Skrypt do przedmiotu

EKONOMETRIA I

AUTORZY:

Michał Rubaszek

Katarzyna Bech-Wysocka

Piotr Dybka

Michał Gradzewicz

Karolina Konopczak

Jakub Mućk

Karol Szafranek

Marcin Topolewski

Bartłomiej Wiśnicki

Zuzanna Wośko

Redakcja i koordynacja

Michał Rubaszek

2020

Szkoła Główna Handlowa w Warszawie

Spis treści

1 Wprowadzenie **1**

2 MNK **25**

3 Istotność zmiennych **51**

4 Specyfikacja modelu **67**

5 Współliniowość **87**

6 Heteroskedastyczność **105**

7 Autokorelacja **121**

8 Modele dynamiczne **139**

9 Niestacjonarność **155**

10 Prognoza **177**

11 Modele zmiennej jakościowej **201**

12 Endogeniczność **219**

13 Metoda zmiennych instrumentalnych **235**

14 Testy w MZI **249**

Wstęp

Ten skrypt zawiera materiały przygotowane przez wykładowców Szkoły Głównej Handlowej w Warszawie w celu prowadzenia przedmiotu "Ekonometria I".

Zajęcia są prowadzone z wykorzystaniem:

- darmowego pakietu ekonometrycznego GRETL:
  `http://gretl.sourceforge.net/`
- plików z danymi umieszczonych na stronie przedmiotu:
  `http://web.sgh.waw.pl/~mrubas/`

Materiały te w dużej mierze są oparte na:

- opracowaniu R.C. Hill, W.E. Griffiths i G.C. Lim "Principles of Econometrics"
  `https://www.principlesofeconometrics.com/`
- skryptu L. Adkins "Using gretl for Principles of Econometrics"
  `http://www.learneconometrics.com/gretl/index.html`

Temat 1

Wprowadzenie do ekonometrii

Katarzyna Bech-Wysocka i Piotr Dybka

- Czym się zajmuje ekonometria
- Model ekonometryczny
- Rodzaje danych
- Źródła danych
- Działania na macierzach
- Zmienna losowa
- Rozkład prawdopodobieństwa
- Rozkłady statystyczne
- Pakiety ekonometryczne: Gretl

Czym jest „Ekonometria”?

EKONOMETRIA

Zastosowanie **matematyki** i **statystyki** do analizy ilościowych związków zachodzących między obserwowanymi zmiennymi ekonomicznymi

Ciekawostki:
- słowo „Ekonometria” zostało wprowadzone do literatury przez Pawła Ciompę w pracy "Zarys ekonometryi i teorya buchalterii" opublikowanej w 1910 roku we Lwowie .
- Za ojców współczesnej ekonometrii uważa się laureatów nagrody Nobla z ekonomii: **Ragnara Frischa** i **Jana Tinbergena**.

DO CZEGO SŁUŻY EKONOMETRIA?

- Do **weryfikacji** hipotez ekonomicznych
- Do **kwantyfikacj**i siły zależności między zmiennymi
- Do **prognozowania**

Ekonometria polega na połączeniu teorii i danych ekonomicznych, finansowych, demograficznych itp. z narzędziami statystycznymi w celu odpowiedzi na pytanie „**ile**?”.

Przykłady wyzwań dla ekonometryka

1. Rada miasta zastanawia się, jak zmniejszyć się przestępczość.
   Pytanie: jak liczba policjantów wpływa na przestępczość?

2. Właściciel restauracji zastanawia się jaką kwotę wydać na reklamę w lokalnej gazecie.
   Pytanie: jak wydatki na reklamę wpływają na liczbę klientów?

3. Uniwersytet planuje zwiększyć opłaty za czesne.
   Pytanie: jak wyższy poziom czesnego wpłynie na liczbę studentów w kolejnych latach?

4. Firma kosmetyczna zastanawia się nad budową nowej fabryki.
   Pytanie: jaka jest prognozowana wartość popytu w najbliższej dekadzie?

Czym się różni Ekonometria od data science?

EKONOMETRIA a DATA SCIENCE

Obie dziedziny istotnie się przenikają, jednak:

- Ekonometria w większym stopniu koncentruje się na badaniu zależności przyczynowo-skutkowych
- Inżynieria danych (data science) na poszukiwaniu zależności symptomatycznych, np. korelacyjnych

Ponadto w przypadku modeli ekonometrycznych, badacz określa formalną strukturę modelu (specyfikację), natomiast data science obejmuje także metody w których nie ma zmiennej objaśnianej, np. problemy klasyfikacyjne, w których celem jest podział danych na grupy o podobnym profilu (metody uczenia nienadzorowanego).

A co jeśli chciałbym lepiej poznać metody związane z inżynierią danych?

Ciekawe wprowadzenie można znaleźć na stronie:
https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/

Model ekonometryczny a model ekonomiczny

Ekonometrycy i ekonomiści inaczej zapisują zależność między zmiennymi, czyli tzw. **model**

Rozważmy zależność między konsumpcją (C) i dochodem (Y):

Ekonomista: model opisuje deterministyczną zależność między zmiennymi :

$$C = a + bY$$

Ekonometryk: model opisuje stochastyczną zależność między zmiennymi:

$$C_i = \alpha + \beta Y_i + \varepsilon_i$$

poprzez dodanie składnika losowego ε.

Dodatkowo, podkreślane jest, że model ekonometryczny dotyczy każdej obserwacji $i = 1,2, \dots, N$

Ekonometria pozwala sprawdzić na ile teoretyczne zależności, np. opisane przez modele ekonomiczne, dobrze opisują obserwowane zjawiska. W tym celu potrzebujemy stworzyć **zbiór danych**.

Model ekonometryczny

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. Weryfikacja
6. Zastosowanie

Specyfikacja liniowego modelu ekonometrycznego

Model ekonometryczny:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \text{ dla } i = 1,2,\dots,N$$

y_i	zmienna zależna / objaśniana (dependent variable)
x_{ki}	zmienne niezależne / regresory / zmienne objaśniające (explanatory variables)
ε_i	składnik losowy (error lub disturbance term)
β_k	(nieznane) parametry strukturalne

Indeks dolny i wskazuje, że model jest prawdziwy dla każdej obserwacji $i = 1,2,\dots,N$.

Model empiryczny (po oszacowaniu parametrów):

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_K} x_{Ki}$$

$\hat{y}_i$	wartości teoretyczna z modelu
$\widehat{\beta_k}$	(znane) oszacowania parametrów strukturalnych

UWAGA: aby uzyskać model empiryczny należy zebrać obserwacje dla y_i oraz x_{ki}

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. **Zebranie danych**
4. Estymacja
5. Weryfikacja
6. Zastosowanie

Rodzaje danych

Kryterium pochodzenia:

- Dane mikroekonomiczne (np. dochody gospodarstw domowych, przychody firm)
- Dane makroekonomiczne (poziom PKB, stopa bezrobocia)
- Dane ankietowe (np. preferencje polityczne)
- Dane eksperymentalne (wyniki działania szczepionki)

Kryterium typu danych:

- Jakościowe (np. płeć)
- Ilościowe (np. dochód)
- Tekstowe (np. adres zamieszkania)

Kryterium obserwacji:

- Dane przekrojowe (PKB w krajach UE)
- Szeregi czasowe (PKB w Polsce w okresie 1995-2020)
- Dane panelowe (PKB w krajach UE w latach 1995-2020)

Przykład 1.1. Rodzaje danych

PKB per capita

Dane przekrojowe:

- Obserwacje dla różnych podmiotów z tego samego okresu
- Indeksowanie: y_i dla $i = 1,2,\dots,N$

time	Poland
2006	7200
2007	8200
2008	9600
2009	8200
2010	9400
2011	9900
2012	10100
2013	10300
2014	10700
2015	11200
2016	11100
2017	12200

Szeregi czasowe:

- Obserwacje dla tego samego podmiotu z różnych okresów
- Indeksowanie: y_t dla $t = 1,2,\dots,T$

geo	2017
Belgium	38700
Bulgaria	7300
Czechia	18100
Denmark	50100
Germany	39600
Estonia	18000
Ireland	61200
Greece	16700
Spain	25100
France	34200
Croatia	11800
Italy	28500
Cyprus	22800
Latvia	13900
Lithuania	14900
Luxembou	92600
Hungary	12700
Malta	23800
Netherland	43000
Austria	42100
Poland	12200
Portugal	18900
Romania	9600
Slovenia	20800
Slovakia	15600
...	
Norway	67100

Dane panelowe:

- Obserwacje dla różnych podmiotów z różnych okresów
- Indeksowanie: y_{it} dla $i = 1,2,\dots,N$ oraz $t = 1,2,\dots,T$

geo\time	2006	2007	2008	2009	2010	...	2014	2015	2016	2017
Belgium	31000	32400	33100	32300	33500		35800	36600	37600	38700
Bulgaria	3500	4200	4900	4900	5100		5900	6300	6800	7300
Czechia	12100	13400	15500	14200	14900		14900	16000	16700	18100
Denmark	41500	42700	44000	41900	43800		47100	47800	48400	50100
Germany	29500	31000	31700	30600	32100		36300	37300	38400	39600
...										
Norway	59100	62300	66500	57700	66300		73300	67100	64100	67100

Źródła danych

Eurostat: https://ec.europa.eu/eurostat/data/database
Dane dla krajów Unii Europejskiej

Przydatne pojęcia:

1. Sekcja *National accounts*:
   dane o rachunkach narodowych, np. PKB
2. Sekcja *Balance of payments*:
   dane o inwestycjach zagr. czy wymianie handlowej
3. Sekcja P*opulation and social conditions*:
   dane o bezrobociu i wynagrodzeniach

Po otwarciu wybranej bazy w prawym górnym rogu mamy do wyboru następujące opcje:

Jeżeli wielokrotnie korzystamy z danej bazy, warto korzystać z opcji bookmark – tworzy ona trwały link do tej bazy wraz z naszymi ustawieniami (wyboru państwa, zakresu danych itp.).

Źródła danych

The Economic Network's website:
http://www.economicsnetwork.ac.uk

Penn World Tables:
https://pwt.sas.upenn.edu

FRED – Federal Reserve Economic Data:
https://fred.stlouisfed.org/

Bank Światowy (World Bank):
http://data.worldbank.org

OECD:
http://www.oecd-ilibrary.org/statistics

Organizacja Narodów Zjednoczonych (UNCTAD):
https://unctadstat.unctad.org

Przykład 1.2. Model z jedną zmienną objaśniającą

Model ekonometryczny:

$$score_i = \beta_0 + \beta_1 hours_i + \varepsilon_i,$$

gdzie $score_i$ to ocena z egzaminu z Ekonometrii I, a $hours_i$ to liczba godzin spędzonych na powtórzeniu materiału przed egzaminem.

Model empiryczny (po oszacowaniu parametrów):

$$\widehat{score}_i = 12{,}0 + 4{,}5 hours_i$$

Pytania:
- Jakie dane są potrzebne do oszacowania parametrów modelu?
- Czy uwzględniono wszystkie czynniki wpływające na ocenę z ekonometrii?
- Jaka jest interpretacja oszacowań $\widehat{\beta_0} = 12{,}0$ oraz $\widehat{\beta_1} = 4{,}5$?

Przykład 1.3. Model z wieloma zmiennymi objaśniającymi

Model ekonometryczny:

$$score_i = \beta_0 + \beta_1 hours_i + \beta_2 sex_i + \beta_3 IQ_i + \varepsilon_i,$$

gdzie $score_i$ to ocena z egzaminu z Ekonometrii I, $hours_i$ to liczba godzin spędzonych na powtórzeniu materiału przed egzaminem, sex_i to płeć (1 dla kobiet), a IQ_i to miara ilorazu inteligencji.

Model empiryczny (po oszacowaniu parametrów):

$$\widehat{score}_i = -48{,}0 + 5{,}0 hours_i + 0{,}7 sex_i + 0{,}6 IQ_i$$

Pytania:
- Jakiego typu dane występują w modelu?
- Czy teraz uwzględniono wszystkie czynniki wpływające na ocenę z ekonometrii?
- Dlaczego wartość oszacowania $\widehat{\beta_0}$ tak mocno się zmieniła w porównaniu z przykładem 1.2?

Specyfikacja modelu: zapis macierzowy

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \text{ dla } i = 1,2,\ldots,N$$

W trakcie zajęć powyższy model będziemy zapisywali w postaci macierzowej:

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{K1} \\ 1 & x_{12} & \ldots & x_{K2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1N} & \ldots & x_{KN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_N \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{y}$	wektor $N \times 1$ obserwacji zmiennej zależnej
$\mathbf{X}$	macierz $N \times (K+1)$ zmiennych objaśniających
$\boldsymbol{\varepsilon}$	wektor $N \times 1$ składników losowych
$\boldsymbol{\beta}$	wektor $(K+1) \times 1$ parametrów strukturalnych
N	liczba obserwacji (T dla szeregów czasowych)
$K+1$	liczba parametrów

Działania na macierzach

Macierze

Macierz

Macierz *A* to zbiór elementów ułożonych w m wierszach i n kolumnach.
Wtedy mówimy, że macierz A jest o wymiarze $m \times n$.

Rodzaje macierzy

Niech *A* będzie macierzą o wymiarze $m \times n$.

- A jest macierzą kwadratową jeżeli $m = n$
- A jest macierzą symetryczną jeżeli jest macierzą kwadratową oraz $a_{ij} = a_{ji}$ dla każdej pary (i, j). Innymi słowy, $A = A'$ (o transpozycji za chwilę)
- A jest macierzą jednostkową, w zapisie $A = I_n$, jeżeli jest macierzą kwadratową, elementy na głównej przekątnej są równe $a_{ii} = 1$, oraz elementy poza główną przekątną wynoszą $a_{ij} = 0$ dla $i \neq j$.

UWAGA: dla macierzy *B* o odpowiednich wymiarach $I_n B = B I_n = B$.

Macierze

Proste operacje na macierzach

Niech A i B będą macierzami wymiarach $m \times n$. Wtedy:

- **Suma** $C = A + B$ jest macierzą o wymiarze $m \times n$ z elementami $c_{ij} = a_{ij} + b_{ij}$
- **Iloczyn ze skalarem λ** $C = \lambda A$ jest macierzą $m \times n$ o elementach $c_{ij} = \lambda a_{ij}$
- **Transpozycja** $C = A'$ to macierz o wymiarach $n \times m$ powstała przez zmianę wierszy w kolumny, a kolumn w wiersze. Własności:

$$(A + B)' = A' + B'$$
$$(A')' = A$$
$$(\lambda A)' = \lambda A'$$
$$(AB)' = B'A'$$

Iloczyn macierzy

Niech *A* i *B* będą macierzami o wymiarach odpowiednio $m \times n$ i $\mathrm{n} \times p$.
Iloczyn $C = AB$ jest macierzą o wymiarach $m \times p$ i elementach:

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} ,$$

gdzie $i = 1, \dots, m$ oraz $k = 1, \dots, p$.
UWAGA: $AB \neq BA$

Macierze

Macierz odwrotna

Niech A będzie kwadratową macierzą o wymiarach $n \times n$. Macierz B jest macierzą odwrotną do A jeżeli $AB = BA = I_n$. Jeżeli taka macierz B istnieje, to macierz A jest odwracalna. Warunkiem odwracalności jest niezerowa wartość wyznacznika ($|A| \neq 0$). Warto dodać, że istnieje co najwyżej jedna macierz odwrotna. Jej własności są następujące:

$$(A^{-1})^{-1} = A$$
$$(A')^{-1} = (A^{-1})'$$
$$(AB)^{-1} = B^{-1}A^{-1}$$

Niezależność liniowa, czyli rząd macierzy

- Zbiór wektorów jest **liniowo niezależny**, jeżeli żadnego z nich nie można przedstawić jako liniowej kombinacji pozostałych wektorów.
- Dla macierzy A o wymiarach $m \times n$ rząd wierszowy ($raw\ rank$) opisuje liczbę liniowo niezależnych wierszy, zaś rząd kolumnowy ($column\ rank$) liczbę liniowo niezależnych kolumn
- Odwracalna macierz kwadratowa $n \times n$ musi mieć pełny rząd (*full rank*), czyli rząd wierszowy oraz rząd kolumnowy wynosi n

Podstawy statystyki

Zmienna losowa

W modelu ekonometrycznym składnik losowy ε jest zmienną losową. Co to oznacza?

Nieformalna definicja zmiennej losowej.
Zmienna, której wartości nie znamy, dopóki tej wartości nie zaobserwujemy.

Ilustracja:

- Temperatura przy wejściu do budynku G o godz. 12:00 1 stycznia 2030 r. (zmienna losowa)
- Temperatura przy wejściu do budynku G o godz. 12:00 1 stycznia 2020 r. (realizacja)
- Długość dnia 1 stycznia 2030 r. (zmienna deterministyczna)

Formalna definicja zmiennej losowej.

Zmienne losowe to funkcje mierzalne względem przestrzeni probabilistycznych, które przypisują zdarzeniom elementarnym wartości liczbowe (prawdopodobieństwa)

Podział zmiennych losowych ze względu na zbiór zdarzeń elementarnych:

- **Dyskretna zmienna losowa:** przyjmuje skończoną liczbę wartości, szczególnym przypadkiem jest **zmienna binarna** przyjmująca wartości 0 i 1
- **Ciągła zmienna losowa:** przyjmuje nieskończoną liczbę wartości (zazwyczaj ze zbioru liczb rzeczywistych)

Zmienne losowe: rozkład prawdopodobieństwa

Prawdopodobieństwo występowania poszczególnych wartości zmiennej losowej jest opisane przez:

- **funkcję prawdopodobieństwa** (dla zmiennych dyskretnych, ang. *probability mass function*)
- **funkcję gęstości prawdopodobieństwa** (dla zmiennych ciągłych, ang. *probability density function*)

Funkcja prawdopodobieństwa

Dla dyskretnej zmiennej losowej X wartość **funkcji prawdopodobieństwa** wynosi:

$$f(x) = P(X = x)$$

Funkcja ta przyjmuje niezerowe wartości jedynie dla n możliwych realizacji zmiennej X:

$$f(x_i) = p_i \geq 0 \text{ dla } i = 1,2,\dots,n$$

Dodatkowo, zachodzi warunek:

$$\sum_{i=1}^{n} p_i = 1$$

Funkcja gęstości prawdopodobieństwa

Dla ciągłej zmiennej losowej X nie określamy prawdopodobieństwa pojedynczych wydarzeń, ale prawdopodobieństwo, że zmienna losowa znajduje się w pewnym przedziale:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

gdzie $f(x)$ to **funkcja gęstości prawdopodobieństwa**, która spełnia warunek

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Zmienne losowe: rozkład prawdopodobieństwa

Dystrybuanta

Dystrybuanta (ang. *cumulative distribution function*) zmiennej losowej X, oznaczana jako $F(a)$, jest zdefiniowana jako prawdopodobieństwo, że X jest mniejsze bądź równe określonej wartości a:

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

Warto zauważyć, że:

- $F(a)$ jest funkcją niemalejącą
- $F(a) \in [0,1]$
- $\int_a^b f(x)dx = F(b) - F(a)$
- $F'(x) = f(x)$.

Zależności między zmiennymi

Dla dwóch zmiennych losowych X i Y:

- **Rozkład łączny** określa prawdopodobieństwo wystąpienia dwóch zdarzeń jednocześnie:

$$f_{X,Y}(x,y) = P(X = x, Y = y).$$

- **Rozkład brzegowy (bezwarunkowy)** opisuje prawdopodobieństwo dla indywidualnych zmiennych:

$f_X(x) = P(X = x) = \sum_y f_{X,Y}(x,y)$ [zmienne dyskretne]

$f_X(x) = \int f(x,y)dy$ [zmienne ciągłe]

- **Rozkład warunkowy** opisuje prawdopodobieństwo wystąpienia zdarzenia $X = x$ pod warunkiem, że wystąpiło zdarzenie $Y = y$:.

$$f(x|y) = P(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- Dwie zmienne losowe są **statystycznie niezależne**, jeżeli rozkład warunkowy jest taki sam jak rozkład brzegowy:

$$f(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = f_X(x) \quad \Leftrightarrow \quad f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Momenty rozkładu prawdopodobieństwa

Wartość oczekiwana

Definicja **wartości oczekiwanej** zmiennej losowej X:

$$\mu = E(X) = \sum_i x_i f(x_i) \quad \text{[dyskretna zmienna losowa]}$$

$$\mu = E(X) = \int x f(x) dx \quad \text{[ciągła zmienna losowa]}$$

Interpretacja: jakiej wartości X oczekujemy przed zaobserwowaniem jej realizacji
Ważne: wartość oczekiwana μ to nie to samo co średnia w próbie ($\hat{\mu} = \bar{x}$), którą możemy policzyć dopiero po zaobserwowaniu realizacji

Warunkowa wartość oczekiwana to:

$$\mu_{X|Y} = E(X|Y = y) = \sum x_i f(x_i|y) \quad \text{[dyskretna zmienna losowa]}$$

$$\mu_{X|Y} = E(X|Y = y) = \int x f(x|y) dx \quad \text{[ciągła zmienna losowa]}$$

Interpretacja: jakiej wartości X oczekujemy przed zaobserwowaniem jej realizacji, jeżeli posiadamy dodatkową informację, a mianowicie, że $Y = y$

Momenty rozkładu prawdopodobieństwa

Wariancja

Wariancja zmiennej losowej *X* (dyskretnej lub ciągłej) to:

$$Var(X) = \sigma_X^2 = E[X - E(X)]^2 = \int f(x)(x - \mu)^2 \, dx$$

$$Var(X) = \sum_i (x_i - \mu)^2 f(x_i) \quad \text{[dyskretna zmienna losowa]}$$

$$Var(X) = \int (x - \mu)^2 f(x) dx \quad \text{[ciągła zmienna losowa]}$$

Przydatny wzór:

$$Var(X) = E([X^2]) - (E[X])^2.$$

Pierwiastek (kwadratowy) z wariancji nazywamy **odchyleniem standardowym**, σ_X

Kowariancja i korelacja

Kowariancja między zmiennymi X i Y to:

$$cov(X, Y) = \sigma_{XY} = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$$

Korelacja między zmiennymi to natomiast:

$$corr(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

WAŻNE: brak korelacji nie oznacza niezależności!

Momenty rozkładu prawdopodobieństwa

Wybrane reguły

- Mnożenie przez skalar / dodawanie skalaru:

$$E(aX+b) = aE(X) + b$$

$$Var(aX+b) = a^2 Var(X)$$

- Suma zmiennych losowych:

$$E(X+Y) = E(X) + E(Y)$$

$$Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y)$$

- Jeżeli $g(X)$ jest funkcją zmiennej losowej X, to też jest zmienną losową:

$$E[g(X)] = \sum_i g(x_i) f(x_i) \qquad \text{[dyskretna zmienna losowa]}$$

$$E[g(X)] = \int g(x) f(x) dx \qquad \text{[ciągła zmienna losowa]}$$

- Prawo iteracyjnych oczekiwań (law of iterated expectations, szerzej w Temat 12):

$$E(X) = E\,[E(X|Y)]$$

$$E(XY) = E_X[XE(Y|X)]$$

Rozkłady statystyczne

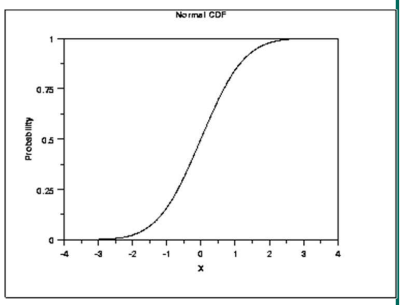
Rozkład normalny

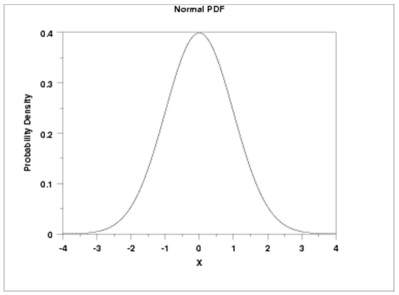
O zmiennej losowej X mówimy, że ma rozkład normalny o $E(X) = \mu$ i $Var(X) = \sigma^2$:

$$X \sim N(\mu, \sigma^2)$$

jeżeli funkcja gęstości wynosi:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{\sigma^2}\right\}$$





Standaryzacja do rozkładu $N(0,1)$ polega na:

$$Z = \frac{X-\mu}{\sigma}$$

Obliczanie prawdopodobieństwa zdarzenia:

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

gdzie Φ oznacza dystrybuantę rozkładu $N(0,1)$.

