwarunkowania macierzy CN = 1,3539368e-009
```

# Przykład


gretl: macierz korelacji
Współczynniki korelacji, wykorzystane obserwacje 1 - 47
Wartość krytyczna (przy dwustronnym 5% obszarze krytycznym) = 0,2876 dla n = 47
r age ed ex0 u2
1,0000 -0,0895 0,3228 0,6876 0,1773 r
1,0000 -0,5302 -0,5057 -0,2448 age
1,0000 0,4830 -0,2157 ed
1,0000 0,1851 ex0
1,0000 u2
w
0,4413 r
-0,6701 age
0,7360 ed
0,7872 ex0
0,0921 u2
1,0000 w

# Pytania teoretyczne

1. Jakie są skutki pominięcia w równaniu regresji istotnych zmiennych objaśniających?
2. Jakie są efekty dodania w równaniu regresji nieistotnych zmiennych objaśniających?
3. Kiedy mówimy, że zmienne w modelu są dokładnie współliniowe? Jak można rozwiązać ten problem?
4. Jakie są konsekwencje niedokładnej współliniowości? Za pomocą jakiej statystyki można wykryć niedokładną współliniowość w modelu?
5. Co to jest obserwacja nietypowa? Kiedy obserwację nietypową można uznać za błędną?
6. Jakich statystyk używamy do wykrywania obserwacji nietypowych i błędnych?

# Dziękuję za uwagę