Rozkłady statystyczne

Rozkład normalny – przedział ufności

Dla zmiennej $X \sim N(\mu, \sigma^2)$:

- przedział $\mu \pm \sigma$ zawiera około 68% obserwacji
- przedział $\mu \pm 2\sigma$ zawiera około 95% obserwacji
- przedział $\mu \pm 3\sigma$ zawiera prawie wszystkie obserwacje

Załóżmy, że procentowy wynik testu z matematyki dla studentów pierwszego roku ma rozkład normalny o wartości oczekiwanej $\mu = 64$ i odchyleniu standardowym $\sigma = 10$. Oznacza to, że:

- około 68% studentów uzyskało między 54 a 74 pkt.
- około 95% studentów uzyskało między 44 a 84 pkt.
- prawie wszyscy studenci uzyskali wynik między 34 a 94 pkt.

Pytanie: A co jeżeli empiryczne wyniki testu są inne?

Popularne rozkłady

Rozklad chi-kwadrat (χ^2)

Dla niezależnych zmiennych $X_i \sim N(0,1)$ zmienna:

$$V = X_1^2 + \cdots + X_m^2 \sim \chi^2(m)$$

ma rozkład chi-kwadrat z *m* stopniami swobody, $V \sim \chi^2(m)$, gdzie $E(V) = m \; i \; Var(V) = 2m$

Rozklad t-Studenta

Dla niezależnych zmiennych $X \sim N(0,1) \; oraz \; V \sim \chi^2(m)$ zmienna:

$$t = \frac{X}{\sqrt{V/m}}$$

ma rozkład t-studenta z *m* stopniami swobody, $t \sim t(m)$, gdzie $E(t) = 0 \; i \; Var(t) = \frac{v}{v-2}$

Rozkład F-Snedecora

Dla niezależnych zmiennych $V_1 \sim \chi^2(m) \; oraz \; V_2 \sim \chi^2(k)$ zmienna

$$F = \frac{V_1/m}{V_2/k}$$

ma rozkład F z (m, k) stopniami swobody, $F \sim F(m, k)$

WAŻNE:

- jeżeli $m \to \infty$ to rozkład $t(m)$ zbiega do rozkładu $N(0,1)$, zaś $F(k, m)$ zbiega do $\chi^2(k)$
- jeżeli $t \sim t(m)$ to $t^2 \sim F(1, m)$

Centralne twierdzenie graniczne

Jeżeli X_i są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, takiej samej wartości oczekiwanej $\mu = E(x_i)$ oraz (skończonej) wariancji $\sigma^2 = Var(X_i)$ to zmienna losowa

$$Z_n = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$$

gdzie $\bar{x}_n = \frac{1}{n}\sum x_i$, zbiega wraz z liczebnością próby n do rozkładu $N(0,1)$

Dzięki centralnemu twierdzeniu granicznemu, przy dużej próbie możliwe jest uproszczenie obliczeń dzięki przejściu na rozkład normalny.

Pakiety ekonometryczne

Jakie narzędzia mogę wykorzystywać?

GRETL

Darmowy program, w którym można oszacować większość popularnych typów modeli ekonometrycznych. Do pobrania na stronie:
http://www.kufel.torun.pl/

Excel

Niektóre podstawowe modele ekonometryczne można też oszacować w Excelu.
Należy aktywować dodatek Analysis ToolPak
Plik → Opcje → Dodatki → Zarządzaj: Dodatki programu Excel → Przejdź i zaznaczyć Analysis ToolPak
Następnie należy wybrać Dane → Analiza danych.

Jakie narzędzia mogę wykorzystywać?

Specjalistyczne języki programowania

R

- Bardzo popularny język wśród ekonometryków
- Umożliwia oszacowanie najróżniejszych modeli ze względu na bardzo szeroki zakres bibliotek
- Posiada bogatą ofertę dotyczącą graficznej prezentacji wyników
- Uzupełniony o narzędzie Rstudio, które jest wygodnym kompilatorem kodu
- Wadą jest prędkość obliczeń i duże obciążenie pamięci RAM
- Przykładowe kursy:
  https://www.datacamp.com/courses/free-introduction-to-r
  https://www.coursera.org/learn/r-programming
- Programiści piszący w R uśmiechają się najczęściej[1]

Python

- Język bardziej popularnych u osób zajmujących się inżynierią danych (data science)
- Wiele bibliotek z zakresu uczenia maszynowego

Julia

- Relatywnie nowy język, który intensywnie się rozwija
- Jego zaletą jest szybkość, link do analizy porównującej R, Python i Julia:
  https://voxeu.org/content/which-numerical-computing-language-best-julia-matlab-python-or-r

Inne, np. Matlab

[1] https://medium.com/swlh/what-programming-language-has-the-happiest-developers-f0636b08e898

Zadania

Zadanie 1.1

a. Wymyśl co najmniej dwa przykłady zmiennych każdego typu:
   - zmienna ciągła
   - zmienna dyskretna
   - zmienna binarna
   - zmienna kategoryczna / uporządkowana.

b. Wybierz jedną ze zmiennych z punktu a. i zaproponuj model ekonometryczny, w którym będzie ona zmienną zależną

c. Wybierz inną ze zmiennych z punktu a. i zaproponuj model ekonometryczny, w którym będzie ona zmienną objaśniającą

d. Czy pozyskanie danych dla zmiennych z punktu a. jest możliwe?

Zadanie 1.2

Realizacja zmiennej losowej X jest następująca:
$x_1 = 1, x_2 = 3, x_3 = 5, x_4 = 3.$
Oblicz i zinterpretuj:

a. Średnią arytmetyczną $\bar{x} = \sum_{i=1}^{4} \frac{x_i}{4}$

b. Wyrażenie $\sum_{i=1}^{4}(x_i - \bar{x})$

c. Wyrażenie $\sum_{i=1}^{4}(x_i - \bar{x})^2$

d. Wyrażenie $\left(\sum_{i=1}^{4} x_i^2\right) - 4\bar{x}^2$

Zadanie 1.3

Dla każdego z poniższych punktów spróbuj określić specyfikacje modelu oraz zmienne, dla których należy znaleźć obserwacje. Jakie jest potencjalne źródło pozyskania tych obserwacji?

a. Naukowcy pragną ustalić jaka jest temperatura ciała zdrowego człowieka

b. Sieć wodociągowa planuje ustalić od czego zależy miesięczne zużycie wody przez gospodarstwa domowe

c. Ministerstwo zdrowia jest zainteresowane od czego zależy czas trwania zarażenia wirusem COVID-19 u pacjentów

d. Sprzedawca chce dowiedzieć się jaka jest żywotność żarówek, które ma w ofercie

Zadanie 1.4

Liczba oddechów na minutę wśród studentów w trakcie egzaminu ma rozkład normalny z wartością oczekiwaną równą 12 i odchyleniem standardowym równym 2,3. Jaka jest proporcja studentów, którzy oddychają z wartościami z poniższych przedziałów?

a. 9,7 do 14,3 wdechów na minutę

b. 7,4 do 16,6 wdechów na minutę

c. 9,7 do 16,6 wdechów na minutę

d. mniej niż 5,1 lub więcej niż 18,9 wdechów na minutę.

Zadanie 1.5

Wśród 60 studentów wiemy, że:

- 9 nie mieszka w akademiku
- 36 to studenci studiów licencjackich
- 3 studentów studiów licencjackich nie mieszka w akademiku

Niech X oraz Y określają binarne zmienne losowe, określające uczestnictwo w studiach licencjackich oraz mieszkanie w akademiku.

a. Oszacuj i zinterpretuj $P(X = 1, Y = 0)$ oraz $P(X = 1|Y = 0)$
b. Czy $P(Y = 1|X = 1)$ jest takie samo jak $P(Y = 1)$?
c. Znajdź udział studentów studiów magisterskich, którzy mieszkają w akademiku
d. Czy zmienne X i Y są niezależne?

Zadanie 1.6

Korzystając ze strony Eurostatu(https://ec.europa.eu/eurostat/data/database)
wykonaj następujące polecenia:

- znajdź dane dla inflacji HICP (r/r) dla wybranego kraju
- wgraj dane do excela
- zainmportuj dane do Gretla
- stwórz wykres danych w Gretlu

W DOMU (trudniejsze) spróbuj pobrać te same dane za pomocą dodatku w Gretlu DB.NOMICS

Zadanie 1.7

Otwórz Gretla. Zaimportuj dane z pliku cps5.gdt i wykonaj następujące polecenia:

a. Sprawdź funkcjonalności pod
Narzędzia—Tablice statystyczne, Narzędzia—wartość p, Narzędzia—Testy parametryczne.
b. Wygeneruj i opisz statystyki opisowe dla zmiennej WAGE.
c. Stwórz wykres WAGE do EDUC. Opisz tę zależność
d. Znajdź macierz współczynników korelacji między zmiennymi WAGE, EDUC i EXPER.
e. Utwórz nowe zmienne: $EDUC^2, \ln(EDUC), \sqrt{EDUC}, {}^{EDUC}/_{EXPER}$.
f. Utwórz nową zmienną, która przyjmuje wartość 1 dla pierwszych 300 obserwacji, 0 dla pozostałych. Zmień ostatnią obserwację w próbie, tak aby wartość tej nowej zmiennej była równa 1.
g. Ogranicz zakres próby do obserwacji 1-500.
h. Zapisz teoretyczny model wyjaśniający zmienność płac. Wybierz zmienne objaśniające na podstawie teorii ekonomicznej poznanej do tej pory oraz własnej logiki.
i. Jakie masz podejrzenia odnośnie znaków zmiennych objaśniającyh w Twoim modelu? Zastanów się czy wpływ tych zmiennych na zarobki jest dodatni czy ujemny i dlaczego?

Zadanie 1.8.

Wykonaj następujące ćwiczenia dotyczące operacji na macierzach.

a. Rozwiń iloczyn macierzy

$$X = ((AB + (CD)')((EF)^{-1} + GH))'.$$

Przyjmij, że wszystkie macierze są kwadratowe oraz, że E i F są odwracalne.

b. Niech X będzie niepustą macierzą o wymiarze $n \times k$, gdzie $n \geq k$.
Pokaż, że macierz $X'X$ jest symetryczna.

W domu

W ramach **pracy domowej**

- powtórz materiał z sekcji **Podstawy Statystyczne** i **Algebra Liniowa**. Możesz wrócić do swoich notatek z poprzednich przedmiotów matematycznych i statystycznych.
- zainstaluj **Gretla** na prywatnym komputerze i poćwicz podstawowe funkcjonalności pakietu

Temat 2

Metoda najmniejszych kwadratów

Katarzyna Bech-Wysocka i Piotr Dybka

- Model regresji liniowej
- Estymacja parametrów
- Metoda Najmniejszych Kwadratów (MNK)
- Założenia klasycznego modelu regresji liniowej
- Własności estymatora MNK
- Twierdzenie Gaussa-Markova
- Precyzja oszacowań: wariancja estymatora MNK
- Dopasowanie modelu do danych: współczynnik determinacji R^2

Model regresji liniowej

Rozważmy model regresji z jedną zmienną objaśniającą:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

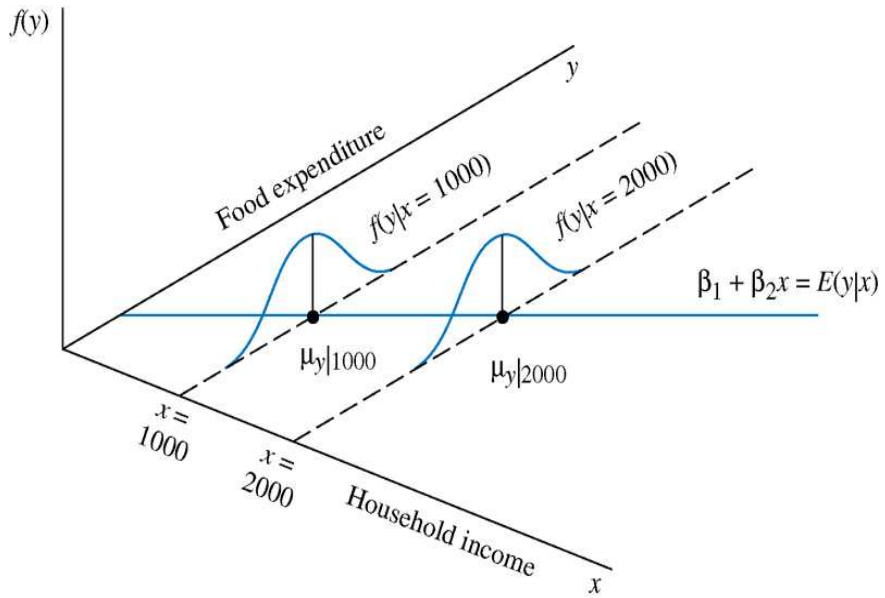
- Ekonomiści wskazują na występowanie (zazwyczaj deterministycznych) związków między zmiennymi, np. że wydatki na jedzenie (y) zależą od dochodu (x)

- Dla ekonometryka:
  - y jest **zmienną losową**, której wartość zależy od x (część deterministyczna), ale także od składnika losowego ε (część stochastyczna)
  - model ekonometryczny może być wykorzystany, aby ustalić warunkowy rozkład y, czyli:

warunkową wartość oczekiwaną $E(y|x) = \mu_{y|x}$
warunkową wariancję $\mathrm{Var}(y|x) = \sigma^2_{y|x}$

- Parametry β_0 oraz β_1 nie są znane, ale można oszacować ich wartości na podstawie realizacji dla y_i oraz x_i, gdzie $i = 1,2,\ldots,N$

Model regresji liniowej



Źródło: Principles of Econometrics, R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

Model regresji liniowej

- Dla modelu regresji y względem x (np. wydatków na jedzenie względem dochodu)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

warunkowa wartość oczekiwana, czyli część deterministyczna, wynosi:

$$E(y|x) = \mu_{y|x} = \beta_0 + \beta_1 x$$

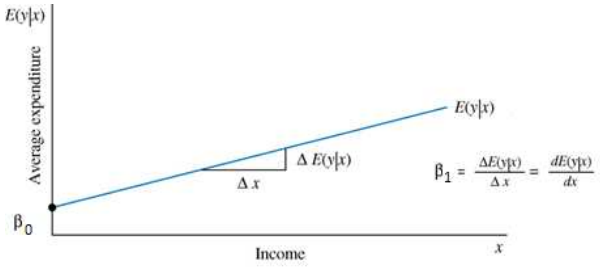
gdzie β_0 to wyraz wolny, a β_1 współczynnik kierunkowy (funkcji liniowej).

- Współczynnik kierunkowy opisuje:

$$\beta_1 = \frac{\Delta E(y|x)}{\Delta x} = \frac{dE(y|x)}{dx}$$

pochodną warunkowej wartości oczekiwanej y względem x.

Jak interpretujemy ten współczynnik?



Źródło: Principles of Econometrics,
R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. **Estymacja**
5. Weryfikacja
6. Zastosowanie

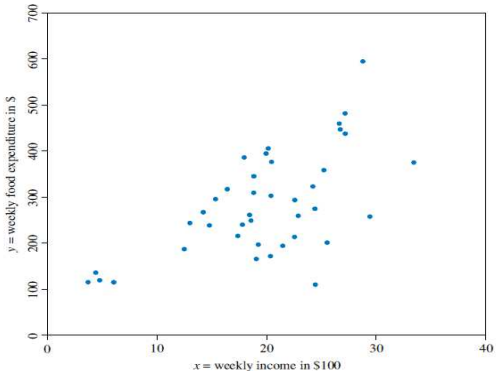
Metoda Najmniejszych Kwadratów

Estymacja parametrów regresji

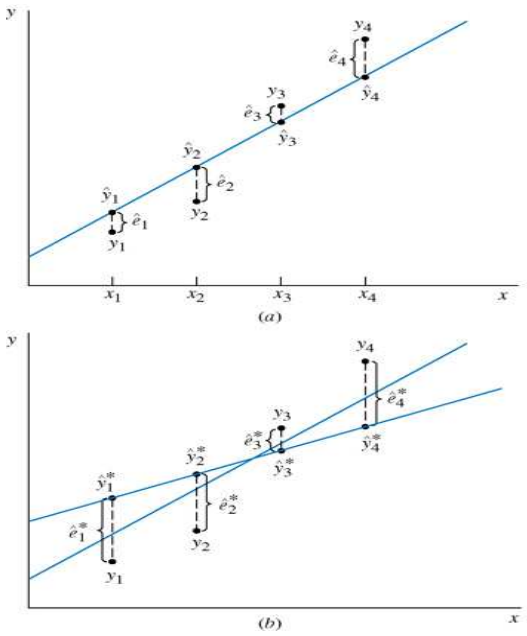
- Dla modelu $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ parametry β_0 oraz β_1 nie są znane, ale mogą zostać oszacowane na podstawie obserwacji dla zmiennych y_i i x_i, gdzie $i = 1,2, \dots, N$
- Idea estymacji parametrów: znalezienie kombinacji liniowej $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$, która najlepiej przybliża wartości y_i dla każdego $i = 1,2, \dots, N$

Jak wybrać kryterium "najlepszego przybliżenia"?

Observation (household)	Food expenditure ($)	Weekly income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
⋮	⋮	
39	257.95	29.40
40	375.73	33.40
	Summary statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. Dev.	112.6752	6.8478



Estymacja parametrów regresji



Źródło: Principles of Econometrics, R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

- Dla dowolnych wartości $\widehat{\beta_0}$ i $\widehat{\beta_1}$ możemy policzyć kombinację $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$, która określa linię regresji
- Wartości $\widehat{\beta_0}$ i $\widehat{\beta_1}$ dobierz tak, żeby odległości między linią regresji i realizacjami y_i były małe
- W celu określenia co oznacza „małe odległości" możemy wykorzystać różne miary

Estymator MNK

Jak znaleźć optymalne Wartości $\widehat{\beta_0}$ i $\widehat{\beta_1}$? Metoda Najmniejszych Kwadratów (MNK)

- Wartości teoretyczne / dopasowane, $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$, możemy porównać z realizacją, y_i. W ten sposób otrzymujemy reszty modelu, czyli realizację składnika losowego:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$$

- W **metodzie najmniejszych kwadratów** kryterium optymalizacji przy obliczaniu wartości $\widehat{\beta_0}$ oraz $\widehat{\beta_1}$ jest **minimalizacja sumy kwadratów reszt**:

$$SSE = \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \sum_{i=1}^{N} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 = SSE(\widehat{\beta_0}, \widehat{\beta_1})$$

- Wartości $\widehat{\beta_0}$ oraz $\widehat{\beta_1}$ wyznaczane są poprzez rozwiązanie układu równań opisanych przez warunki pierwszego rzędu:

$$\frac{\partial SSE}{\partial \widehat{\beta_0}} = \frac{\partial \sum_{i=1}^{N} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2}{\partial \widehat{\beta_0}} = 0$$

$$\frac{\partial SSE}{\partial \widehat{\beta_1}} = \frac{\partial \sum_{i=1}^{N} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2}{\partial \widehat{\beta_1}} = 0$$

Estymator MNK

Rozwiązaniem układu równań opisanych przez warunki pierwszego rzędu są wartości:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

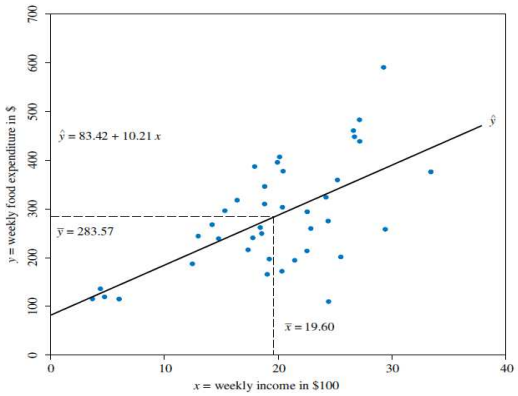
Są to wzory na estymator MNK
w prostym modelu regresji liniowej

Uwaga:
- estymator jest zmienną losową
- po podstawieniu realizacji x_i oraz y_i do wzorów otrzymujemy oszacowania, czyli liczby
- estymator ≠ oszacowanie!

Przykład 2.1. Estymacja parametrów regresji

Rozważmy model ekonometryczny, w którym wydatki na żywność ($food\_exp$, USD) zależą od dochodu ($income$, 100USD). Na podstawie danych z pliku `food.gdt` uzyskano następującą zależność:

$$\widehat{food\_exp_i} = 83.42 + 10.21 income_i.$$



Pytanie: Jak zinterpretować oszacowane wartości parametrów tego modelu?

Przykład 2.2. Estymacja parametrów regresji

Na podstawie danych z pliku `bweight.gdt` oszacowano wpływ wieku matki (*mage*, w latach) na wagę urodzeniową noworodka (*bweight*, w gramach).

Model 1: OLS, using observations 1-4642
Dependent variable: bweight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3074.06	40.7441	75.45	<0.0001	***
mage	10.8519	1.50383	7.216	<0.0001	***

Mean dependent var	3361.680	S.D. dependent var	578.8196
Sum squared resid	1.54e+09	S.E. of regression	575.6608
R-squared	0.011098	Adjusted R-squared	0.010885
F(1, 4640)	52.07250	P-value(F)	6.22e-13
Log-likelihood	−36088.03	Akaike criterion	72180.06
Schwarz criterion	72192.95	Hannan-Quinn	72184.59

Pytania:

- Czy oszacowanie dla wyrazu wolnego ma interpretację?
- O ile zmieni się waga urodzeniowa dziecka, jeżeli wiek matki rok?

Estymacja parametrów: regresja wieloraka

- W modelu regresji wielorakiej występuje K zmiennych objaśniających (x_k dla $k = 1,2,\dots,K$):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

gdzie parameter β_k mierzy jak zmiana wartości x_k wpływa na warunkową wartość oczekiwaną y, przy założeniu, że pozostałe zmienne objaśniające nie zmieniają się (zasada ***ceteris paribus***)

- Wzór na estymator MNK wyprowadza się korzystając z notacji macierzowej (zob. Temat 1)

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

- W tym przypadku suma kwadratów reszt wynosi:

$$SSE = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

zaś warunki pierwszego rzędu można zapisać jako:

$$\frac{\partial SSE}{\partial \widehat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = 0$$

i pozwalają one na uzyskanie wzoru na estymator MNK:

Estymator MNK w regresji wielorakiej

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Przykład 2.3. Estymacja parametrów regresji

Załóżmy, że posiadamy dodatkowe informacje o zmiennych wpływających na wydatki żywnościowe:

- $fPriceIndex$: wskaźnik cen żywności,
- $farmer$: zmienna zero-jedynkowa, przyjmująca wartośc 1 dla rolników.

Zmienne te uwzględniamy w specyfikacji modelu i uzyskujemy następujące oszacowania parametrów:

$$\widehat{food\_exp}_i = 75.14 + 9.12 income_i - 0.12 fPriceIndex_i - 16.83 farmer_i,$$

W zapisie wektorowym oszacowania (nie mylić z estymatorem!) MNK wynoszą:

$$\hat{\beta} = \begin{bmatrix} 75{,}14 \\ 9{,}12 \\ -0{,}12 \\ -16{,}83 \end{bmatrix}.$$

Pytanie: Jaka jest interpretacja oszacowań parametrów tego modelu?

Przykład 2.4. Estymacja parametrów regresji

Na podstawie danych z pliku `bweight.gdt` oszacowano wpływ wieku matki ($mage$, w latach), wieku ojca ($fage$, w latach) oraz pochodzenia matki ($mhisp = 1$ jeżeli matka jest Latynoską, $mrace = 1$ jeżeli matka jest biała) na wagę urodzeniową noworodka ($bweight$, w gramach).

Model 2: OLS, using observations 1-4642
Dependent variable: bweight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	2901.60	42.3930	68.45	<0.0001	***
mage	5.93910	1.80970	3.282	0.0010	***
fage	1.79069	1.08443	1.651	0.0987	*
mhisp	−34.4813	46.1281	−0.7475	0.4548	
mrace	303.374	23.0942	13.14	<0.0001	***

Mean dependent var	3361.680	S.D. dependent var	578.8196
Sum squared resid	1.48e+09	S.E. of regression	564.9135
R-squared	0.048294	Adjusted R-squared	0.047473
F(4, 4637)	58.82549	P-value(F)	1.63e-48
Log-likelihood	−35999.05	Akaike criterion	72008.09
Schwarz criterion	72040.31	Hannan-Quinn	72019.42

Pytania:

- Czy oszacowany wyraz wolny ma ekonomiczną interpretację?
- O ile zmieni się waga urodzeniowa dziecka, jeżeli wiek matki wzrośnie o 1 jednostkę (1 rok)?
- Jak zinterpretować pozostałe oszacowania?

Założenia i własności estymatora MNK

Założenia klasycznego modelu regresji liniowej

Założenie A1

Prawdziwy model jest następujący:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

A1 oznacza, że:

- poprawnie dobrano postać funkcyjną modelu
- Odpowiednio dobrano zbiór regresorów, tj.:
  - nie pominięto żadnej istotnej zmiennej objaśniającej
  - nie włączono do zbioru regresorów niepotrzebnych zmiennych (szerzej w Temat 4).
- Założenie A1 jest szczegółowo omawiane w części Temat 4

Założenie A2

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ oraz } E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$$

Niespełnienie A2 oznacza, że zmienne objaśniające są endogeniczne, co ma poważne skutki dla własności estymatora MNK. Problem ten jest omawiany w części Temat 12 – Temat 14.

Założenia klasycznego modelu regresji liniowej

Założenie A3

$$Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

Spełnienie A3 oznacza, że nie występuje problem:

- heteroskedastyczności wariancji składnika losowego (Temat 6) lub
- autokorelacji składnika losowego (Temat 7)

Założenie A4

$\mathbf{X}$ jest nielosową macierzą o wymiarach $n \times (K+1)$, której rząd wynosi $rank(X) = (K+1) < N$.

Spełnienie A4 oznacza, że nie występuje problem współliniowości regresorów (Temat 5)

Założenie A5 (opcjonalne)

$$\varepsilon \sim N(0, \sigma^2 I)$$

Założenie A5 nie jest konieczne do zapewnienia odpowiednich własności estymatorów MNK, ale jest potrzebne do przeprowadzania testów (w małych próbach) (Temat 5)

Własności estymatora MNK

Jeżeli spełnione są założenia A1-A4 możemy ustalić, jakie są własności estymatora MNK.
Będziemy szukać odpowiedzi na następujące pytania:

1. Skoro estymator MNK jest zmienną losową, to jaka jest jego wartość oczekiwana, wariancja i ogólnie rozkład prawdopodobieństwa?
2. Jak własności estymatora MNK wyglądają na tle własności innych estymatorów?

- Zaczynamy od wartości oczekiwanej.

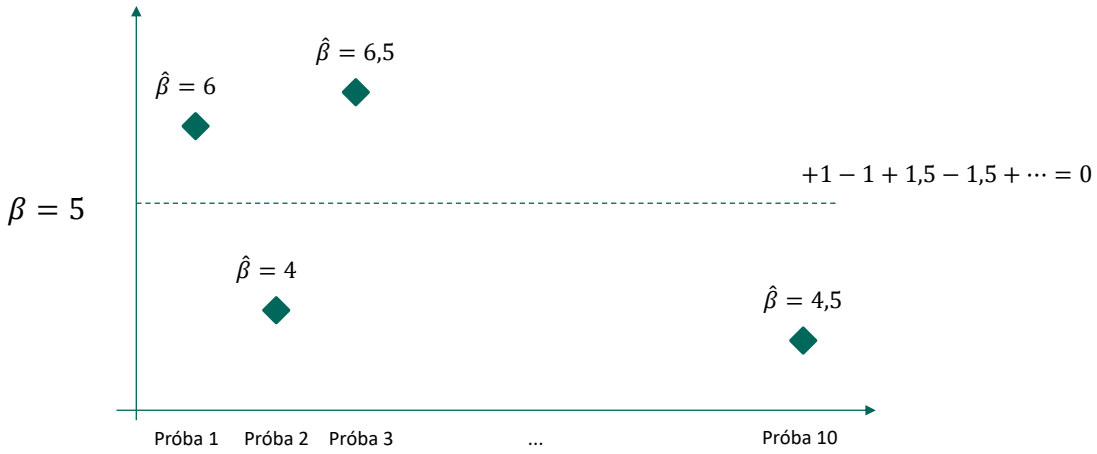
$$E(\widehat{\boldsymbol{\beta}}) = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] \overset{A1}{=} E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] \overset{A2}{=} \boldsymbol{\beta}$$

Jeżeli spełnione są założenia A1-A2 to
wartość oczekiwana $\widehat{\boldsymbol{\beta}}$ jest równa prawdziwej wartości parametru $\boldsymbol{\beta}$
A zatem estymator MNK jest nieobciążony

- **WAŻNE:** nieobciążoność nie oznacza, że wartość oszacowania na podstawie jednej próby losowej jest taka sama jak prawdziwa wartość parametru! (estymator ≠ oszacowanie).

Własności estymatora MNK

- Nieobciążoność estymatora oznacza, że jeżeli powtórzymy estymację z wykorzystaniem różnych obserwacji, to „średnio" oszacowania będą kształtowały blisko prawdziwej wartości parametru.
- W celu ilustracji, załóżmy, że wylosowaliśmy z populacji 10 prób losowych obserwacji.
  Dla każdej próby estymujemy wartość parametru β. Wiemy, że prawdziwa wartość to $\beta = 5$.
  Nieobciążoność możemy przedstawić następująco:



Własności estymatora MNK: wariancja

- A jaka jest wariancja estymatora MNK?

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] \overset{A3}{=} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

  Zauważ, że wykorzystaliśmy wzór z poprzednich slajdów: $\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$

- Znamy zatem pełny rozkład dla estymatora MNK (gdy spełnione są **A1**-**A5**):

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\beta}})$$

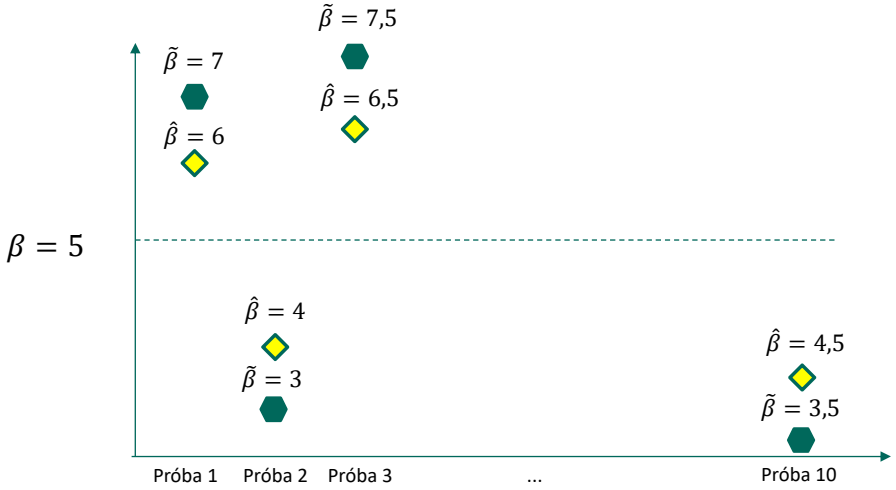
- Wariancje poszczególnych $\widehat{\beta_k}$ to elementy na głównej przekątnej macierzy $\boldsymbol{\Sigma}_{\hat{\beta}}$.

Reguły:

1. Im większa wariancja składnika losowego σ^2, tym większy wpływ części stochastycznej modelu ekonometrycznego, co jest odzwierciedlone w większej wariancji estymatora MNK
2. Im większa próba N, tym mniejsza wariancja estymatora MNK [wpływ przez $(\mathbf{X}'\mathbf{X})^{-1}$].

Własności estymatorów MNK - wariancja

- Załóżmy, że mamy do wyboru dwa nieobciążone estymatory $\hat{\beta}$ oraz $\tilde{\beta}$
- Szukamy wtedy tego, który ma mniejszą wariancję, czyli jest **efektywniejszy**



Twierdzenie Gaussa - Markova

- Jeżeli spełnione są założenia A1-A4, to estymator MNK ma najmniejszą wariancję wśród wszystkich liniowych, nieobciążonych estymatorów (jest **najefektywniejszy**)
- Mówimy wtedy, że jest **Best Linear Unbiased Estimators (BLUE).**

Zauważ, że:

1. Estymator MNK jest „najlepszy" w porównaniu do innych liniowych, nieobciążonych estymatorów. Twierdzenie nic nie mówi o wszystkich możliwych estymatorach.
2. Estymator MNK jest „najlepszy", bo ma najmniejszą wariancję.
3. Twierdzenie jest prawdziwe tylko wtedy, gdy spełnione są założenia A1-A4. Jeżeli którekolwiek z nich jest niespełnione, to estymatory MNK nie są BLUE.

CIEKAWOSTKA. Jeżeli spełnione jest także założenie A5 to estymator MNK ma taki sam wzór jak estymator Metody Największej Wiarygodności (MNW, *ang. Maximum Likelihood*). Możemy wtedy wykorzystać metodę dolnej granicy Cramera-Rao, aby udowodnić, że estymator MNW jest **BUE- Best Unbiased Estimators**, czyli najefektywniejszy wśród wszystkich nieobciążonych estymatorów (nie tylko liniowych).

Precyzja estymatora MNK

Wariancja składnika losowego

- We wzorze na wariancję estymatora MNK:

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

pojawia się wariancja składnika losowego σ^2. Niestety, tej wartości zazwyczaj nie znamy, a zatem jest to dodatkowy parametr, który musimy oszacować.

- Intuicyjnym estymatorem dla σ^2 jest średnia arytmetyczna kwadratów reszt:

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i^2$$

Niestety, ten estymator jest obciążony. Powodem jest to, że reszty pochodzą z modelu, w którym liczba stopni swobody wynosi $N - (K + 1)$, co jest równe liczbie obserwacji pomniejszonej o liczbę estymowanych parametrów. Intuicja jest taka, że za każdy oszacowany parametr tracimy stopień swobody. Przykładowo, jakie są reszty w modelu, w którym $N = 2$ i $K = 1$?

- Nieobciążony estymator dany jest wzorem:

$$s^2 = \frac{1}{N - (K + 1)}\sum_{i=1}^{N} \hat{\varepsilon}_i^2$$

- Pierwiastek kwadratowy z s^2 nazywamy **błędem standardowym regresji.**

Wariancja estymatora MNK

- Jeżeli we wzorze na wariancję estymatora MNK, $\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, nieznaną wartość σ^2 zastąpimy przez oszacowanie s^2, to uzyskamy wzór:

$$\widehat{\boldsymbol{\Sigma}_{\hat{\beta}}} = \widehat{Var(\hat{\boldsymbol{\beta}})} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Zauważmy, że za parametr σ^2 podstawiliśmy zmienną losową o rozkładzie:

$$\frac{(N-(K+1))s^2}{\sigma^2} \sim \chi^2_{N-(K+1)}$$

- Podstawienie to sprawia, że jeżeli korzystamy z $\widehat{\boldsymbol{\Sigma}_{\hat{\beta}}}$, to rozkład estymatora MNK zamienia się na wielowymiarowy rozkład t-Studenta o $v = N-(K+1)$ stopniach swobody (por. Temat 1). Pojedyncze parametry mają natomiast jednowymiarowy rozkład t-Studenta:

$$\frac{\widehat{\beta_k} - \beta_k}{S_{\hat{\beta}_k}} \sim t_{N-(K+1)}$$

gdzie $S_{\hat{\beta}_k}$ jest średnim błędem szacunku (szczegóły na kolejnym slajdzie)

Średni błąd szacunku

$$\widehat{\boldsymbol{\Sigma}_{\hat{\beta}}} = \widehat{Var(\hat{\boldsymbol{\beta}})} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Na głównej przekątnej macierzy $\widehat{\boldsymbol{\Sigma}_{\hat{\beta}}}$ znajdują się wariancje estymatora MNK dla indywidualnych parametrów (poza przekątną są kowariancje):

$$d_{kk} = \widehat{var(\hat{\beta}_k)}$$

- Pierwiastki kwadratowe wariancji opisują **błędy standardowe** estymatora MNK, które określamy jako **błędy szacunku**:

$$S_{\hat{\beta}_k} = s(\hat{\beta}_k) = \sqrt{d_{kk}}$$

- Błędy szacunku określają precyzję oszacowań MNK. Można wykorzystać również **relatywny błąd standardowy** (średni względny błąd szacunku):

$$V_{\hat{\beta}_k} = \frac{S_{\hat{\beta}_k}}{|\hat{\beta}_k|} \times 100\%$$

- Jeżeli $V_{\hat{\beta}_k} < 50\%$ to mówimy, że zmienna x_k istotnie wpływa na y (szerzej w Temat 3).

Przykład 2.4 cd. Błędy szacunku parametrów

Na podstawie danych z pliku `bweight.gdt` oszacowano wpływ wieku matki ($mage$, w latach), wieku ojca ($fage$, w latach) oraz pochodzenia matki ($mhisp = 1$ jeżeli matka jest Latynoską, $mrace = 1$ jeżeli matka jest biała) na wagę urodzeniową noworodka ($bweight$, w gramach).

Błędy szacunku

Model 2: OLS, using observations 1-4642
Dependent variable: bweight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	2901.60	42.3930	68.45	<0.0001	***
mage	5.93910	1.80970	3.282	0.0010	***
fage	1.79069	1.08443	1.651	0.0987	*
mhisp	−34.4813	46.1281	−0.7475	0.4548	
mrace	303.374	23.0942	13.14	<0.0001	***

Mean dependent var	3361.680	S.D. dependent var	578.8196
Sum squared resid	1.48e+09	S.E. of regression	564.9135
R-squared	0.048294	Adjusted R-squared	0.047473
F(4, 4637)	58.82549	P-value(F)	1.63e-48
Log-likelihood	−35999.05	Akaike criterion	72008.09
Schwarz criterion	72040.31	Hannan-Quinn	72019.42

Przykładowy względny błąd standardowy (ocena precyzji oszacowania):

$$V_{\hat{\beta}_1} = \frac{S_{\hat{\beta}_1}}{|\hat{\beta}_1|} \times 100\% = \frac{1{,}8097}{|5{,}9391|} \times 100\% = 30{,}5\% < 50\%.$$

Oszacowanie przedziałowe

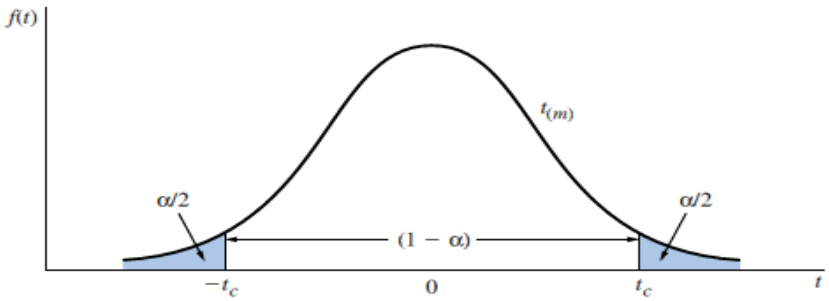
Błędy szacunku mogą być również wykorzystane do konstrukcji przedziałów ufności dla parametru, zwanych również **oszacowaniem przedziałowym.** Jest to przedział w którym, z określonym prawdopodobieństwem, znajduje się prawdziwa wartość parametru. Zauważmy, że:

$$\frac{\hat{\beta}_k - \beta_k}{S_{\hat{\beta}_k}} \sim t_{N-(K+1)}$$

co oznacza:

$$P\left(\hat{\beta}_k - t_c S_{\hat{\beta}_k} \leq \beta_k \leq \hat{\beta}_k + t_c S_{\hat{\beta}_k}\right) = 1 - \alpha$$

Przedział $\hat{\beta}_k \pm t_c S_{\hat{\beta}_k}$ nazywamy $(1-\alpha)$ przedziałem ufności dla parametru β_k.



Źródło: Principles of Econometrics, R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

Przykład 2.4 cd. Oszacowanie przedziałowe

Na podstawie danych z pliku `bweight.gdt` oszacowano wpływ wieku matki ($mage$, w latach), wieku ojca ($fage$, w latach) oraz pochodzenia matki ($mhisp = 1$ jeżeli matka jest Latynoską, $mrace = 1$ jeżeli matka jest biała) na wagę urodzeniową noworodka ($bweight$, w gramach).

Model 1: OLS, using observations 1-4642
Dependent variable: bweight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3074.06	40.7441	75.45	<0.0001	***
mage	10.8519	1.50383	7.216	<0.0001	***

Mean dependent var	3361.680	S.D. dependent var	578.8196
Sum squared resid	1.54e+09	S.E. of regression	575.6608
R-squared	0.011098	Adjusted R-squared	0.010885
F(1, 4640)	52.07250	P-value(F)	6.22e-13
Log-likelihood	−36088.03	Akaike criterion	72180.06
Schwarz criterion	72192.95	Hannan-Quinn	72184.59

95% przedział ufności dla β_{mage} to $(7.90, 13.8)$

Pytanie: Jak obliczono ten przedział?

Dopasowanie modelu do danych

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Etapy weryfikacji modelu

1. Oceny parametrów i ich znaki
2. Istotność parametrów
3. **Dopasowanie modelu do danych**
4. Specyfikacja modelu / postać funkcyjna
5. Własności składnika losowego
6. Stabilność parametrów

32

Dopasowanie modelu: współczynnik R-kwadrat

- Jak ocenić, czy model ekonometryczny dobrze opisuje obserwacje dla y? Zauważmy, że:

$$y_i = \widehat{y_i} + \widehat{\varepsilon_i}$$

gdzie $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_K} x_{Ki}$ to część wyjaśniona przez model

- Przedstawmy powyższe równanie jako odchylenie od średniej:

$$y_i - \bar{y} = (\widehat{y_i} - \bar{y}) + \widehat{\varepsilon_i}.$$

- Biorąc pod uwagę, że $\sum \widehat{y_i}\widehat{\varepsilon_i}$, można pokazać, że:

$$\sum (y_i - \bar{y})^2 = \sum (\widehat{y_i} - \bar{y})^2 + \sum \widehat{\varepsilon_i}^2$$

- Powyższe równanie pozwala na dekompozycję całkowitej zmienności y_i (TSS) na część objaśnioną przez model (SSR) oraz pozostałe czynniki (SSE)
- **Współczynnik determinacji R^2** określa proporcję zmienności y wyjaśnioną przez model:

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}.$$

Dopasowanie modelu: skorygowany R-kwadrat

- Własnością miary R^2 jest to, że jest wartość rośnie (a przynajmniej nie maleje), jeżeli dodamy do modelu kolejne regresory. Dlatego faworyzuje ona "duże modele"
- W celu porównywania dopasowania alternatywnych modeli warto skorygować wartość R^2 o liczbę szacowanych parametrów. Wartość skorygowanego $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{TSS/(N-1)}$$

Pamiętaj: gdy porównujemy alternatywne specyfikacje modelu, wybierz tę z wyższą wartością skorygowanego R-kwadrat.

- Można też porównywać modele wykorzystując kryteria informacyjne: Akaike Information Criterion (AIC), Baysian- Schwartz Information Criterion (BIC) or Hannan-Quinn Information Criterion (HIC). Ich wartości są sumą miary dopasowania do danych oraz kary za liczbę parametrów.

Pamiętaj: gdy porównujemy alternatywne specyfikacje modelu, wybierz tę z niższą wartością kryteriów informacyjnych.

Przykład 2.4 cd. Porównywanie modeli

Na podstawie danych w `bweight.gdt` otrzymano 2 konkurujące modele:

Model 1: OLS, using observations 1-4642
Dependent variable: bweight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3074.06	40.7441	75.45	<0.0001	***
mage	10.8519	1.50383	7.216	<0.0001	***

Mean dependent var	3361.680	S.D. dependent var	578.8196
Sum squared resid	1.54e+09	S.E. of regression	575.6608
R-squared	0.011098	Adjusted R-squared	0.010885
F(1, 4640)	52.07250	P-value(F)	6.22e-13
Log-likelihood	−36088.03	Akaike criterion	72180.06
Schwarz criterion	72192.95	Hannan-Quinn	72184.59

Model 2: OLS, using observations 1-4642
Dependent variable: bweight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	2901.60	42.3930	68.45	<0.0001	***
mage	5.93910	1.80970	3.282	0.0010	***
fage	1.79069	1.08443	1.651	0.0987	*
mhisp	−34.4813	46.1281	−0.7475	0.4548	
mrace	303.374	23.0942	13.14	<0.0001	***

Mean dependent var	3361.680	S.D. dependent var	578.8196
Sum squared resid	1.48e+09	S.E. of regression	564.9135
R-squared	0.048294	Adjusted R-squared	0.047473
F(4, 4637)	58.82549	P-value(F)	1.63e-48
Log-likelihood	−35999.05	Akaike criterion	72008.09
Schwarz criterion	72040.31	Hannan-Quinn	72019.42

Pytania:

- Który model jest lepiej dopasowany do danych?
- Wykorzystaj skorygowany R-kwadrat oraz kryteria informacyjne.

Zadania

Zadanie 2.1

Niełatwo jest zrozumieć, że **estymator MNK to zmienna losowa**, zaś jej realizacja zależy od zbioru danych, z którym pracujemy. Aby to zilustrować, za pomocą pakietu ekonometrycznego:

a. Wygeneruj syntetyczne obserwacje z następującego procesu:

$$x_i \sim N(5,2)$$
$$\varepsilon_i \sim N(0,1)$$
$$y_i = 5 + 0.5x_i + \varepsilon_i$$

dla $i = 1,2, \dots, 50$. Przyjmij, że liczebność próby wynosi $N = 50$.

b. Oszacuj parametry modelu:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

i zapisz otrzymane wartości oszacowań MNK. Dlaczego $\widehat{\beta_1} \neq 0.5$?

c. Czy jesteś w stanie ocenić z jakiego rozkładu jest losowana wartość $\widehat{\beta_1}$?

Zadanie 2.2

Jak **zmiana jednostek miary** zmiennych wpływa na oszacowania parametrów?

Załóżmy, że szacujemy parametru prostego modelu liniowego:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Co stanie się z wartościami oszacowań MNK parametrów β_0 i β_1 oraz z oszacowaniami wariancji składnika losowego, jeżeli:

a. Wartości x_i pomnożono przez 10, zaś wartości y_i nie zmieniły się.

b. Wartości y_i pomnożono przez 10, zaś wartości x_i nie zmieniły się.

Zadanie 2.3

Lorraine Cake jest dyrektorem firmy produkującej ciasteczka. Poprosiła swojego asystenta o zebranie danych dotyczących produktywności pracowników firmy. Zebrano informacje o:

- produktywności (procentowe odchylenie od średniej),
- poziomie wykształcenia (zmienna kategoryczna z 7 wartościami, gdzie 1 to najniższy poziom),
- inteligencji (IQ, punktowe odchylenie od średniej),
- płci (zmienna zero-jedynkowa, 1 dla kobiet),
- stanie cywilnym (zmienna zero-jedynkowa, 1 dla zamężnych/żonatych).

Lorraine chce wykorzystać dane, aby sprawdzić, czy single są równie produktywni co pracownicy w związkach małżeńskich. W tym celu szacuje parametry modelu:

$$productivity_i = \beta_0 + \beta_1 education_i + \beta_2 IQ_i + \beta_3 married_i + \varepsilon_i.$$

Zadanie 2.3 cd

$$productivity_i = \beta_0 + \beta_1 education_i + \beta_2 IQ_i + \beta_3 married_i + \varepsilon_i$$

Wyniki (na podstawie 2649 obserwacji) to:

	coefficient	standard error
$\widehat{\beta_0}$	-0.3281	0.0255
$\widehat{\beta_1}$	0.1080	0.0082
$\widehat{\beta_2}$	0.0054	0.0011
$\widehat{\beta_3}$	0.0622	0.0177

a. Zinterpretuj oszacowania parametrów (wzrost $productivity$ o 0.01 oznacz wzrost o 1%)

b. Skoro Lorraine chce głównie mierzyć różnice w produktywności singli i osób w związkach małżeńskich mogłaby oszacować prostszy model:

$$productivity_i = \beta_0 + \beta_1 married_i + \varepsilon_i.$$

Wyjaśnij dlaczego to może być zły pomysł.

c. Lorraine otrzymała $R^2 = 0.1401$ i $\bar{R}^2 = 0.1391$. Jak możemy zinterpretowac te wartości? Dlaczego są inne?

Zadanie 2.3 cd

d. Lorraine dodała do modelu zmienną „płeć”:

$$productivity_i = \beta_0 + \beta_1 education_i + \beta_2 IQ_i + \beta_3 married_i + \beta_4 gender_i + \varepsilon_i.$$

Otrzymała następujące wyniki:

	coefficient	standard error
$\widehat{\beta_0}$	-0.2960	0.0255
$\widehat{\beta_1}$	0.1093	0.0081
$\widehat{\beta_2}$	0.0051	0.0011
$\widehat{\beta_3}$	0.0604	0.0178
$\widehat{\beta_4}$	-0.0690	0.0167

Co możemy powiedzieć na temat produktywności kobiet?

Zadanie 2.3 cd

e. Lorraine oszacowała ponownie oryginalny model

$$productivity_i = \beta_0 + \beta_1 education_i + \beta_2 IQ_i + \beta_3 married_i + \varepsilon_i$$

ale tylko dla kobiet i otrzymała:

	coefficient	standard error
$\widehat{\beta_0}$	-0.2859	0.0291
$\widehat{\beta_1}$	0.0813	0.0093
$\widehat{\beta_2}$	0.0052	0.0012
$\widehat{\beta_3}$	0.0525	0.0195

Oszacowania dla pełnej próby

	coefficient	standard error
$\widehat{\beta_0}$	-0.3281	0.0255
$\widehat{\beta_1}$	0.1080	0.0082
$\widehat{\beta_2}$	0.0054	0.0011
$\widehat{\beta_3}$	0.0622	0.0177

Porównując te wyniki do modelu wyjściowego (tabela po prawej stronie), co możemy powiedzieć o zmiennej $married$? A jakiego oszacowania możemy oczekiwać dla mężczyzn?

Zadanie 2.4

Postanowiono oszacować wpływ przeciętnego dochodu w gospodarstwach domowych (I_t, w 1000USD) i ceny (P_t, w USD) na konsumpcję czekolady na osobę ($Choc_t$ - w 100g). Postać modelu ekonometrycznego jest następująca:

$$Choc_t = \beta_0 + \beta_1 I_t + \beta_2 P_t + \varepsilon_i$$

Otrzymano następujące wyniki:

$$N = 27$$

$$\widehat{Choc_t} = 1.17 + 0.4 I_t - 0.95 P_t$$

$$s^2(X'X)^{-1} = \begin{bmatrix} 0.11 & -0.02 & 0.002 \\ -0.02 & 0.02 & -0.01 \\ 0.002 & -0.01 & 0.01 \end{bmatrix}$$

a. Zinterpretuj wartości oszacowań.
b. Dla każdego parametru oblicz błąd szacunku (także względny) i określ precyzję oszacowania.
c. Podaj oszacowania przedziałowe dla $1 - \alpha = 0.99$ dla β_1 oraz β_2

Zadanie 2.5

Anna jest naukowcem zajmującym się badaniem zdolności językowych dzieci. Stawia hipotezę, że zasób słownictwa wykorzystywanego przez dzieci zależy od sposoby w jaki matka mówi do dziecka. Anna przez 5 lat zbierała informacje na temat dwóch interesujących zmiennych. Po pierwsze, zebrała informację o liczbie różnych słów wypowiadanych przez matkę do dziecka w pierwszym roku jego życia – zmienna W. Po drugie, zebrała dane o wyniku testu słownictwa dzieci, który odbywa się w pierwszym roku szkoły – zmienna S (mierzona w skali 1-100). Dane znajdują się w pliku `Q3_data.xlsx`

a. Zapisz model regresji pozwalający na zbadanie związku, którym Anna jest zainteresowana.
b. Na podstawie danych zebranych przez Annę, oszacuj parametry tego modelu korzystając ze wzoru:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Podpowiedź: Wykorzystaj funkcje tablicowe w Excelu:
MACIERZ.ILOCZYN() – mnożenie macierzy
TRANSPONUJ() – transpozycja macierzy
MACIERZ.ODW() – odwracanie macierzy
Porównaj wyniki otrzymane z automatycznymi funkcjami szacującymi MNK (w Excelu)

c. Zinterpretuj wyniki.

Zadanie 2.5 cd

Lola również zajmuje się badaniem zdolności językowych dzieci. Lola wykorzystuje dane zgromadzone przez Annę, ale zamiast mierzyć wynik testu dzieci w skali 1-100 używa skali 1-60.

d. Na podstawie danych Loli oszacuj model z punktu (a).

e. Opisz zależność między oszacowaniami z punktu (b) i (d).

Maria jest kolejnym badaczem zdolności językowych dzieci, który również korzysta z danych zebranych przez Annę. Maria nie pracuje jednak bezpośrednio ze zmienną W, ale używa odchylenia wartości od średniej – zmienna $W^* = W_i - \bar{W}$.

f. Na podstawie danych Marii oszacuj parametry modelu.

g. Porównaj wyniki z (f) z wynikami z (b) oraz (d).

Zadanie 2.6

Poniższa tabela zawiera informacje o połowie sardeli (w milionach ton) oraz średniej cenie ryb (w $ za tonę) w latach 1965-1978.

	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978
Cena (y)	190	160	134	129	172	197	167	239	542	372	245	376	454	410
Połów (x)	7,23	8,53	9,82	10,26	8,96	12,27	10,28	4,45	1,78	4,0	3,3	4,3	0,8	0,5

a. Wprowadź dane do pakietu GRETL

b. Narysuj wykres zmian w czasie

c. Znajdź wartości oszacowań MNK dla modelu określającego zależność między ceną a połowem

d. Zinterpretuj wartości oszacowań.

e. Oblicz wartość oszacowania wariancji składnika losowego s^2

f. Oblicz błąd szacunku dla $\widehat{\beta_1}$. Oblicz średni względny błąd szacunku i oceń precyzję oszacowania.

g. Oblicz 99% oszacowanie przedziałowego dla parametru β_1

Zadanie 2.7

Jak edukacja wpływa na zarobki? Plik `cps5.gdt` zawiera dane o stawce godzinowej, wykształceniu i innych zmiennych zebranych w Current Population Survey (CPS) z 2008 roku.

a. Oblicz statystyki opisowe i zbuduj histogramy dla zmiennych $WAGE$ i $EDUC$.
Opisz charakterystykę tych danych.

b. Oszacuj model liniowy wpływu wykształcenia na zarobki. Zinterpretuj wyniki.

c. Oszacuj reszty i zbuduj wykres reszt względem wykształcenia.
Czy coś na tym wykresie wygląda niepokojąco? Czy obserwujemy jakiś wzór?
Jeżeli spełnione są A1-A4, to czy powinnismy obserwować jakiś wzór w rozkładzie reszt?

d. Dodaj zmienne *black*, *exper*, *female*, *faminc* oraz *south* jako dodatkowe zmienne objaśniające.
Oszacuj parametry tego modelu i zinterpretuj wpływ poszczególnych zmiennych na zarobki.

e. Dla każdego oszacowania oblicz względny błąd szacunku i oceń prezycję tych oszacowań.

f. Porównaj skorygowany R-kwadrat oraz kryterium AIC między modelami z punktów b. i d.
Który model jest lepiej dopasowany do danych?

Zadanie 2.8

Dane o nieruchomościach sprzedawanych w Stockton, California zawarte są w pliku `stockton5.gdt`. Dostępne zmienne to $SPRICE$ ($) – cena domu, $LIVAREA$ (hundreds of square feet) - powierzchnia, $BEDS$- liczba sypialni, $BATHS$ – liczba łazienek, $LGELOT = 1$, jeżeli powierzchnia działki jest większa niż 0.5 ara, AGE – wiek domu i $POOL = 1$, jeżeli jest basen.

a. Stwórz histogram dla zmiennej $PRICE$. Co obserwujesz?

b. Oszacuj parametry modelu objaśniającego $PRICE$ przez pozostałe zmienne.
Zinterpretuj oszacowania.

c. Zinterpretuj wartość R-kwadrat. Jeżeli mielibyśmy dostęp do innych zmiennych, to jakie czynniki (inne niż te wykorzystane w zadaniu) mają wpływ na cenę mieszkań? Jak możemy je zmierzyć?

d. Dla każdego regresora, podaj 95% przedział ufności dla parametru. Formalnie zinterpretuj te przedziały.

Zadanie 2.9

W pliku `TaylorRule.gdt` zawarte są dane o poziomie stopy procentowej (*IR*, w %), inflacji rocznej (*INF*, %) oraz indeksu aktywności gospodarczej (*Y*, 100 jeżeli normalny poziom aktywności) dla wybranych krajów OECD. Badania ekonomiczne wskazują, że banki centralne ustalają poziom stopy procentowej w zależności od poziomu inflacji oraz aktywności gospodarczej

$$IR_t = \beta_0 + \beta_1 INF_t + \beta_2 Y_t + \varepsilon_t$$

a. Wybierz kraj, który będziesz analizował

b. Oszacuj parametry modelu dla stopy procentowej

c. Dokonaj interpretacji oszacowań parametrów β_1 i β_2

d. Podaj 95% przedział ufności dla parametru β_1

e. Oblicz i zinterpretuj wartość współczynnika R^2

Temat 3

Istotność zmiennych objaśniających

Marcin Topolewski

- Hipoteza statystyczna
- Budowa testu statystycznego
- Błędy I i II rodzaju
- Wartość krytyczna, wartość-p
- Test istotności t-Studenta
- Uogólniony test Walda istotności modelu

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Etapy weryfikacji modelu

1. Oceny parametrów i ich znaki
2. **Istotność parametrów**
3. Dopasowanie modelu do danych
4. Specyfikacja modelu / postać funkcyjna
5. Własności składnika losowego
6. Stabilność parametrów

2

Statystyczna weryfikacja hipotez

Test statystyczny pozwala zweryfikować przyjętą hipotezę dotyczącą populacji na podstawie informacji zawartej w próbie (dostępnych obserwacji).

W modelu ekonometrycznym hipotezy są zwykle pewnymi założeniami dotyczącymi wartości parametrów modelu.

Należy podkreślić, że **hipotezy dotyczą zawsze parametrów, a nie ich oszacowań** (własności populacji, a nie próby), ale własności parametrów weryfikuje się właśnie na podstawie ich oszacowań (wnioskowanie statystyczne).

BUDOWA TESTU STATYSTYCZNEGO

1. Hipoteza zerowa H_0
2. Hipoteza alternatywna H_1
3. Statystyka testowa
4. Obszar odrzuceń (lub wartość p)
5. Decyzja (wniosek)

Testowanie hipotez

HIPOTEZA ZEROWA

Hipoteza zerowa, H_0, zwykle dotyczy wartości parametru i może zawierać znak równości, na przykład:

$$H_0: \beta_k = c$$

gdzie stała c jest określoną wartością, specyficzną dla badanego modelu statystycznego.

H_0 zostaje odrzucona tyko wtedy, gdy są do tego wystarczająco silne przesłanki statystyczne.

(Nierówności nieostre $\beta_k \geq c$ lub $\beta_k \leq c$ również mogą stanowić hipotezę zerową)

HIPOTEZA ALTERNATYWNA

Hipoteza alternatywna, H_1, jest logicznym dopełnieniem hipotezy zerowej, zatem opisuje zdarzenie komplementarne do opisanego przez hipotezę zerową, na przykład:

$$H_1: \beta_k \neq c$$

(Jeżeli hipotezę zerową stanowią nierówności nieostre $\beta_k \geq c$ lub $\beta_k \leq c$, hipoteza alternatywna określona jest przez nierówności ostre, odpowiednio $\beta_k < c$ lub $\beta_k > c$)

Testowanie hipotez

STATYSTYKA TESTOWA

Statystyka testowa to wartość obliczona na podstawie obserwacji w próbie, na podstawie której podejmujemy decyzję o ewentualnym odrzuceniu hipotezy zerowej.

Co ważne, przy założeniu prawdziwości hipotezy zerowej rozkład statystyki testowej jest znany. Pozwala to na obliczenie prawdopodobieństwa popełnienia błędu odrzucenia prawdziwej hipotezy zerowej.

OBSZAR ODRZUCEŃ

Obszar odrzuceń to obszar wartości nietypowych dla statystyki testowej – przy założeniu prawdziwości hipotezy zerowej, występujących z ustalonym prawdopodobieństwem.

Aby określić obszar odrzuceń musimy znać:

- Statystykę testową, której rozkład jest znany przy założeniu prawdziwości hipotezy zerowej
- Hipotezę alternatywną
- **Poziom istotności** α

Poziom istotności testu α, to prawdopodobieństwo popełnienia ***błędu I-go rodzaju***, to znaczy odrzucenia prawdziwej hipotezy zerowej. Poziom istotności α ustala się na odpowiednio niskim poziomie, konkretnie 0,01, 0,05 lub 0,10.

Poziom istotności wyznacza obszar odrzuceń.

Testowanie hipotez

Błędy I-go i II-go rodzaju

	Stan faktyczny w populacji	
Decyzja na podstawie próby	H_0 prawdziwa	H_0 fałszywa
Brak odrzucenia H_0	**Decyzja prawidłowa** (prawdopodobieństwo = 1 - α)	**Błąd II-go rodzaju** Brak odrzucenia H_0 gdy jest fałszywa (prawdopodobieństwo = β)
Odrzucenie H_0	**Błąd I-go rodzaju** Odrzucenie H_0 gdy jest prawdziwa (prawdopodobieństwo = α)	**Decyzja prawidłowa** Określa moc testu (prawdopodobieństwo = 1 - β)

Moc testu (1 - β) to prawdopodobieństwo odrzucenia H_0, gdy jest ona fałszywa.
Mocy testu zazwyczaj nie możemy wyznaczyć analitycznie, jedynie za pomocą symulacji.

Testowanie hipotez

DECYZJA (WNIOSEK)

W teście statystycznym są tylko dwie możliwe decyzje:

- Odrzucić hipotezę zerową
- Nie odrzucać hipotezy zerowej

Jeżeli statystyka testowa przyjmuje wartość z obszaru odrzuceń, jest mało prawdopodobne, że hipoteza zerowa jest prawdziwa, a zatem należy ją odrzucić, w przeciwnym przypadku nie odrzucamy H_0. Ostatecznie należy wyjaśnić, jakie znaczenie ma wynik testu w kontekście badanego problemu i jaka jest jego interpretacja (np. ekonomiczna).

WAŻNE: brak podstaw do odrzucenia hipotezy zerowej nie oznacza, że jest ona prawdziwa

PROCEDURA STATYSTYCZNEGO TESTOWANIA HIPOTEZ

1. Określ hipotezę zerową i alternatywną.
2. Ustal statystykę testową i jej rozkład (w przypadku prawdziwości hipotezy zerowej).
3. Wybierz poziom istotności i ustal obszar odrzuceń.
4. Oblicz wartość statystyki testowej na podstawie próby.
5. Podejmij decyzję odnośnie hipotezy zerowej.

Wartość-*p* (*p*-Value)

- W praktyce decyzje w teście statystycznym podejmuje się zwykle na podstawie **wartości-p**, czyli tzw. empirycznego poziomu istotności.
- **Wartość-*p*** to graniczny poziom istotności przy którym odrzucamy hipotezę zerową (jeżeli $\alpha \leq$wartość-*p to pozostajemy przy* H_0). Inaczej mówiąc wartość-*p* to prawdopodobieństwo popełnienia błędu I-go rodzaju przy odrzuceniu H_0.
- Zatem H_0 należy odrzucić tylko, jeśli wartość-*p* jest odpowiednio niska (np. poniżej $\alpha = 0{,}05$).

REGUŁA WARTOŚCI-*p*:

Hipotezę zerową należy odrzucić tylko, jeśli wartość-*p* jest równa lub niższa od przyjętego poziomu istotności α, czyli:

jeżeli wartość-*p* $\leq \alpha \Rightarrow$ należy odrzucić H_0.

jeżeli wartość-*p* $> \alpha \Rightarrow$ nie ma podstaw do odrzucenia H_0.

Oznacza to, że odrzucamy hipotezę zerową, tylko gdy prawdopodobieństwo popełnienia błędu I-go rodzaju jest mniejsze niż α.

Test istotności pojedynczej zmiennej objaśniającej

- W modelu regresji weryfikujemy istotność wpływu poszczególnych regresorów x_k na zmienną zależną y poprzez weryfikacje hipotezy, czy parametr β_k jest statystycznie różny od zera.
- W tym celu używamy testu *t-studenta*.

TEST *t-studenta*

Hipoteza zerowa

$H_0: \beta_k = 0$ (wpływ zmiennej x_k jest nieistotny statystycznie)

Hipoteza alternatywna

$H_1: \beta_k \neq 0$ (wpływ zmiennej x_k jest istotny statystycznie)

Statystyka testowa

$$t = \frac{\hat{\beta}_k}{S(\hat{\beta}_k)} \sim t_{(N-K-1)}$$

Wartość krytyczna

$$t_c = t_{(1-\alpha/2, N-K-1)}$$

Decyzja

odrzucić H_0 jeśli $|t| \geq t_c$ (w przeciwnym przypadku, nie odrzucać H_0) lub

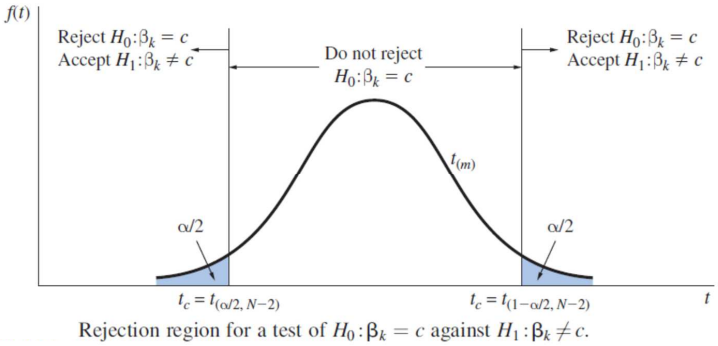
odrzucić H_0 jeśli $p \leq \alpha$ (w przeciwnym przypadku, nie odrzucać H_0)

Test istotności pojedynczej zmiennej objaśniającej

Interpretacja graficzna testu *t-studenta*

Reguły decyzyjne można zilustrować graficznie

(wartości krytyczne dla modelu ze stałą i jedną zmienną objaśniającą)



Rejection region for a test of $H_0: \beta_k = c$ against $H_1: \beta_k \neq c$.

Źródło: Principles of Econometrics, R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

Przykład 3.1: Test t-studenta

Na podstawie danych `andy.gdt` zbudowano model regresji liniowej wyjaśniający jak miesięczna wartość sprzedaży w `Big Andy's Burger Barn*` zależy od cen i wydatków na reklamę.

```
Model : Estymacja KMNK, wykorzystane obserwacje 1-75
Zmienna zależna (Y): sales

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const      118,914          6,35164         18,72     2,21e-029 ***
  price       -7,90785        1,09599         -7,215    4,42e-010 ***
  advert       1,86258        0,683195         2,726    0,0080    ***
```

Pytania:

1. Jaka jest wartość krytyczna
2. Jak jest decyzja na podstawie wartości krytycznej
3. Jak jest decyzja na podstawie wartości-*p*
4. Zinterpretuj wyniki testów

Testowanie wartości pojedynczego współczynnika

Można również testować, czy określony współczynnik jest równy zakładanej wcześniej wartości, powiedzmy c. Wówczas test *t-studenta* wygląda następująco

t-student test

Hipoteza zerowa

$$H_0: \beta_k = c$$

Hipoteza alternatywna

$$H_1: \beta_k \neq \text{c}$$

Statystyka testowa

$$t = \frac{\widehat{\beta}_k - c}{S(\widehat{\beta}_k)} \sim t_{(N-K-1)}$$

Wartość krytyczna

$$t_c = t_{(1-\alpha/2, N-K-1)}$$

Decyzja

odrzucić H_0 jeśli $|t| \geq t_c$ (w przeciwnym przypadku, nie odrzucać H_0) lub

odrzucić H_0 jeśli $p \leq \alpha$ (w przeciwnym przypadku, nie odrzucać H_0)

Testowanie hipotez łącznych

- Możemy również testować hipotezy zawierające przypuszczenia odnośnie więcej niż jednego parametru (z więcej niż jednym znakiem równości), czyli tak zwane **hipotezy łączne**.
- Zwykle jesteśmy zainteresowani, czy grupa zmiennych objaśniających powinna znaleźć się w specyfikacji modelu
- Wówczas porównujemy model z restrykcjami z modelem bez restrykcji.

Model bez restrykcji

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \beta_{K+1} x_{K+1i} + \cdots + \beta_S x_{Si} + \varepsilon_i$$

Model z restrykcjami

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Restrykcje

$$\beta_{K+1} = \cdots = \beta_s = 0$$

czyli zmienne od x_{K+1} do x_S są nieistotne statystycznie

Testowanie hipotez łącznych

Wprowadźmy oznaczenia:

SSE_U	suma kwadratów reszt dla modelu bez restrykcji
SSE_R	suma kwadratów reszt dla modelu z restrykcjami
$J = S - K$	liczba restrykcji zerowych

Test łącznej istotności zmiennych

Hipoteza zerowa

$H_0: \beta_{K+1} = \cdots = \beta_s = 0$ (dodatkowe zmienne są nieistotne)

Hipoteza alternatywna

$H_1: \beta_{K+1} \neq 0$ lub ... lub $\beta_s \neq 0$ (przynajmniej jedna zmienna dodatkowa jest istotna)

Statystyka testowa

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K-1)} \sim F_{(J,N-K-1)}$$

Wartość krytyczna

$$F_c = F_{(1-\alpha,J,N-K-1)}$$

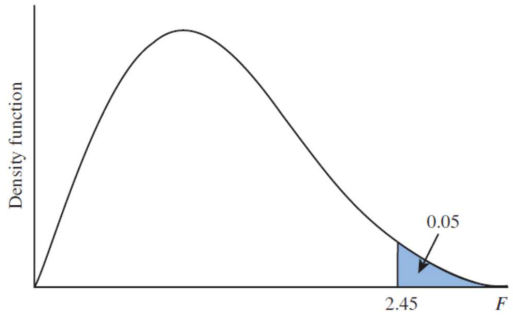
Decyzja

odrzucić H_0 jeśli $F \geq F_c$ (w przeciwnym przypadku, nie odrzucać H_0) lub

odrzucić H_0 jeśli $p \leq \alpha$ (w przeciwnym przypadku, nie odrzucać H_0)

Testowanie hipotez łącznych

Interpretacja graficzna testu-*F*



The probability density function of an $F_{(8,20)}$ random variable.

Źródło: Principles of Econometrics, R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

Przykład 3.2: Testowanie hipotez łącznych

Używając danych `andy.gdt` zbudowano model regresji liniowej wyjaśniający jak miesięczna wartość sprzedaży w `Big Andy's Burger Barn`* zależy od cen, wydatków na reklamę i kwadratu wydatków na reklamę. Oceń łączny wpływ wydatków na reklamę na wartość sprzedaży.

Model bez restrykcji

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	109,719	6,79905	16,14	1,87e-025	***
price	−7,64000	1,04594	−7,304	3,24e-010	***
advert	12,1512	3,55616	3,417	0,0011	***
sq_advert	−2,76796	0,940624	−2,943	0,0044	***

Średn.aryt.zm.zależnej	77,37467	Odch.stand.zm.zależnej	6,488537
Suma kwadratów reszt	1532,084	Błąd standardowy reszt	4,645283
Wsp. determ. R-kwadrat	0,508235	Skorygowany R-kwadrat	0,487456

Model z restrykcjami

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	121,900	6,52629	18,68	1,59e-029	***
price	−7,82907	1,14286	−6,850	1,97e-09	***

Średn.aryt.zm.zależnej	77,37467	Odch.stand.zm.zależnej	6,488537
Suma kwadratów reszt	1896,391	Błąd standardowy reszt	5,096858
Wsp. determ. R-kwadrat	0,391301	Skorygowany R-kwadrat	0,382963

Przykład 3.2 : Testowanie hipotez łącznych c.d.

Pytanie. Czy wydatki na reklamę istotnie wpływają na sprzedaż?

Test-*F*

Hipotezy testowe:

$H_0: \beta_2 = \beta_3 = 0$ (zmienne `advert` i `sq_advert` są nieistotne statystycznie)
$H_1: \beta_2 \neq 0 \; or \; \beta_3 \neq 0$ (przynajmniej jedna z tych zmiennych jest istotna statystycznie)

Statystyka testowa:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K-1)} = \frac{(1896.391 - 1532.084)/2}{1532.084/(75-4)} = 8.44136$$

Wartość krytyczna:

$$F_c = F_{(0.95,2,71)} = 3.12576$$

Decyzja:

ponieważ $F > F_c$ odrzucamy hipotezę zerową.
Przynajmniej jedna ze zmiennych `advert`, `sq_advert` jest istotna statystycznie.
Wydatki na reklamę mają istotny wpływ na wartość sprzedaży.

Przykład 3.2: Testowanie hipotez łącznych c.d.

Pytanie. Czy wydatki na reklamę istotnie wpływają na sprzedaż?

Gretl: Test pominiętych/dodanych zmiennych

W `Gretl`:

Zbadaj łączną istotność wydatków na reklamę testem „pominiętych zmiennych” w modelu bez restrykcji:

->Testy->Test pominiętych zmiennych

1. Sprawdź statystykę testową i porównaj z poprzednim przykładem
2. Sprawdź wartość-*p*
3. Podejmij decyzje na podstawie wartości-*p*

```
Test for omission of variables -
  Null hypothesis: parameters are zero for the variables
    advert
    sq_advert
  Test statistic: F(2, 71) = 8.44136
  with p-value = P(F(2, 71) > 8.44136) = 0.000514159
```

Przykład 3.2: Testowanie hipotez łącznych c.d.

Pytanie. Czy wydatki na reklamę istotnie wpływają na sprzedaż?

Gretl: Test pominiętych/dodanych zmiennych

W `Gretl`:

Zbadaj łączną istotność wydatków na reklamę testem „dodanych zmiennych" dla modelu z restrykcjami:

->Testy->Test dodanych zmiennych

1. Sprawdź statystykę testową i porównaj z poprzednim przykładem
2. Sprawdź wartość-*p*
3. Podejmij decyzje na podstawie wartości-*p*

```
Test for addition of variables -
  Null hypothesis: parameters are zero for the variables
    advert
    sq_advert
  Test statistic: F(2, 71) = 8.44136
  with p-value = P(F(2, 71) > 8.44136) = 0.000514159
```

Testowanie istotności modelu

Za pomocą testu-*F* można również badać istotność całego modelu regresji. Ta wersja testu nosi nazwę **uogólnionego testu Walda** i może być interpretowana jako test istotności współczynnika determinacji R^2.

Uogólniony test Walda

Hipoteza zerowa

$H_0: \beta_1 = \cdots = \beta_K = 0$ (wszystkie zmienne w modelu są nieistotne)

Hipoteza alternatywna

$H_1: \beta_1 \neq 0 \ \vee \ \ldots \vee \beta_K \neq 0$ (co najmniej jedna zmienna w modelu jest istotna)

Statystyka testowa

$F = \frac{(SSE_R - SSE_U)/K}{SSE_U/(N-K-1)} \sim F_{(K,N-K-1)}$ (alternatywnie $F = \frac{R^2/K}{(1-R^2)/(N-K-1)} \sim F_{(K,N-K-1)}$)

Wartość krytyczna

$F_c = F_{(1-\alpha,K,N-K-1)}$

Decyzja

odrzucić H_0 jeśli $F \geq F_c$ (w przeciwnym przypadku, nie odrzucać H_0) lub

odrzucić H_0 jeśli $p \leq \alpha$ (w przeciwnym przypadku, nie odrzucać H_0)

Przykład 3.3: Uogólniony test Walda

Test na istotność modelu sprzedaży z przykładu 3.1.

Test istotności modelu

Model 1: OLS, using observations 1-75

Dependent variable: sales

	Coefficient	Std. Error	t-ratio	p-value	
const	118.914	6.35164	18.72	<0.0001	***
Price	−7.90785	1.09599	−7.215	<0.0001	***
advert	1.86258	0.683195	2.726	0.0080	***

Mean dependent var	77.37467	S.D. dependent var	6.488537
Sum squared resid	1718.943	S.E. of regression	4.886124
R-squared	0.448258	Adjusted R-squared	0.432932
F(2, 72)	29.24786	P-value(F)	5.04e-10
Log-likelihood	−223.8695	Akaike criterion	453.7390
Schwarz criterion	460.6915	Hannan-Quinn	456.5151

Zadania

Zadanie 3.1

Na podstawie 5 rocznych obserwacji dla wydatków pewnego gospodarstwa domowego [tys PLN]: $y = [10\ 8\ 10\ 16\ 26]'$:

a. Oblicz wartości macierzy $\mathbf{X'X}$, $\mathbf{X'y}$ oraz $(\mathbf{X'X})^{-1}$

b. Oszacuj parametry modelu trendu
$$y_t = \beta_0 + \beta_1\, t + \varepsilon_t$$

c. Oblicz reszty modelu oraz oszacowanie wariancji składnika losowego s^2

d. Oblicz oszacowanie wariancji estymatora MNK: $\widehat{\boldsymbol{\Sigma}_{\hat{\beta}}} = \widehat{Var(\hat{\boldsymbol{\beta}})} = s^2(\mathbf{X'X})^{-1}$ oraz średnie błędy szacunku

e. Dokonaj weryfikacji hipotezy H_0: $\beta_1 = 0$ [wartość krytyczna to $t_{3,5\%} = 3.18$]

f. Wpisz dane do programu GRETL i sprawdź, czy uzyskałeś takie same wyniki

Zadanie 3.2

Departament pożyczek Banku Warszawskiego zamierza ustalić jak wartość kredytów hipotecznych (*M*, € per capita) zależy od aktywności gospodarczej (*PKB*, € per capita) i referencyjnej stopy procentowej (*R*, %) w różnych krajach europejskich. Na podstawie danych dla 24 krajów analitycy oszacowali parametry modelu:

$$\widehat{M}_i = 23278 + 5.75PKB_i - 1251R_i$$
$$(se) \quad (1479) \quad (3.47) \qquad (498)$$

a. Zinterpretuj oszacowania parametrów.

b. Uzupełnij zdania:
   - Oszacowanie parametru β_1 wynosi…
   - Błąd szacunku parametru β_1 wynosi…
   - Statystyka testu istotności parametru β_1 wynosi…

c. Zweryfikuj hipotezę, że PKB nie ma wpływu na wartość kredytów hipotecznych.

d. Zweryfikuj hipotezę, że stopa procentowa nie ma wpływu na wartość kredytów hipotecznych.

e. Zweryfikuj hipotezę, że parametr przy stropie procentowej wynosi -2500.

Zadanie 3.3

Plik `cps5.gdt` zawiera dane o stawce godzinowej, wykształceniu i innych zmiennych zebranych w Current Population Survey (CPS) z 2008 roku.

a. Oszacuj parametry równania dochodów

$$WAGE_i = \beta_0 + \beta_1 EDUC_i + \beta_2 EXPER_i + \varepsilon_i$$

b. Zbadaj istotność modelu oraz każdej ze zmiennych osobno

c. Czy na podstawie danych możemy odrzucić hipotezę, że z każdy rokiem wykształcenia prowadzi do wzrostu stawki godzinowej o 2.5USD?

d. Rozważ wprowadzenie do modelu kwadratów zmiennych $EDUC$ i $EXPER$

$$WAGE_i = \beta_0 + \beta_1 EDUC_i + \beta_2 EXPER_i + \beta_3 EDUC_i^2 + \beta_4 EXPER_i^2 + \varepsilon_i$$

oraz zbadaj łączną istotność wykształcenia oraz doświadczenia.

e. Czy położenie geograficzne jest istotnym czynnikiem wyjaśniającym płace? (jakich zmiennych należy użyć?)

Zadanie 3.4

Dane o nieruchomościach sprzedawanych w Stockton, California zawarte są w pliku `stockton5.gdt`. Dostępne zmienne to $SPRICE$ ($) – cena domu, $LIVAREA$ (hundreds of square feet) - powierzchnia, $BEDS$- liczba sypialni, $BATHS$ – liczba łazienek, $LGELOT = 1$, jeżeli powierzchnia działki jest większa niż 0.5 ara, AGE – wiek domu i $POOL = 1$, jeżeli jest basen.

a. Zaproponuj i oszacuj parametry modelu objaśniającego ceny nieruchomości

b. Zbadaj istotność modelu i istotność każdej zmiennej osobno

c. Zweryfikuj hipotezę, że cena jednostki powierzchni (100ft2) wynosi 10000USD

d. Porównaj dopasowanie pełnego modelu z modelem z jedną zmienną objaśniającą

$$SPRICE_i = \beta_0 + \beta_1 LIVAREA + \varepsilon_t$$

Zadanie 3.5

W pliku `TaylorRule.gdt` zawarte są dane o poziomie stopy procentowej (*IR*, w %), inflacji rocznej (*INF*, %) oraz indeksu aktywności gospodarczej (*Y*, 100 jeżeli normalny poziom aktywności) dla wybranych krajów OECD. Badania ekonomiczne wskazują, że banki centralne ustalają poziom stopy procentowej w zależności od poziomu inflacji oraz aktywności gospodarczej

a. Wybierz kraj, który będziesz analizował

b. Oszacuj parametry następującego modelu

$$IR_t = \beta_0 + \beta_1 INF_t + \beta_2 Y_t + \varepsilon_t$$

c. Oceń, które zmienne są statystycznie istotne.

d. Czy cały model jest statystycznie istotny?

e. Dokonaj weryfikacji hipotezy $H_0: \beta_1 = 1{,}5$

f. Dokonaj weryfikacji hipotezy $H_0: \beta_1 = 1{,}5 \ \wedge \beta_2 = 0{,}5$?

Zadanie 3.6

W pliku `PhillipsCurve.gdt` zawarte są dane o inflacji rocznej (INF, %) oraz stopy bezrobocia (U, %) dla wybranych krajów UE. Teoria ekonomii wskazuje na ujemną zależność między obydwoma zmiennymi.

a. Wybierz kraj, który będziesz analizował

b. Oszacuj parametry następującego modelu

$$INF_t = \beta_0 + \beta_1 U_t + \varepsilon_t$$

c. Oceń istotność zmiennej U_t.

d. Rozszerz specyfikacje modelu o zmienną opisującą inflację w Niemczech:

$$INF_t = \beta_0 + \beta_1 U_t + \beta_2 INF_t^{DE} \varepsilon_t$$

oraz porównaj dopasowanie obydwu modeli do danych.

e. Dla rozszerzonego modelu dokonaj weryfikacji hipotezy $H_0: \beta_1 = 0 \ \wedge \beta_2 = 1$?

Temat 4

Specyfikacja modelu ekonometrycznego

Karolina Konopczak i Michał Rubaszek

- Model nieliniowy
- Błąd specyfikacji modelu
- Efekt krańcowy i elastyczność
- Modele wielomianowe
- Modele z logarytmami zmiennych
- Test specyfikacji Ramsey'a (RESET)
- Zmienne binarne i zmienne interakcyjne

Modele nieliniowe

- Teoria ekonomiczna rzadko określa dokładną formę funkcyjną zależności między zmiennymi.
- **Specyfikacja liniowa jest często punktem wyjścia do modelowania zależności między zmiennymi ekonomicznych**. Jest ona jednak jedynie przybliżeniem dla rzeczywistych relacji. Przybliżenie to jest zazwyczaj wystarczające, jeżeli wnioskowanie jest prowadzone dla wąskiego przedziału zmienności.
- Ale: nieuwzględnienie nieliniowości może stanowić błąd specyfikacji

Zagadnienia omawiane w Temat 4:

1. Jak wykryć błąd specyfikacji?
2. Jak usunąć błąd specyfikacji poprzez zmianę postaci funkcyjnej modelu?
3. Jak porównać konkurencyjne specyfikacje nieliniowe?
4. Jak interpretować parametry w modelach nieliniowych?
5. Jak uwzględnić niestabilność zależności między zmiennymi w próbie?

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. **Wybór postaci funkcyjnej**
3. Zebranie danych
4. Estymacja
5. Weryfikacja
6. Zastosowanie

Błąd specyfikacji

Przyczyny złej specyfikacji modelu:

- pominięcie istotnej zmiennej objaśniającej (zbyt uboga specyfikacja)
- włączenie nieistotnej zmiennej objaśniającej (zbyt obszerna specyfikacja)
- zła postać funkcyjna modelu
- błąd pomiaru zmiennych

Skutki błędu specyfikacji:

- nietrafne prognozy (szerzej w Temat 10)
- obciążenie estymatora parametrów (szerzej w Temat 12-14)

Jak wykryć błąd specyfikacji:

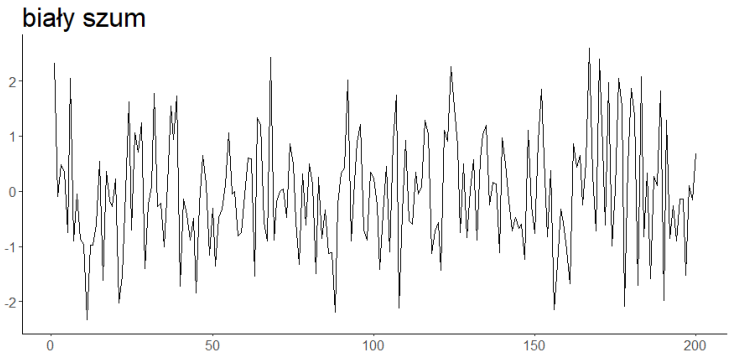
- Test RESET, tj. test Ramsey'a
- Test pominiętej zmiennej (ang. *omitted variable*)
- Analiza wykresów zależności między zmiennymi

Błąd specyfikacji

- Modelowanie danych wymaga wydobycia informacji z szumu
- Model ekonometryczny działa jak filtr.
- To, co pozostało po przefiltrowaniu danych, powinno być czysto losowe:

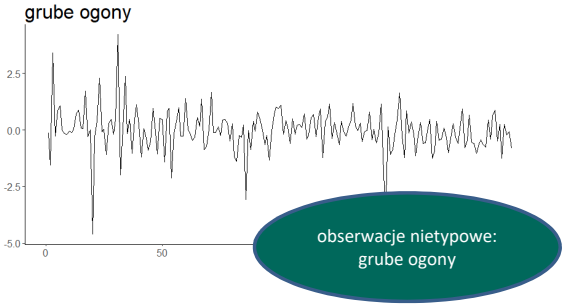
$$\varepsilon_t \sim IID\ N(0, \sigma^2)$$

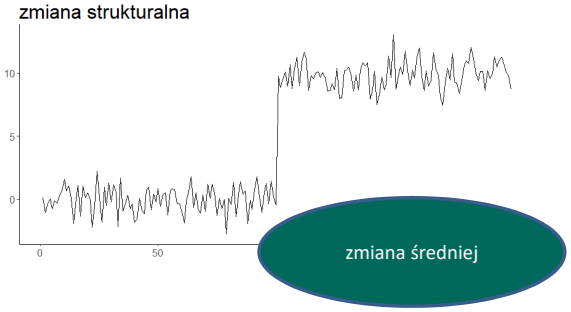
- Wszelkie nielosowe wzorce dla reszt modelu wskazują na błąd specyfikacji.



Stała średnia,
stała wariancja,
brak zmiany strukturalnej

Błąd specyfikacji





Nieliniowości

Rodzaje nieliniowości:

- "krzywoliniowość" (pochodna y względem x nie jest stałą)
- asymetria reakcji na wzrosty / spadki
- zmiany strukturalne
  - w czasie
  - między klasami obiektów

Metody uwzględnienia nieliniowości w modelu ekonometrycznym:

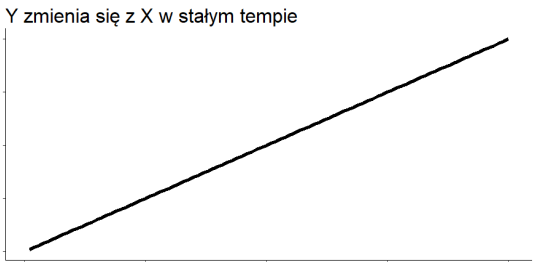
- przekształcanie zmiennych w ramach regresji liniowej (potęgi, logarytmy, odwrotności)
- wprowadzenie zmiennych binarnych lub interakcyjnych
- modele przełącznikowe (switching models)
- modele progowe (threshold models)
- regresje nieparametryczne
- inne

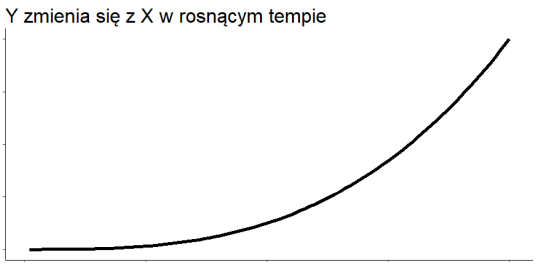
Nieliniowości

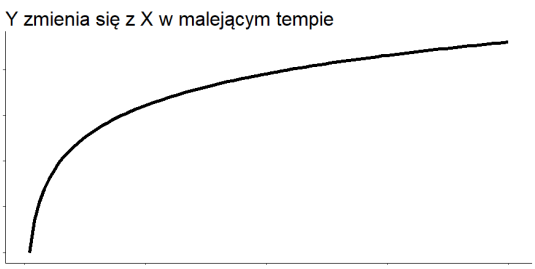
Teoria często wskazuje na możliwe odstępstwa od liniowości w ekonomii, np .:

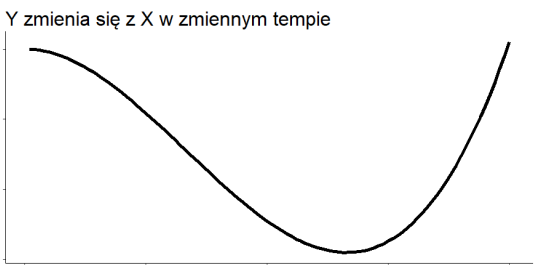
- malejąca krańcowa skłonność do konsumpcji
  wpływ dodatkowego dochodu na konsumpcję spada wraz z dochodem
- Krzywa Laffera
  "zgarbiona" zależność między stawkami podatkowymi a dochodami budżetu
- Metoda „rakiet i piór” przy ustalaniu ceny
  ceny reagują szybciej („jak rakiety”) na wzrost kosztów niż („jak pióra”) na ich spadek
- „Krzywa uśmiechu” w globalnych łańcuchach wartości dodanej (global value chains)
  dwa końce łańcucha - badania i rozwój oraz marketing – dają większą marżę niż środkowa część łańcucha - produkcja

Nieliniowości









Podstawowe miary zależności nieliniowych

Efekt krańcowy

Interpretacja: zmiana Y wywołana jednostkową zmianą X

$$\mathrm{ME}_{Y/X} = \frac{\partial Y}{\partial X}$$

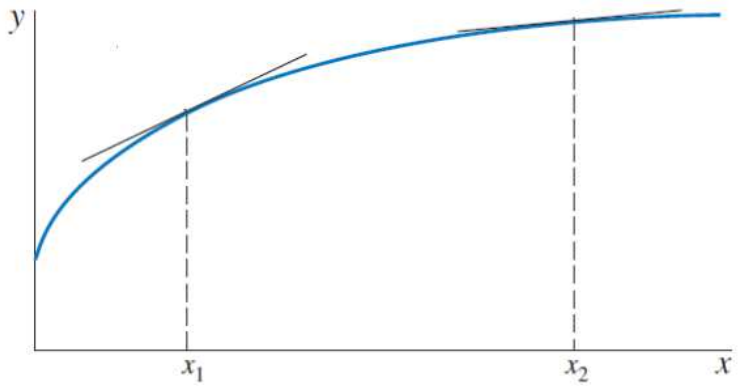
Elastyczność

Interpretacja: procentowa zmiana Y wywołana procentową zmianą X

$$\mathrm{E}_{Y/X} = \frac{\partial Y/Y}{\partial X/X} = \frac{\partial \ln Y}{\partial \ln X} = \frac{\partial Y}{\partial X} \times \frac{X}{Y} = \mathrm{ME}_{Y/X} \times \frac{X}{Y}$$

Efekt krańcowy - ilustracja

Efekt krańcowy: nachylenie stycznej do krzywej (tj. pochodnej) w danym punkcie::



W relacjach liniowych efekt krańcowy jest stały, w nieliniowych zmienny

Nieliniowe zależności w modelu ekonometrycznym

Czy nieliniowości można włączyć do regresji MNK?

- nieliniowość względem zmiennych → TAK

  np. funkcja kwadratowa: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$
- nieliniowość względem parametrów → NIE

  np. wykładnicza funkcja wzrostu: $y_t = e^{\beta_0 + \beta_1 t} + \varepsilon_t$

Popularne transformacje stosowane w regresjach MNK:

- potęgi: x_i^2, x_i^3, …
- logarytmy: $\ln(x_i)$
- odwrotności: $\frac{1}{x_i}$

Jak oszacować parametry w modelach nieliniowych względem zmiennych, np.:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

- zauważ, że $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$, gdzie $z_i = x_i^2$
- użyj estymatora MNK

Jak wybrać właściwą postać nieliniową modelu?

- **Łatwo, jeśli modele są zagnieżdżone, np .:**

  model A: $y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \varepsilon_i$

  model B: $y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3 + \varepsilon_i$
  - testujemy $H_0: \alpha_3 = 0$
- **Trudniej, jeśli modele nie są zagnieżdżone, np .:**

  model A: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

  model B: $y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$
  - kryteria informacyjne (AIC, BIC)
  - skorygowane R^2
- **Uwagi ogólne:**
  1. Nigdy nie poznamy prawidłowej zależności funkcjonalnej
  2. Próbujemy wybrać postać, która jest zgodna z teorią ekonomii, zaś model jest dobrze dopasowana do danych
  3. Modele można porównać na podstawie kryteriów wyboru tylko wtedy, gdy mają tę samą zmienną zależną i zakres próby

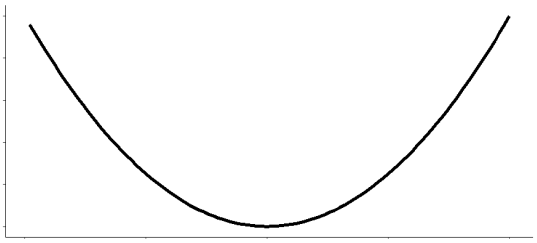
Modele wielomianowe

Zależność kwadratowa:

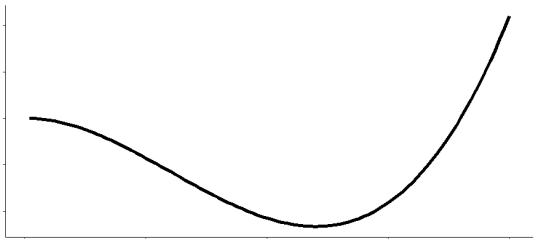
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

$$ME_{y_i/x_i} = \frac{\partial y_i}{\partial x_i} = \beta_1 + 2\beta_2 x_i = f(x_i)$$

Zależność kwadratowa



Zależność sześcienna



Zależność sześcienna:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

$$ME_{y_i/x_i} = \frac{\partial y_i}{\partial x_i} = \beta_1 + 2\beta_2 x_i + 3\beta_3 x_i^2 = f(x_i)$$

Wielomiany wyższego rzędu: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$

Przykład 4.1.
Modele wielomianowe

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry modelu objaśniającego stawkę godzinową (*wage*, USD), przez wykształcenie (*educ*, lata) oraz wiek (*age*, lata).
Wyniki regresji to:

```
Model: Estymacja KMNK, wykorzystane obserwacje 1-9799
Zmienna zależna (Y): wage

            współczynnik  błąd std     t-Studenta  wartość p
  ------------------------------------------------------------
  const     −35,022        1,658         −21,12    7,03e-097 ***
  educ        2,342        0,052          44,62    0,0000    ***
  age         1,038        0,073          14,15    5,23e-045 ***
  sq_age     −0,010        0,000         −11,65    3,56e-031 ***

Średn.aryt.zm.zależnej  23,46008   Odch.stand.zm.zależnej  16,07305
Suma kwadratów reszt     2006839   Błąd standardowy reszt  14,31377
Wsp. determ. R-kwadrat  0,207172   Skorygowany R-kwadrat   0,206929
```

Pytania:

1. jaki jest wpływ wieku na wynagrodzenie dla osoby w wieku 50 lat?
2. Dla jakiego wieku płace są najwyższe?
3. Jak najlepiej zmierzyć związek między płacą a wiekiem?

Logarytmy

Właściwości funkcji logarytmicznych

- Dodatnie argumenty: $\ln A,\ A > 0.$
- Iloczyn: $\ln(AB) = \ln A + \ln B$
- Iloraz: $\ln(A/B) = \ln A - \ln B$
- Potęga: $\ln\left(A^k\right) = k \times \ln A$
- Funkcja wykładnicza: $\ln(e^x) = x \times \ln e = x$ and $e^{\ln x} = x$

Zmiany logarytmów

Interpretacja $\Delta \ln X$: procentowa zmiana X.

Dlaczego?

$$\frac{\partial \ln X}{\partial X} = \frac{1}{X} \rightarrow \partial \ln X = \frac{\partial X}{X} \rightarrow \Delta \ln X \approx \frac{\Delta X}{X}$$

A zatem dla małych zmian: $\Delta \ln X \approx \frac{\Delta X}{X}$

Logarytmy

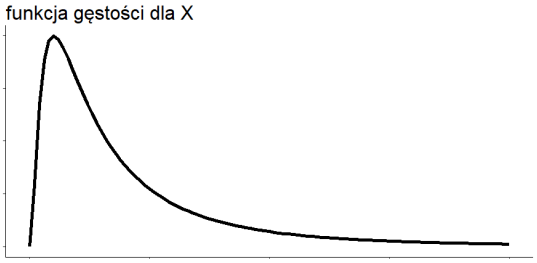
Logarytmy często stosuje się w modelowaniu :

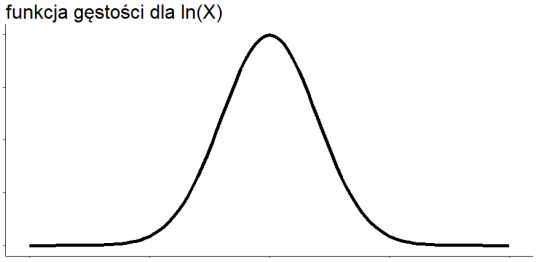
- płac
- dochodu
- cen
- sprzedaży
- wydatki

tj., zmiennych, których wartości są:

- dodatnie
- dodatnio skośne
  (z długim ogonem po prawej)

Po transformacji logarytmicznej takie zmienne mają rozkład normalny





Logarytmy

Model liniowy

Specyfikacja:	$y_i = \alpha + \beta x_i + \varepsilon_i$
Postać wyjściowa:	$y_i = \alpha + \beta x_i + \varepsilon_i$
Efekt krańcowy:	$ME_{y_i/x_i} = \beta$
Elastyczność:	$E_{y_i/x_i} = \beta \times x_i/y_i$
Interpretacja β:	$\Delta x_i = 1 \rightarrow \Delta y_i = \beta$

Model Log-liniowy (log-lin)

Specyfikacja:	$\ln y_i = \alpha + \beta x_i + \varepsilon_i$
Postać wyjściowa:	$y_i = e^{\alpha+\beta x_i+\varepsilon_i}$
Efekt krańcowy:	$ME_{y_i/x_i} = \beta \times y_i$
Elastyczność:	$E_{y_i/x_i} = \beta \times x_i$
Interpretacja β:	$\Delta x_i = 1 \rightarrow \Delta y_i = \beta \times y_i \rightarrow \frac{\Delta y_i}{y_i} = \beta \times 100\%$

Logarytmy

Model liniowo-logarytmiczny (lin-log)

Specyfikacja:	$y_i = \alpha + \beta \ln x_i + \varepsilon_i$
Postać wyjściowa:	brak
Efekt krańcowy:	$ME_{y_i/x_i} = \beta/x_i$
Elastyczność:	$E_{y_i/x_i} = \beta/y_i$
Interpretacja β:	$\Delta \ln x_i \left(\approx \frac{\Delta x_i}{x_i}\right) = 0.01 = 1\% \rightarrow \Delta y_i = \beta/100$

Model log-log

Specyfikacja:	$\ln y_i = \alpha + \beta \ln x_i + \varepsilon_i$
Postać wyjściowa:	$y_i = e^{\alpha+\beta \ln x_i+\varepsilon_i} = e^{\alpha} \times x_i^{\beta} \times e^{\varepsilon_i}$
Efekt krańcowy:	$ME_{y_i/x_i} = \beta \times y_i/x_i$
Elastyczność:	$E_{y_i/x_i} = \beta$
Interpretacja β:	$\frac{\Delta x_i}{x_i} = 1\% \rightarrow \frac{\Delta y_i}{y_i} = \beta\%$

Przykład 4.2.
Logarytmy

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry 3 modeli wyjaśniających stawki godzinowe (*wage*, USD) przez wykształcenie (*educ*, `lata`). Wyniki regresji to:

```
Model A: Zmienna zależna (Y): wage

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       −49,7269        1,84009         −27,02    3,50e-155 ***
  l_educ       27,7830        0,695974         39,92    0,0000    ***

Model B: Zmienna zależna (Y): l_wage

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       1,61483         0,0262872        61,43      0,0000   ***
  educ        0,09674         0,00181574       53,28      0,0000   ***

Model C: Zmienna zależna (Y): l_wage

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       0,00379         0,0625762        0,06071    0,9516
  l_educ      1,13360         0,0236680       47,90       0,0000   ***
```

Pytania:

1. Jaka jest interpretacja oszacowań parametrów dla zmiennych *educ* / *l_educ* ?
2. Jakie są efekty krańcowe / elastyczność dla osoby z 10-letnim wykształceniem?

Test specyfikacji Ramsey'a (RESET)

RESET = regression specification error test (Ramsey, 1969)

Etapy testu RESET

1. Oszacuj model wyjściowy: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$
2. Oblicz wartości dopasowane: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i$
3. Oszacuj model pomocniczy: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \varepsilon_i$
4. Zweryfikuj hipotezę zerową: $\mathrm{H}_0: \gamma_1 = \gamma_2 = 0$
   za pomocą testu F-Walda (lub testu LM)

Uwagi:

- $\hat{y}_i^2$ i $\hat{y}_i^3$ są nieliniowymi funkcjami regresorów x_i oraz z_i, które dodajemy zamiast przekształconych regresorów aby ograniczyć spadek liczby stopni swobody
- jeśli liczba obserwacji jest wysoka, to można przeprowadzić następującą regresję pomocniczą zamiast testu RESET:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \gamma_1 x_i^2 + \gamma_2 x_i^3 + \gamma_3 z_i^2 + \gamma_4 z_i^3 + \gamma_5 x_i z_i + \gamma_6 x_i^2 z_i + \gamma_7 x_i z_i^2 + \varepsilon_i$$
$$H_0: \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7 = 0$$

Przykład 4.3.
RESET test

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry 3 modeli dla stawek godzinowych ($wage$, USD). Wyniki regresji testu RESET są następujące:

```
Pomocnicze równanie regresji dla testu specyfikacji RESET
Estymacja KMNK, wykorzystane obserwacje 1-9799
Zmienna zależna (Y): wage

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const      13,7418         2,62361            5,238    1,66e-07  ***
  educ       -0,649960       0,275975          -2,355    0,0185    **
  exper      -0,0649621      0,0245678         -2,644    0,0082    ***
  yhat^2      0,0486381      0,00563241         8,635    6,77e-018 ***
  yhat^3     -0,000536277    9,34970e-05       -5,736    1,00e-08  ***

Statystyka testu: F = 83,270688,
z wartością p = P(F(2,9794) > 83,2707) = 1,38e-036
```

Pytanie: jaka jest decyzja na podstawie przeprowadzonego testu?

Zmienna binarna / zero-jedynkowa

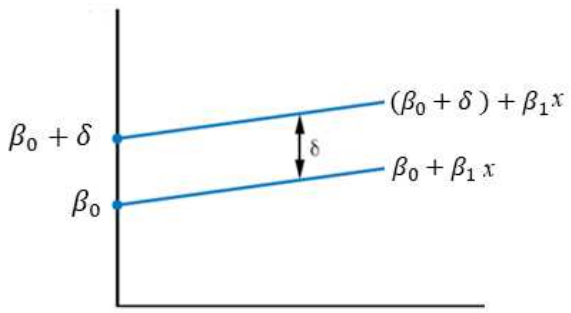
- Zmienne binarne są wprowadzane, jeśli zakładamy, że wartości stałej różnią się podzbioru obserwacji S w próbie.
- Przykłady podzbioru S: kobiety, cudzoziemcy, osoby posiadające dzieci.
- Zmienna binarna / zero-jedynkowa przyjmuje postać:

$$D_i = \begin{cases} 1 & \text{dla } i \in S \\ 0 & \text{dla } i \notin S \end{cases}$$

Dla modelu

$$y_i = \beta_0 + \beta_1 x_i + \delta D_i + \varepsilon_i,$$

parametr δ mierzy różnicę wartości w dwóch podpróbach



Zmienna binarna / zero-jedynkowa

Rozważmy model dla płac:

$$wage_i = \beta_0 + \delta\, D_i + \beta_1 educ_i + \varepsilon_i$$

$$D_i = \begin{cases} 1 & \text{dla mężczyzn} \\ 0 & \text{dla kobiet} \end{cases}$$

- δ mierzy różnicę między średnim wynagrodzeniem mężczyzn i kobiet, tj. opisuje stopień dyskryminacji na rynku pracy ze względu na płeć

$$E(wage_i) = \begin{cases} \beta_0 + \delta + \beta_1 educ_i & \text{dla mężczyzn} \\ \beta + \beta_1 educ_i & \text{dla kobiet} \end{cases}$$

- Możemy zatem sprawdzić, czy istnieją znaczące różnice w średniej płacy między mężczyznami i kobietami, których nie można wyjaśnić różnicami w poziomie wykształcenia

$$H_0: \delta = 0$$

Zmienna binarna / zero-jedynkowa

Zdefiniujmy $D_i' = 1 - D_i = \begin{cases} 0 & \text{dla mężczyzn} \\ 1 & \text{dla kobiet} \end{cases}$ oraz oszacujmy:

$$wage_i = \beta_0 + \delta' D_i' + \beta_1 educ_i + \varepsilon_i$$

uzyskamy: $\hat{\delta}' = -\hat{\delta}$.

Dlaczego? Wytłumacz.

- Czy możemy oszacować?
  i. $wage_i = \beta_0 + \delta\, D_i + \delta' D_i' + \alpha_1 educ_i + \varepsilon_i$
  ii. $wage_i = \delta\, D_i + \delta' D_i' + \alpha_1 educ_i + \varepsilon_i$
- Jakich oszacowań oczekujesz?
- Zauważ, że $D_i + D_i' = 1$ (szerzej o współliniowości w Temat 5)

Przykład 4.4.

Zmienna binarna / zero-jedynkowa

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry modelu wyjaśniającego stawki godzinowe ($wage$, USD), przez wykształcenie ($educ$, lata) oraz płeć ($female/male$). Wyniki regresji są następujące:

Model A: Zmienna zależna (Y): wage

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−9,99992	0,769702	−12,99	2,81e-038	***
educ	2,48242	0,0531762	46,68	0,0000	***
female	−4,07411	0,295410	−13,79	7,21e-043	***

Model B: Zmienna zależna (Y): wage

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−14,0740	0,800559	−17,58	3,86e-068	***
educ	2,48242	0,0531762	46,68	0,0000	***
male	4,07411	0,295410	13,79	7,21e-043	***

Pytanie: jaki jest związek między parametrami modeli A i B?

Zmienne interakcyjne

- Skomplikujmy model i wprowadźmy zmienną interakcyjną:

$$x_i^* = x_i \times D_i = \begin{cases} x_i & \text{dla } i \in S \\ 0 & \text{dla } i \notin S \end{cases}$$

Nowa postać modelu:

$$y_i = \beta_0 + \beta_1 x_i + \delta D_i + \gamma(x_i \times D_i) + \varepsilon_i$$

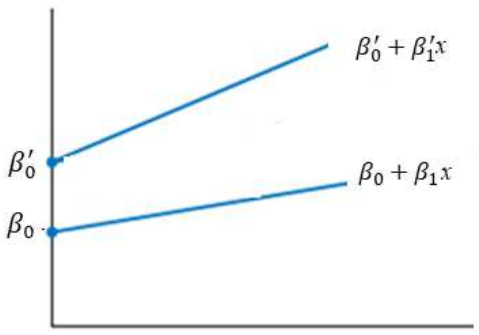
- Wartości dopasowane:

$$E(y_i) = \begin{cases} \beta_0' + \beta_1' x_i & \text{if } i \in S \\ \beta_0 + \beta_1 x_i & \text{if } i \notin S \end{cases}$$

gdzie:

$\beta_0' = \beta_0 + \delta$: przesunięcie stałej

$\beta_1' = \beta_1 + \gamma$: przesunięcie nachylenia



- Zmienne interakcyjne służą do modyfikowania parametru nachylenia, gdy zakładamy, że charakter zależności między zmiennymi różni się między podpróbkami.

Zmienne interakcyjne

Rozważmy model dla płac:

$$wage_i = \beta_0 + \beta_1 educ_i + \delta D_i + \gamma(D_i \times educ_i) + \varepsilon_i$$

gdzie $D_i = \begin{cases} 1 & \text{dla mężczyzn} \\ 0 & \text{dla kobiet} \end{cases}$

Parametr γ to różnica w średnich zwrotach z edukacji między mężczyznami i kobietami:

$$E(wage_i) = \begin{cases} \beta_0 + \delta + (\beta_1 + \gamma) educ_i & \text{dla mężczyzn} \\ \beta_0 + \beta_1 educ_i & \text{dla kobiet} \end{cases}$$

$$\frac{\partial wage}{\partial educ} = \begin{cases} \beta_1 + \gamma & \text{dla mężczyzn} \\ \beta_1 & \text{dla kobiet} \end{cases}$$

Przykład 4.5.
Zmienne interakcyjne

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry modelu wyjaśniającego stawki godzinowe ($wage$, USD) przez wykształcenie ($educ$, lata) oraz płeć ($female$).
Wyniki regresji są następujące:

Model: Zmienna zależna (Y): wage

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−9,50978	0,98150	−9,689	4,22e-022	***
educ	2,44739	0,06871	35,62	2,20e-261	***
female	−5,32158	1,57788	−3,373	0,0007	***
female_educ	0,08732	0,10849	0,8048	0,4209	

Pytanie: jak płeć wpływa na zależność między płacą i wykształceniem?

*Dodatek: Interakcje zmiennych binarnych

- Zdefiniujmy $D1_i = \begin{cases} 0 & \text{dla mężczyzn} \\ 1 & \text{dla kobiet} \end{cases}$, $D2_i = \begin{cases} 0 & \text{dla krajana} \\ 1 & \text{dla imigranta} \end{cases}$

  a zatem: $D1_i \times \mathrm{D2}_i = \begin{cases} 1 & \text{imigrantka} \\ 0 & \text{pozostali} \end{cases}$

- Dla modelu $wage_i = \beta_0 + \delta_1 D1_i + \delta_2 D2_i + \delta_{12} D1_i D2_i + \beta_1 educ_i + \varepsilon_i$

$$E(wage_i) = \begin{cases} \beta_0 + \beta_1 educ_i & \text{krajan} \\ \beta_0 + \beta_1 educ_i + \delta_2 & \text{imigrant} \\ \beta_0 + \beta_1 educ_i + \delta_1 & \text{krajanka} \\ \beta_0 + \beta_1 educ_i + \delta_1 + \delta_2 + \delta_{12} & \text{imigrantka} \end{cases}$$

Krajanie (mężczyźni) są grupą referencyjną.

- δ_1- mierzy efekt płci,
- δ_2- mierzy efekt miejsca urodzenia,
- δ_{12}- mierzy dodatkowy efekt miejsca urodzenia dla kobiet

Zadania

Zadanie 4.1

Student oszacował, że średnia liczba zadań ekonometrycznych, które może rozwiązać w ciągu godziny (efektywność, z), zależy od liczby godzin nauki w ciągu dnia (czas, h):

$$z = 0.04 + 0.32h - 0.02h^2$$

a. Jaka jest średnia efektywność w rozwiązywaniu ćwiczeń, jeśli czas wynosi $h = 4$?
b. Jaki jest krańcowy wpływ godzin na efektywność, $ME_{z/h}$, jeśli $h = 4$?
c. Jaka jest elastyczność efektywności względem czasu, $E_{z/h}$, przy $h = 4$?
d. Przy jakim czasie nauki efektywność jest najwyższa?
e. Przelicz punkty a-d dla $h = 1$ i 5.

Zadanie 4.2

Popyt na pączki (Y) w zależności od ceny (P) wynosi:

$$Y = 3 + 6/P$$

a. Oblicz poziom sprzedaży zakładając, że P = 2.
b. Jaki jest krańcowy wpływ ceny na sprzedaż, $ME_{Y/P}$, dla $P = 2$.
c. Oblicz hipotetyczny poziom sprzedaży dla $P = 3$ wykorzystując informacje z punktów a i b.
d. Oblicz teoretyczny poziom sprzedaży z modelu dla $P = 3$. Porównaj wyniki z punktem c.
e. Oblicz elastyczność sprzedaży względem ceny dla $P = 2$ oraz $P = 3$.

Zadanie 4.3

Oszacowano model, w którym wyniki egzaminu maturalnego z matematyki (SCORE) są wyjaśnione przez:

FATHER_EDU:	wykształcenie ojca (1 – wyższe, 0 - inne)
LN_INCOME:	logarytm dochodu *per capita* gospodarstwa domowego
GENDER:	płeć (1 – mężczyzna, 0 - kobieta)
PRIVATE_SCHOOL:	rodzaj szkoły (1 – prywatna, 0 -publiczna)
PRIVATE_GENDER:	iloczyn GENDER i PRIVATE_SCHOOL
TUTORING:	liczba godzin korepetycji przed egzaminem

```
------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  father_edu |    130.414   35.037       3.72   0.001      59.654     201.173
   ln_income |    125.156   34.027       3.68   0.001      56.436     193.877
      gender |   -112.086   43.445      -2.58   0.014    -199.826     -24.346
private_school |    -40.075   48.370      -0.83   0.412    -137.762      57.611
private_gender |    292.293   83.103       3.52   0.001     124.462     460.125
     tutoring |     39.137     .599      65.33   0.000      37.927      40.347
  tutoring_2 |      -.015     .001      -8.39   0.000       -.018       -.011
       _cons |    338.794  208.830       1.62   0.112     -82.948     760.536
------------------------------------------------------------------------------
```

```
-------------------------
       Gender |       Mean           |
          0   |   2437.777
          1   |   2440.632
-------------------------
```

Zadanie 4.3 cd.

a. Zinterpretuj oszacowanie parametru przy FATHER_EDU.

b. Zinterpretuj oszacowanie parametru przy LN_INCOME.

c. Czy istnieją jakieś różnice w wynikach egzaminu między chłopcami i dziewczętami?

d. Na podstawie uzyskanych wyników doradź rodzicom, czy powinni wysłać swoje dziecko do prywatnej szkoły. Czy rekomendacja zależy od płci dziecka?

e. Czy rodzice powinni zapewnić dziecku możliwie jak najwięcej godzin korepetycji przed egzaminem?

f. Spróbuj naszkicować szacunkową zależność między liczbą godzin korepetycji a wynikiem uzyskanym na egzaminie.

Zadanie 4.4

W pliku `utown.gdt` znajdują się dane dotyczące cen nieruchomości ($price$, `1000$`), ich powierzchni ($sqft$, `100sq.feet`) oraz lokalizacji w pobliżu uniwersytetu ($utown = 1$)

a. Oszacuj następujące modele i zinterpretuj ich parametry

M1: $price_i = \beta_0 + \beta_1 sqft_i + \varepsilon_i$
M2: $price_i = \beta_0 + \beta_1 sqft_i + \delta \times utown_i + \varepsilon_i$
M3: $price_i = \beta_0 + \beta_1 sqft_i + \delta \times utown_i + \gamma \times (sqft_i \times utown_i) + \varepsilon_i$

b. Jaka jest interpretacja parametrów modelu z logarytmami?

M4: $\ln(price_i) = \beta_0 + \beta_1 \ln(sqft_i) + \varepsilon_i$
M5: $\ln(price_i) = \beta_0 + \beta_1 \ln(sqft_i) + \delta \times utown_i + \varepsilon_i$

c. Czy sensowne jest uwzględnianie zmiennych interakcyjnych w modelu z logarytmami?

M6: $\ln(price_i) = \beta_0 + \beta_1 \ln(sqft_i) + \delta \times utown_i + \gamma \times (sqft_i \times utown_i) + \varepsilon_i$

d. Spróbuj włączyć inne zmienne binarne do specyfikacji modelu.

Zadanie 4.5

Agencja nieruchomości „Na swoim” wprowadza usługę doradczą, pozwalającą ocenić atrakcyjność ofert sprzedaży. Otrzymałeś zadanie zbudowania modelu wyceny nieruchomości, który pozwoli ustalić, czy oferta jest atrakcyjna, a tym samym pomóc klientom w podjęciu decyzji o zakupie. Baza danych, na podstawie której ma być oszacowany model (`housing_market.gdt`), zawiera następujące zmienne:

- cena
- powierzchnia (w metrach kwadratowych)
- piętro (0 oznacza parter, 1 - pierwsze piętro itp.)
- okres budowy (0: przedwojenny; 1: 40s-50s; 2: 60s-80s; 3: 90s; 4: po 2000 r.)
- budynek położony w centrum miasta (1 - tak, 0 - nie)
- w budynku jest winda (1 - tak, 0 - nie)

a. Zbuduj model regresji liniowej dla cen mieszkań. Jakie wnioski płyną z uzyskanych oszacowań.
b. Kierownictwo nie jest zadowolone z twojej pracy: uważa się, że dopasowanie jest zbyt niskie, aby zastosować model w praktyce. Postanawiasz zmienić specyfikację modelu. Obserwując rynek nieruchomości, zauważasz:
   - małe mieszkania wydają się droższe za metr kwadratowy niż mieszkania większe,
   - ludzie nie lubią mieszkać na parterze,
   - mieszkania w centrum miasta wydają się stosunkowo drogie,
   - cena za metr kwadratowy wydaje zależeć od okresu budowy: najtańsze mieszkania są na dużych osiedlach wybudowanych w latach 60. i 80. („wielka płyta”), a najdroższe te z ostatnich lat lub wybudowanych przed wojną

   Zbuduj model, którego specyfikacja uwzględni twoje spostrzeżenia.
c. Czy tym razem kierownictwo będzie zadowolone z pracy?

Zadanie 4.6

Na podstawie pliku `wage2.gdt` ustal, w jaki sposób płace zależą od wykształcenia, płci i narodowości. Użyj zmiennych interakcyjnych, specyfikacji nieliniowych oraz logarytmicznych.

Jaka specyfikacja modelu jest według Ciebie najlepsza?

Definicja zmiennych jest następująca:

`wage`	wynagrodzenie miesięczne wynagrodzenie w PLN
`education`	wykształcenie w latach
`gender`	0 dla mężczyzn, 1 dla kobiet
`nationality`	0 dla Polaków, 1 dla imigrantów

Zadanie 4.7

Plik `cps5.gdt` zawiera dane o stawce godzinowej ($wage$, USD), wykształceniu ($educ$, lata) oraz wieku (age, lata).

a. Zbuduj model, w którym $wage$ zależy od dwóch pozostałych dwóch zmiennych. Porównaj specyfikację liniową względem specyfikacji wielomianowej 2 stopnia. Przeprowadź test RESET.
b. Powtórz czynności z punktu a. dla modelu, w którym zmienną objaśnianą jest $\ln(wage)$
c. Czy zależność między wykształceniem a wynagrodzeniem zależy od zmiennej $female$? Wprowadź zmienne interakcyjne
d. Czy zależność między wykształceniem a wynagrodzeniem zależy od zmiennych $asian, black, white$? Wprowadź zmienne interakcyjne
e. Zbuduj model, który twoim zdaniem najlepiej opisuje zróżnicowanie stawek godzinowych.

Temat 5

Weryfikacja modelu: współliniowość i normalność składnika losowego

Michał Gradzewicz i Michał Rubaszek

- Dokładna współliniowość
- Zmienne kategorialne a współliniowość
- Przybliżona współliniowość
- Czynnik Inflacji Wariancji (CIW)
- Modele z logarytmami zmiennych
- Momenty rozkładu
- Test Jarque-Bera

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Współliniowość

Etapy weryfikacji modelu

1. Oceny parametrów i ich znaki
2. Istotność parametrów
3. Dopasowanie modelu do danych
4. **Specyfikacja modelu / postać funkcyjna**
5. Własności składnika losowego
6. Stabilność parametrów

3

Przypomnienie: Założenia KMNK

Założenia KMNK (przypomnienie z Temat 2)

A1. Prawdziwy model jest następujący: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

A2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ oraz $E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$

A3. $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

A4. $\mathbf{X}$ jest nielosową macierzą, której rząd wynosi $rank(X) = (K+1) < N$

A5. $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$

W Temat 5 skupimy się na **A4** i **A5**.

- **A4** jest niezbędne do uzyskania oszacowań MNK, zaś jego niespełnienie określamy jako **współliniowość**
- Założenie **A5** o **normalności rozkładu** składnika losowego , chociaż niepotrzebne dla twierdzenia Gaussa-Markowa, jest niezbędne aby testy miały odpowiednie rozkłady

Współliniowość zmiennych objaśniających

- Co oznacza założenie **A4**: $rank(\mathbf{X}) = K + 1$?
- Zilustrujmy to na przykładzie, gdy $n = 5$ i $K = 2$:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

- Jeśli jedna z kolumn $\mathbf{X}$ jest liniową kombinacją innych kolumn (tutaj $1 = x_1 + x_2$)
  - występuje **dokładna współliniowość**
  - W konsekwencji:
    - macierz $\mathbf{X}'\mathbf{X}$ jest osobliwa, tj. $\det(\mathbf{X}'\mathbf{X}) = 0$
    - jej odwrotność nie istnieje
    - a zatem nie można policzyć oszacowań:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Przykład 5.1. Dokładna współliniowość

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry modelu wyjaśniającego stawki godzinowe ($wage$, USD), przez wykształcenie ($educ$, lata) oraz płeć ($female/male$). Wyniki regresji są następujące:

Model A: Zmienna zależna (Y): wage

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−9,99992	0,769702	−12,99	2,81e-038	***
educ	2,48242	0,0531762	46,68	0,0000	***
female	−4,07411	0,295410	−13,79	7,21e-043	***

Model B: Zmienna zależna (Y): wage

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−14,0740	0,800559	−17,58	3,86e-068	***
educ	2,48242	0,0531762	46,68	0,0000	***
male	4,07411	0,295410	13,79	7,21e-043	***

Pytanie: Dlaczego nie można oszacować parametrów modelu?

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 female_i + \beta_3 male_i + \varepsilon_i$$

Wspóliniowość a zmienne sezonowe / kategorialne

- Sezonowość w modelu ekonometrycznym jest często uwzględniana w modelu ekonometrycznym za pomocą **sezonowych zmiennych binarnych**. Przykładowo, dla danych kwartalnych:

$$q_{1t}=\begin{cases}1 & \text{dla } Q1\\ 0 & \text{pozostałe}\end{cases},\ q_{2t}=\begin{cases}1 & \text{dla } Q2\\ 0 & \text{pozostałe}\end{cases},\ q_{3t}=\begin{cases}1 & \text{dla } Q3\\ 0 & \text{pozostałe}\end{cases},\ q_{4t}=\begin{cases}1 & \text{dla } Q4\\ 0 & \text{pozostałe}\end{cases}$$

- Należy zauważyć, że wszystkich zmiennych nie możemy wprowadzić do modelu, ponieważ $1 = q_{1t} + q_{2t} + q_{3t} + q_{4t}$. Dlatego musimy wybrać tzw. "kwartał odniesienia", np. jeżeli jest to $Q1$ to model jest postaci:

$$y_t = \beta_0 + \beta_1 x_t + \gamma_2 q_{2t} + \gamma_3 q_{3t} + \gamma_4 q_{4t} + \varepsilon_t$$

Interpretacja γ_s - różnica między średnim poziomem zmiennej zależnej w QS oraz $Q1$

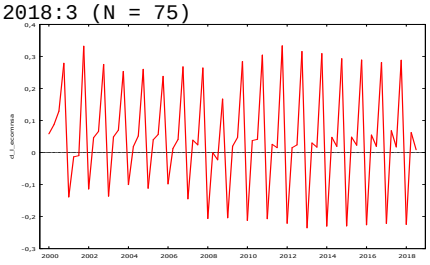
- Podobnie postępujemy dla **zmiennych kategorialnych**, tj. przyjmujących S różnych wartości (np. województwo, pochodzenie). Zamieniamy je na S zmiennych binarnych i wybieramy obiekt odniesienia, którego nie uwzględniamy w modelu by zapobiec współliniowości.

Przykład 5.2. Zmienne sezonowe

Wykorzystując dane do sprzedaży e-commerce oszacowano parametry modelu wyjaśniającego dynamikę sprzedaży, `d_log_sales`, przez zmienne kwartalne. Wyniki oszacowań są następujące:

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 2000:1-2018:3 (N = 75)
Zmienna zależna (Y): d_l_sales

            współ.   błąd std.  t-Stud.  wartość p
  ------------------------------------------------------
  const      0,279   0,0109     25,41    2,26e-037 ***
  q1        −0,448   0,0153    −29,19    2,77e-041 ***
  q2        −0,243   0,0153    −15,86    4,91e-025 ***
  q3        −0,246   0,0153    −16,04    2,65e-025 ***
```



Pytania:

1. Jakie jest średnie tempo wzrostu sprzedaży e-commerce w pierwszym kwartale?
2. Jakie jest średnie tempo wzrostu sprzedaży e-commerce w czwartym kwartale?
3. Jakie byłyby parametry modelu bez zmiennej q1 (a ze zmienną q4)?

Współliniowość zmiennych objaśniających

- W praktyce modelowania problem **dokładnej współliniowości** występuje rzadko. Częstszym problemem jest **przybliżona współliniowość**, czyli sytuacja, kiedy korelacja między parą zmiennych jest bliska 1

- Przykłady **przybliżonej współliniowości**.
  - **Szeregi czasowe**. Zmienne makroekonomiczne (np. PKB, inwestycje, import, konsumpcja) charakteryzuje współzmienność, ponieważ ich zmiany w czasie są skorelowane ze względu na oddziaływanie cyklu koniunkturalnego
  - **Dane przekrojowe**. Obiekty często charakteryzuje tendencja do proporcjonalnych zmian wartości zmiennych objaśniających. W szczególności, obiekty duże charakteryzują się często wysokimi wartościami różnych zmiennych je określających, a obiekty małe - małymi

Konsekwencje przybliżonej współliniowości

- Z Tematu 2 wiemy, że wariancja estymatora MNK wynosi:

$$\boldsymbol{\Sigma}_{\widehat{\beta}} = Var\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Dla dokładnej współliniowości $\det(\mathbf{X}'\mathbf{X}) = 0$, a zatem $\boldsymbol{\Sigma}_{\widehat{\beta}}$ jest nieskończona. Oznacza to, że precyzja szacunku zerowa!
- W przypadku **przybliżonej współliniowości** $\det(\mathbf{X}'\mathbf{X})$ jest bliskie zera, zaś elementy macierzy $\boldsymbol{\Sigma}_{\widehat{\beta}}$ są względnie duże, czyli:
  - błędy szacunku $S_{\widehat{\beta_k}}$ są wysokie, zaś precyzja oszacowań parametrów β_k jest niska
  - testy dla pojedynczych parametrów wskazują na nieistotność zmiennych
  - ale współczynnik determinacji R^2 może być wysoki

Ważne. W przypadku współliniowości dopasowanie modelu do danych jest podobne przy różnych kombinacjach parametrów. Pomyśl np. o modelu, w którym cena mieszkania zależy od powierzchni oraz liczby pokoi.

Konsekwencje przybliżonej współliniowości

Dlaczego przy nieistotnych zmiennych współczynnik determinacji R^2 może być wysoki?

- Można pokazać, że współczynnik determinacji $R^2 = Q_0' Q^{-1} Q_0$,
  - gdzie Q to macierz $K \times K$, której element (k, l) wynosi $q_{kl} = cor(x_k, x_l)$
  - Q_0 jest wektorem $K \times 1$, którego k-ty element wynosi $q_{0k} = cor(y, x_k)$
  - jeśli Q ma elementy pozadiagonalne zbliżone do 1, zwiększa to wartości Q^{-1} i automatycznie wartość współczynnika R^2

Problemy związane ze współliniowością przybliżoną:

- **Oszacowania parametrów są niestabilne w próbie**.
  Innymi słowy, niewielkie zmiany próby (np. obcięcie jej o kilka obserwacji) prowadzą do wyraźnych zmian oszacowań parametrów
- **Problematyczna staje się też interpretacja parametrów modelu.**
  Jeśli zmiana x_k pociąga za sobą niemal automatyczne dostosowanie innych zmiennych objaśniających, to trudno koncepcyjnie interpretować parametr β_k w kategoriach *ceteris paribus,* czyli jako samodzielny efekt x_k na y

Przykład 5.3. Ilustracja problemu współliniowości

Rozpatrzmy 500 obserwacji wylosowanych z:

$x_{1i} \sim N(0, 3^2)$

$x_{2i} = 10x_{1i} + \epsilon_i, \quad \epsilon_i \sim N(0, 0.01^2)$

$y_i = 2 + 5x_{1i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 0.5^2)$

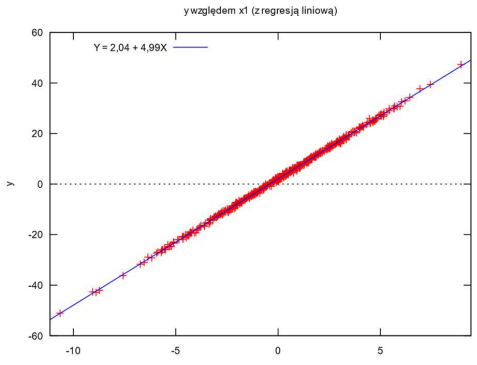
oraz oszacowania modeli:

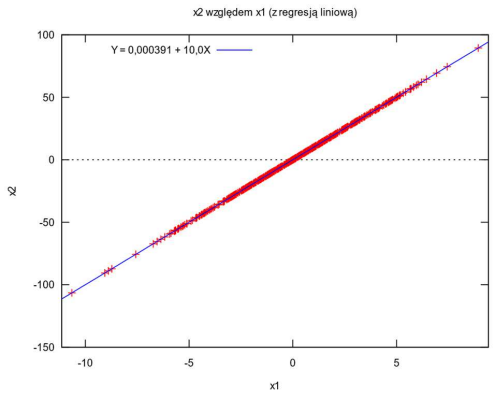
$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

Program Gretla:

```
nulldata 500
series x1 = randgen(N, 0, 3)
series eps1 = randgen(N, 0, 0.5)
series y = 2 + 5*x1 + eps1
ols y 0 x1
series eps2 = randgen(N, 0, 0.01)
series x2 = 10*x1 + eps2
ols y 0 x1 x2
gnuplot y x1 --output=graph1.svg
gnuplot x2 x1 --output=graph2.svg
```





Przykład 5.3. Ilustracja problemu współliniowości

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 1-500
Zmienna zależna (Y): y

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       2,00919        0,0235079         85,47    1,01e-299 ***
  x1          4,99994        0,00808702       618,3     0,0000    ***

Średn.aryt.zm.zależnej  1,909547   Odch.stand.zm.zależnej  14,55785
Suma kwadratów reszt    137,5962   Błąd standardowy reszt  0,525640
Wsp. determ. R-kwadrat  0,998699   Skorygowany R-kwadrat   0,998696
F(1, 498)               382254,2   Wartość p dla testu F   0,000000
Logarytm wiarygodności -386,8980   Kryt. inform. Akaike'a  777,7961
Kryt. bayes. Schwarza   786,2253   Kryt. Hannana-Quinna    781,1037
```

```
Model 2: Estymacja KMNK, wykorzystane obserwacje 1-500
Zmienna zależna (Y): y

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       2,00917         0,0235310      85,38      3,99e-299 ***
  x1          9,25695        23,7452          0,3898    0,6968
  x2         -0,425712        2,37457        -0,1793    0,8578

Średn.aryt.zm.zależnej  1,909547   Odch.stand.zm.zależnej  14,55785
Suma kwadratów reszt    137,5873   Błąd standardowy reszt  0,526152
Wsp. determ. R-kwadrat  0,998699   Skorygowany R-kwadrat   0,998694
F(2, 497)               190755,7   Wartość p dla testu F   0,000000
Logarytm wiarygodności -386,8819   Kryt. inform. Akaike'a  779,7637
Kryt. bayes. Schwarza   792,4076   Kryt. Hannana-Quinna    784,7251
```

Wykrywanie współliniowości

- **Podstawowe wskazanie na problem współliniowości:**
  zmienne w modelu są nieistotne, zaś dopasowanie mierzone współczynnikiem R^2 wysokie

- Formalna diagnoza jest następująca:
  - dla każdego regresora x_k obliczamy **czynnik inflacji wariancji** CIW
    (ang. *Variance Inflation Factor – VIF*)

$$CIW_k = \frac{1}{1-R_k^2}, \quad k = 1,2,\dots,K$$

    gdzie R_k^2 jest współczynnikiem determinacji R^2 modelu, w którym x_k jest objaśniane przez pozostałe $K-1$ zmienne objaśniające
  - Wartości $CIW_k > 10$ są oznaką przybliżone współliniowości

- Przykładowo, dla modelu

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$$

wartość R_2^2 to R^2 z modelu: $x_{2t} = \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{3t} + \eta_t$, zaś $CIW_2 = \frac{1}{1-R_2^2}$

Postępowanie w przypadku współliniowości

W zasadzie istnieją trzy metody postępowania w przypadku występowania współliniowości:

1. **Zmiana specyfikacji modelu:**

- Eliminacja zmiennych powodujących występowanie współliniowości (wysokie wartości CIW)
  - ale usunięcie istotnych zmiennych ze specyfikacji modelu może prowadzić do problemu pominiętych zmiennych (ang. o*mitted variable, zob.* Tematy 12-14)
- Transformacja zmiennych (np. zamiast dochodów miesięcznych - stawka godzinowa)
- Zastosowanie metod redukcji wymiaru, np. **metody głównych składowych**, polegającej na zamianie K skorelowanych zmiennych objaśniających na $M \leq K$ niezależnych czynników**.**

2. **Zmiana metody estymacji**

- Przykładem innej metody estymacji jest **regresja grzbietowa** (*ang. ridge regression)*

$$\widehat{\boldsymbol{\beta}}^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

gdzie $\lambda > 0$ jest skalarem, a $\boldsymbol{I}$ macierzą jednostkową. Metodą ta uzyskujemy obciążony, ale jednocześnie bardziej efektywny estymator niż estymator MNK.

- Inne przykłady to metoda LASSO czy estymacja Bayesowska.

3. **Nic nie robić, co jest uzasadnione gdy rozwiązanie stworzy jeszcze gorsze problemy**

Normalność rozkładu składnika losowego ε_i

Etapy weryfikacji modelu

1. Oceny parametrów i ich znaki
2. Istotność parametrów
3. Dopasowanie modelu do danych
4. Specyfikacja modelu / postać funkcyjna
5. **Własności składnika losowego**
6. Stabilność parametrów

17

Normalność rozkładu składnika losowego

- Normalność rozkładu składnika losowego nie jest niezbędna do wyprowadzenia własności estymatora MNK zawartych w twierdzenie Gaussa-Markowa, ale…
  - jest wymagana, aby statystyki testów miały odpowiednie rozkłady (np. t, χ^2, F)
  - czyli abyśmy mogli korzystać ze standardowo liczonych wartości-p

- Istnieje cała grupa testów normalności zmiennej (zob. http://smarterpoland.pl/index.php/2013/04/wybrane-testy-normalnoci), z których omówimy test Jarque-Bera, oparty o wystandaryzowaną kurtozę i skośność w próbie:

- Powtórzenie ze statystyki: **momenty centralne w próbie** dla zmienne x:

Wariancja: $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\sigma}^2$

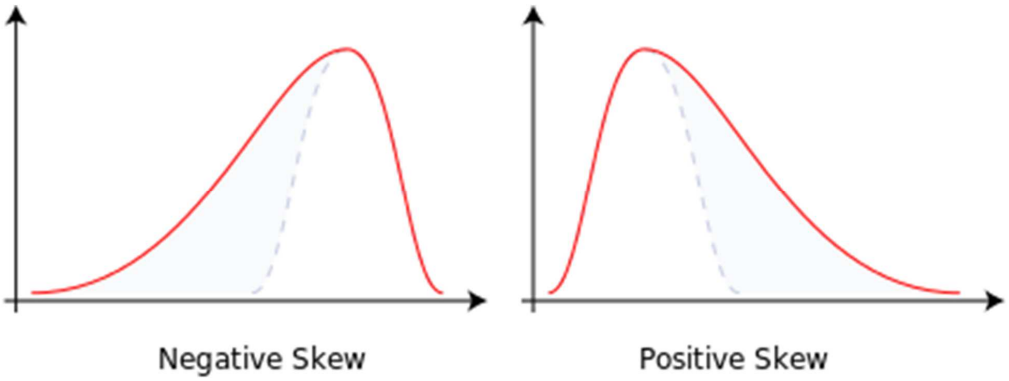
Skośność: $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$

Kurtoza: $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$

Skośność rozkładu

Skośność jest miarą asymetrii rozkładu, czyli sytuacji, gdy masa prawdopodobieństwa rozkładu przesunięta jest na prawo lub lewo od wartości mediana.

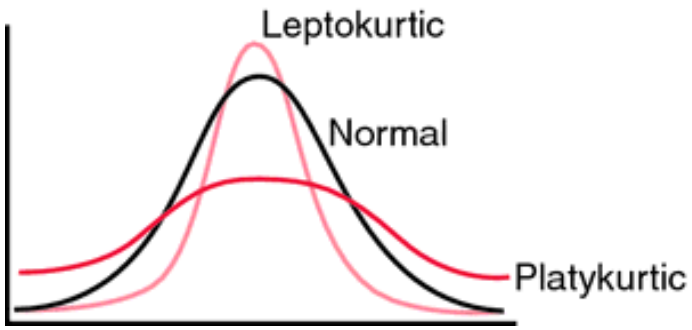
- Rozkład normalny jest symetryczny (czyli skośność jest zerowa)
- Skośność może być ujemna (lewostronna, gdy większość wyników powyżej średniej) lub dodatnia (prawostronna, gdy większość wyników jest poniżej średniej)



Kurtoza rozkładu

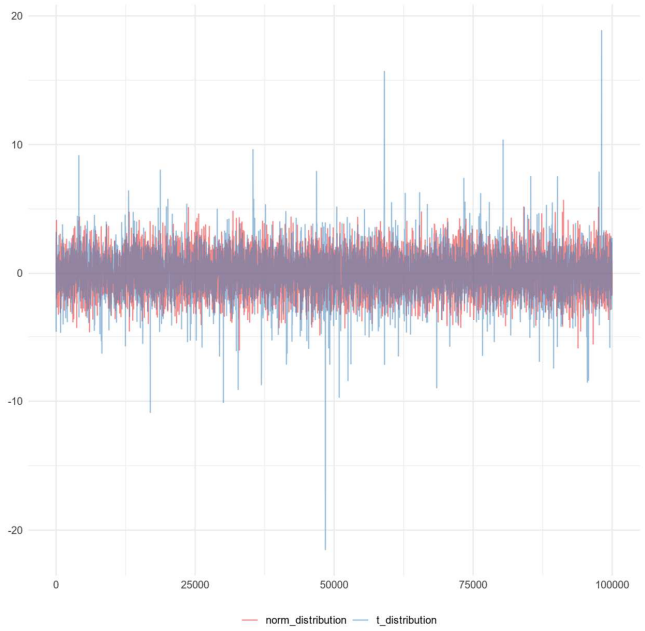
Kurtoza dotyczy ogonów rozkładu, czyli prawdopodobieństwa zdarzeń mocno odbiegających od przeciętnych wartości danego zjawiska

- Dla rozkładu $N(0,1)$ kurtoza w populacji wynosi 3
- Dla rozkładu **leptokurtcznego** (wystandaryzowana) kurtoza jest większa niż 3 (prawdopodobieństwo zdarzeń nietypowych jest wyższe, niż w przypadku rozkładu normalnego, rozkłady takie rozpatruje się w ekonometrii rynków finansowych)
- Dla rozkładu **platykurtycznego** (wystandaryzowana) kurtoza jest mniejsza niż 3 (prawd. zdarzeń nietypowych jest niższe, niż w przypadku rozkładu normalnego)

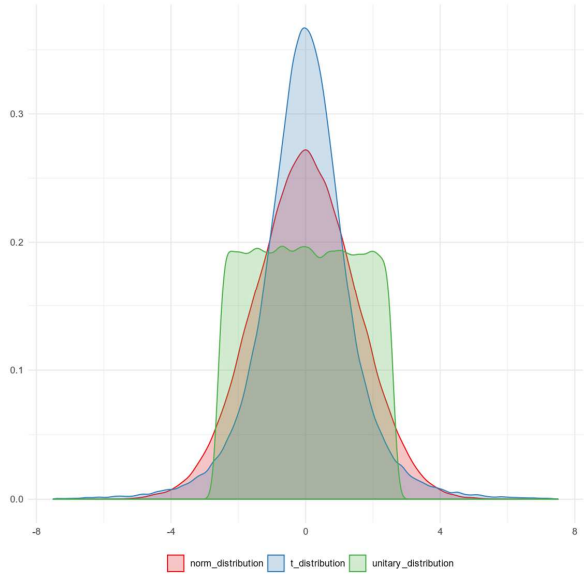


Ilustracja kurtozy

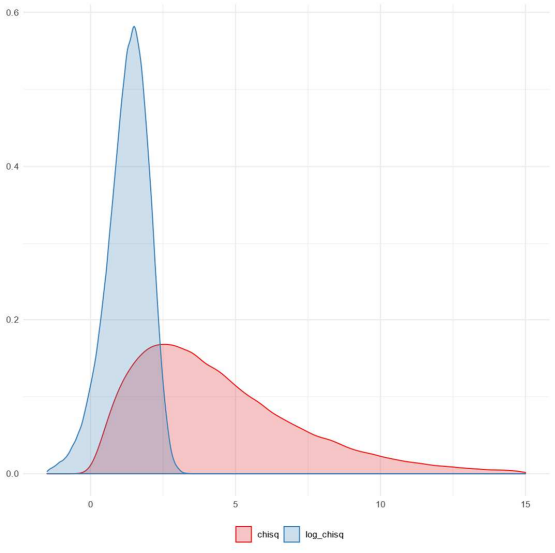
Obserwacje wylosowane z rozkładu normalnego i rozkładu t —Studenta o tej samej wariancji



Przykład 5.4. Rozkłady o różnej kurtozie i skośności



type	mean	sd	skewness	kurtosis
norm_distribution	0	1.5	0	3
t_distribution	0	1.49	-0.23	15.6
unitary_distribution	0	1.5	0	1.8



type	mean	sd	skewness	kurtosis
chisq	4.5	3	1.34	5.7
log_chisq	1.27	0.74	-0.71	3.96

W przypadku skośności prawostronnej logarytm zmiennej często ma bardziej symetryczny rozkład

Test Jarque-Bery

- Zestaw hipotez testowych

  H_0: składnik losowy ε ma rozkład normalny

  H_1: składnik losowy ε nie ma rozkładu normalnego

- Wystandaryzowany współczynnik skośności:

$$S = \frac{m_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}^2\right)^{\frac{3}{2}}}$$

- Wystandaryzowany współczynnik kurtozy:

$$K = \frac{m_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}^2\right)^{\frac{4}{2}}}$$

- Hipoteza zerowa testu Jarque-Berry (JB) opiera się na łącznym teście:

$$H_0\colon S = 0 \ \wedge K = 3$$

Test Jarque-Bery

- Statystyka testu JB:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right) \sim \chi^2(2)$$

  przy prawdziwości H_0 ma rozkład χ^2 z dwoma stopniami swobody (ponieważ testujemy restrykcje na 2 parametry rozkładu)

- Jeśli statystyka JB jest większa od wartości krytycznej (lub wartość-*p* jest niższa niż poziom istotności) to odrzucamy H_0

- Odrzucenie hipotezy zerowej
  - Nie możemy korzystać ze standardowych testów statystycznych
  - W przypadku dużych prób możemy jednak liczyć na własności asymptotyczne testów

Przykład 5.5. Test Jarque-Bera

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry modelu wyjaśniającego stawki godzinowe ($wage$, USD), przez wykształcenie ($educ$, `lata`) oraz płeć ($female$). Wyniki regresji są następujące:

```
Model: Zmienna zależna (Y): wage
            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       −9,99992       0,769702         −12,99    2,81e-038 ***
  educ         2,48242       0,0531762         46,68    0,0000    ***
  female      −4,07411       0,295410         −13,79    7,21e-043 ***

Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny.
Test JB: Chi-kwadrat(2) = 9680,694 z wartością p 0,00000
```

Pytanie: Jakie wnioski na temat rozkładu składnika losowego?

Zadania

Zadanie 5.1

Kwartalna sprzedaż lodów (Y) wynosi:

$$\widehat{Y_t} = 10.0 + 2.0Q_{2t} + 4.0Q_{3t} - 1.5Q_{4t}$$

gdzie Q_{it} jest zmienną binarną, która przyjmuje wartość 1 dla kwartału i.

a. Oblicz teoretyczny poziom sprzedaży w pierwszym kwartale.
b. Dokonaj interpretacji parametru znajdującego się przy zmiennej Q_{2t}.
c. Wyjaśnij, dlaczego nie możemy dodać zmiennej Q_{1t} do specyfikacji modelu.
d. Jakie byłyby oszacowania parametrów, gdybyśmy zastąpili Q_{4t} przez Q_{1t} w zbiorze regresorów
e. A jakie, gdyby to było Q_{2t} lub Q_{3t}?
f. Czy można oszacować model:

$$Y_t = \beta_1 Q_{1t} + \beta_2 Q_{2t} + \beta_3 Q_{3t} + \beta_4 Q_{4t} + \varepsilon_t$$

a. Czy wiesz, jakie byłyby w przybliżeniu oszacowania regresji, w której zmienną objaśnianą jest $\ln Y$ a nie Y, tj.:

$$\ln Y_t = \beta_0 + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} - \gamma_4 Q_{4t} + \varepsilon_t$$

Zadanie 5.2

Plik `cps5.gdt` zawiera dane o stawce godzinowej, wykształceniu i innych zmiennych zebranych w Current Population Survey (CPS) z 2008 roku

a. Skonstruuj histogram zmiennej $wage$ i jej logarytmu. Która wydaje się bliższa rozkładowi normalnemu?
b. Ile wynosi (wystandaryzowana) skośność i kurtoza obu zmiennych? Przeprowadź test JB.
c. Oszacuj parametry modeli:

Model A: $wage_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$
Model B: $\ln(wage_i) = \beta_0 + \beta_1 educ_i + \varepsilon_i$

d. Skonstruuj histogram reszt obu modeli, oblicz skośność i kurtozę oraz przeprowadź test JB.
e. Czy w obu modelach można korzystać z podawanych przez Gretl wartości krytycznych dla różnych testów, np. dla testu t?
f. Oszacuj model dla logarytmu płac uwzględniając zmienne $educ$, $exper$ i age.
Jakie jest oszacowanie parametru przy zmiennej age? Czy potrafisz wyjaśnić, co się stało?

Zadanie 5.3

Przeprowadź następującą symulację w programie Gretl:

Zapisz na kartce parametry (DGP=data generating proces).

$N = 50$:	liczba obserwacji
$\alpha_1 = 10, \sigma_1 = 10$	DGP dla x_1
$\alpha_2 = 10, \sigma_2 = 10$	DGP dla x_2
$\alpha_3 = 10, \sigma_3 = 1$	DGP dla y

a. Stwórz pusty zbiór danych z zakresem N obserwacji i wygeneruj zmienne:

$x_1 = \alpha_1 + \varepsilon_1$, gdzie $\varepsilon_1 \sim N(0, \sigma_1^2)$

$x_2 = 1 + \alpha_2 x_1 + \varepsilon_2$, gdzie $\varepsilon_2 \sim N(0, \sigma_2^2)$

$y = 1 + \alpha_3 x_1 + x_2 + \varepsilon_3$, gdzie $\varepsilon_3 \sim N(0, \sigma_3^2)$.

b. Policz współczynnik korelacji między x_1 a x_2.

c. Oszacuj parametry regresji:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

d. Jak kształtują się wartości R^2 oraz istotności zmiennych modelu?

e. Co wynika z analizy CIW?

f. Narysuj histogram reszt. Przeprowadź test JB dla reszt modelu

g. Pozmieniaj wartości parametrów i powtórz czynności z punktów **a.-f.**

Zadanie 5.4

Zbiór danych *rice5.gdt* zawiera zmienne opisujące produkcję ryżu ($prod$, tony) , w zależności od powierzchni ($area$, hektary), ilości zużytego nawozu ($fert$, kg) oraz nakładów pracy ($labor$, dni robocze)

a. Oblicz jak kształtują się korelacje między zmiennymi: *area, fert, labor, prod*? A jak między ich logarytmami?

b. Oszacuj model opisujący funkcję produkcji dla obserwacji z 1994 roku (ustaw zakres próby):

$$\ln prod = \beta_0 + \beta_1 \ln area + \beta_2 \ln labor + \beta_3 \ln fert + \varepsilon_t.$$

- Jaka to funkcja produkcji?
- Które zmienne są istotne? Przeprowadź test łącznej istotności modelu.
- Dokonaj analizy współliniowości w modelu przy użyciu czynników inflacji wariancji (CIW)
- Wytłumacz, dlaczego R^2 modelu jest wysoki, a zmienne są nieistotne?
- Sprawdź hipotezę zerową wskazującą na stałe korzyści skali

d. Oszacuj ten sam model na próbie dla 1993 roku

- Jak wyglądają wyniki analizy CIW?
- Jak kształtuje się istotność zmiennych i dopasowanie modelu do danych

Zadanie 5.5

W pliku `TaylorRule.gdt` zawarte są dane o poziomie stopy procentowej (*IR*, w %), inflacji rocznej (*INF*, %) oraz indeksu aktywności gospodarczej (*Y*, 100 jeżeli normalny poziom aktywności) dla wybranych krajów OECD.

a. Wybierz kraj EŚW ($c \in \{POL, HUN, CZE\}$), który będziesz analizował
b. Oszacuj parametry modelu, w który poziom stopy jest objaśniany przez główne stopy światowe:

$$IR_t^c = \beta_0 + \beta_1 IR_t^{EA} + \beta_2 IR_t^{USA} + \beta_3 IR_t^{JPN} + \beta_4 IR_t^{GBR} + \beta_5 IR_t^{CHE} + \varepsilon_t$$

c. Spójrz na macierz korelacji między zmiennymi występującymi w modelu oraz oszacowania modelu. Dlaczego znaki korelacji $cor(y, x_k)$ i oszacowań $\widehat{\beta_k}$ w wielu przypadkach są inne?
d. Oblicz i zinterpretuj wskaźniki CIW.
e. Zaproponuj model, który rozwiązuje problem współliniowości.
f. Powtórz punkty **b.-e.** dla zmiennych INF oraz Y.

Temat 6

Weryfikacja modelu: heteroskedastyczność

Zuzanna Wośko i Karol Szafranek

- Definicja heteroskedastyczności
- Konsekwencje heteroskedastyczności dla estymatora MNK
- Testy Breuscha-Pagana, White'a oraz Goldfelda-Quandta
- Ważona / uogólniona MNK
- Błędy odporne na heteroskedastyczność

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej i zbioru zmiennych objaśniających
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Etapy weryfikacji modelu

1. Oceny parametrów i ich znaki
2. Istotność parametrów
3. Dopasowanie modelu do danych
4. Specyfikacja modelu / postać funkcyjna
5. **Własności składnika losowego**
6. Stabilność parametrów

Definicja heteroskedastyczności

Definicja heteroskedastyczności

Heteroskedastyczność, określane również jako **niejednorodność wariancji składnika losowego**, dotyczy następującego założenia MNK:

A3. $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$

Macierz kowariancji składnika losowego

Homoskedastyczny składnik losowy

$$Var(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Heteroskedastyczny składnik losowy

$$Var(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Heteroskedastyczność:
konsekwencje dla estymatora MNK

1. Estymator MNK jest nadal nieobciążony, ale przestaje być najbardziej efektywnym estymatorem liniowym: **istnieje inny estymator liniowy o mniejszej wariancji**.

2. Wzór z Tematu 2 na wariancję estymatora MNK :

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] \overset{A3}{=} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

   przestaje być poprawny, czyli błędy szacunku liczone w standardowy sposób są również niepoprawne.

3. Punkt 2 implikuje, że statystyki testów istotności nie mają rozkładu t-Studenta. Podobnie statystyki testów istotności łącznej nie pochodzą z rozkładu *F* lub χ^2.

Wniosek: występowanie heteroskedastyczności utrudnia weryfikację modelu.

Heteroskedastyczność:
konsekwencje dla estymatora MNK

- Dlaczego Wzór z Tematu 2 na wariancję estymatora MNK jest niepoprawny:

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] \overset{A3}{=} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Zauważmy, że (por. Temat 2):

$$Var(\boldsymbol{\varepsilon}) \quad = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$$
$$\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

- A zatem:

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] =$$
$$= E\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Uwaga: jeżeli $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}$, czyli spełnione jest założenie **A2**, to:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

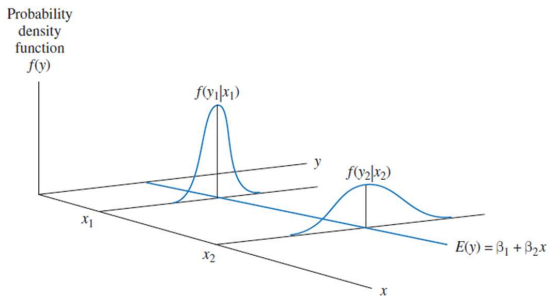
Przykład 6.1. Heteroskedastyczność

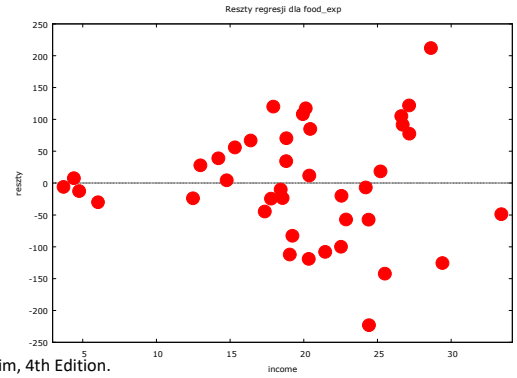
Rozważmy model ekonometryczny, w którym wydatki na żywność ($food\_exp$, USD) zależą od dochodu ($income$, 100USD). Dane pochodzą z pliku `food.gdt`.

Poniższe wykresy ukazuje zależność reszt regresji od dochodu:

- Jeśli wariancje dla wszystkich obserwacji są różne: **heteroskedastyczność**
- Jeśli wariancje dla wszystkich obserwacji są identyczne: **homoskedatyczny**

Pytanie: a jaka jest wariancja w tym modelu?

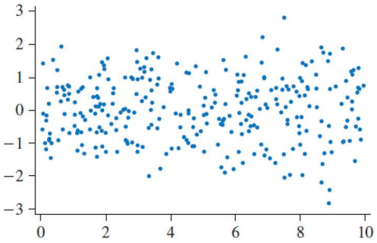




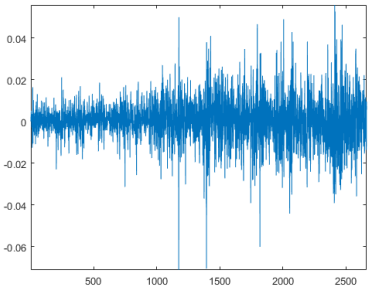
Źródło: Principles of Econometrics, R. Carter Hill, William E. Griffiths and Guay C. Lim, 4th Edition.

Wykrywanie heteroskedastyczności

Wykrywanie heteroskedastyczności
Wykres reszt modelu



- Brak widocznego wzorca kształtowania się reszt w modelu.
- Brak dowodów, aby odrzucić hipotezę o homoskedastyczności



- Widoczny wzorzec kształtowania się reszt
- Może występować problem heteroskedastyczności

Wykrywanie heteroskedastyczności
Test Breuscha-Pagana

Test Breuscha-Pagana weryfikuje następujący zespół hipotez:

$$H_0: Var(\varepsilon_i) = \sigma^2$$
$$H_1: Var(\varepsilon_i) = \sigma_i^2 \neq const.$$

- Przyjmijmy, że wariancje są opisywane przez:

$$\sigma_i^2 = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$$

gdzie z_{is} dla $s = 1,2,\ldots,S$ to wybrane cechy i-tego obiektu, zaś $h()$ jest dowolną funkcją.

- Przy prawdziwości H_0 zachodzi: $h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) = \sigma^2 = const.$

Etapy testu Breuscha-Pagana (przy założeniu liniowej postaci funkcji $h()$) :

1. Wybór determinantów wariancji: z_{is} dla $s = 1,2,\ldots,S$
2. Estymacja parametrów regresji pomocniczej: $\hat{\varepsilon}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$
3. Obliczenie R^2 dla regresji pomocniczej: R^2
4. Obliczenie statystyki testu: $BP = NR^2$
5. Podjęcie decyzji: $BP \overset{H_0}{\sim} \chi^2(N)$

Wykrywanie heteroskedastyczności
Test White'a

Test White'a jest specyficzną wersją testu BP, w którym przyjmuje się następujące założenia:

- $h()$ jest funkcją liniową
- zmienne z_{is} z regresji pomocniczej są potęgami zmiennych z regresji podstawowej

Przykład testu White'a:

- Model podstawowy: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$
- Model pomocniczy: $\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i1}^2 + \alpha_4 x_{i2}^2 + \alpha_5 x_{i1} x_{i2} + v_i$
- Hipoteza testu White'a: $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$
- Weryfikacja H_0: $W = NR^2 \sim \chi^2(5)$

Uwagi:

1. W wersji uproszczonej test White'a występuje bez zmiennych interakcyjnych, tutaj $z_5 = x_1 x_2$
2. Do weryfikacji H_0 możemy wykorzystać test Mnożników Lagrange'a (rozkładzie χ^2) lub uogólniony test Walda (o rozkładzie F, zob. Temat 3)

Przykład 6.2. Test na heteroskedastyczność

Na podstawie danych z pliku `cps5.gdt` oszacowano parametry modelu wyjaśniającego stawki godzinowe (*wage*, USD) przez wykształcenie (*educ*, lata) oraz doświadczenie (*exper*, lata). Wyniki są następujące:

```
Zmienna zależna (Y): wage
            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const      −17,5192        0,857335         −20,43    6,38e-091 ***
  educ         2,57373       0,0535119         48,10    0,0000    ***
  exper        0,193591      0,0114326         16,93    2,03e-063 ***
```

Następnie przeprowadzono test White'a

```
Zmienna zależna (Y): uhat^2
            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const     197,568          576,294          0,3428     0,7317
  educ      −62,3353          65,0715        −0,9580     0,3381
  exper       1,59934         15,8865         0,1007     0,9198
  sq_educ     4,18903          1,96094        2,136      0,0327   **
  X2_X3      −0,167275         0,844903      −0,1980     0,8431
  sq_exper    0,0505901        0,163467       0,3095     0,7570

  Wsp. determ. R-kwadrat = 0,002856
  Statystyka testu: TR^2 = 27,988749, z wartością p = P(Chi-kwadrat(5) > 27,988749) = 0,000037
```

Pytanie: jaki jest wniosek dotyczący heteroskedastyczności?

Wykrywanie heteroskedastyczności
Test Goldfelda-Quandta

Test Goldfelda-Quandta sprawdza, czy wariancja składnika losowego jest taka sama w dwóch, rozłącznych częściach próby, tj. podpróbach A i B.

- Aby przeprowadzić test, musimy podzielić próbę na dwie części. Można do tego wykorzystać zmienne binarne (płeć, lokalizacja, itp.) lub datę zmiany strukturalnej (np. wejście do UE)
- Zespół hipotez testu GQ jest następujący:

$$H_0: \sigma_A^2 = \sigma_B^2$$
$$H_1: \sigma_A^2 \neq \sigma_B^2$$

Etapy testu Breuscha-Pagana:

1. Podzielenie próby na A i B i obliczenie reszt dla każdej podpóby
2. Obliczenie wariancji reszt: $s_A^2 = \frac{\sum_{i \in A} \hat{\varepsilon}_i^2}{N_A - (K_A + 1)}$ oraz $s_B^2 = \frac{\sum_{i \in B} \hat{\varepsilon}_i^2}{N_B - (K_B + 1)}$
3. Obliczenie statystyki testu: $GQ = s_A^2 / s_B^2$
4. Podjęcie decyzji: $GQ \overset{H_0}{\sim} F_{(N_A - (K_A - 1), N_B - (K_B - 1))}$

Heteroskedastyczność: Metody postępowania

Heteroskedastyczność: metody postępowania

W przypadku wykrycia heteroskedastyczności istnieją **4 opcje postępowania**:

1. Zmiana specyfikacji modelu
2. Zmiana metody estymacji
3. Odporne błędy standardowe
4. Pozostawić model bez zmian

Heteroskedastyczność: metody postępowania
Zmiana specyfikacji modelu

Metodą najgłębiej ingerującą w model jest powrót do etapu "specyfikacji" modelu oraz:

1. Wykorzystanie logarytmów zmiennej objaśnianej (wtedy reszty są odchyleniami %)
2. Wyrażanie zmiennych względem miar wielkości (np. cena m^2 zamiast cena mieszkania)
3. Dodanie / usunięcie zmiennych objaśniających

Przykład. Rozważmy funkcję kosztu na kolei (Grilisches 1972):

$$C = \beta_0 + \beta_1 M + \beta_2 X + \varepsilon$$

gdzie C to koszt całkowity, M - długość trakcji kolejowej, zaś X – przychód.

Metoda 1: $\log C = \beta_0 + \beta_1 \log M + \beta_2 \log X + \varepsilon$

Metoda 2: $\frac{C}{M} = \beta_0 + \beta_1 \frac{1}{M} + \beta_2 \frac{X}{M} + \varepsilon$

Heteroskedastyczność: metody postępowania
Zmiana metody estymacji

Ważona MNK

- Rozważmy model o heteroskedastycznym składniku losowym:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \;\; Var(\varepsilon_i) = \sigma_i^2$$

- Załóżmy, że znamy wartości σ_i^2.
  Czy możemy wykorzystać tą informację przy estymacji parametrów modelu?
  Odpowiedź brzmi: TAK. Wystarczy podzielić obustronnie przez σ_i:

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_{1i}}{\sigma_i} + \beta_2 \frac{x_{2i}}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}, \; Var\left(\frac{\varepsilon_i}{\sigma_i}\right) = 1$$

- Przy estymacji modelu po transformacji minimalizujemy sumę kwadratów przeskalowanych reszt:

$$SSE = \sum_{i=1}^{N} \left(\frac{1}{\sigma_i}\hat{\varepsilon}\right)^2 = \sum_{i=1}^{N} (w_i \hat{\varepsilon})^2$$

  gdzie wagi $w_i = 1/\sigma_i$.

- Taka metoda jest określana jako ważona MNK.

UWAGA: W praktyce nie znamy wartości σ_i^2. Jak temu zaradzić?

Heteroskedastyczność: metody postępowania
Zmiana metody estymacji

Jak ustalić wagi ważonej MNK, gdy nie znamy wartości σ_i^2?

Metoda 1: Przyjąć, że σ_i jest proporcjonalne względem wybranej miary wielkości (przykładowo, w modelu cen mieszkań jest to powierzchnia: $\sigma_i^2 \propto \sigma \times size_i$)

Metoda 2: Oszacować wartości σ_i przy pomocy następującego modelu ekonometrycznego.

Etapy metody 2, czyli tzw. uogólnionej MNK (ang. generalized LS, GLS):

1. Oszacuj parametry regresji:

$$\ln \hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_S z_{Si} + \nu_i$$

2. Policz oszacowania wariancji ze wzoru:

$$\hat{\sigma}_i^2 = \exp(\hat{\alpha}_0 + \hat{\alpha}_1 z_{1i} + \hat{\alpha}_2 z_{2i} + \cdots + \hat{\alpha}_S z_{Si})$$

3. Zastosuj ważoną MNK, gdzie wagi wynoszą $w_i = 1/\hat{\sigma}_i$

UWAGI:

- przejście między etapami 1 i 2 wykorzystuje zależność: $E\left(\varepsilon_i^2\right) = \sigma_i^2$
- logarytm w etapie 1 jest wprowadzony, aby spełniony był warunek: $\hat{\sigma}_i^2 > 0$

Heteroskedastyczność: metody postępowania
Odporne błędy standardowe

- Wspomnieliśmy, że w przypadku heteroskedastyczności estymator MNK pozostaje nieobciążony. Problemem jest **niepoprawny wzór na błędy szacunku**.
- Trzecią metodą postępowania jest pozostawienie oszacowań MNK oraz policzenie "poprawnych" błędów szacunku. Wystarczy wykorzystać wzór (por. początek tego Tematu):

$$\boldsymbol{\Sigma}_{\widehat{\beta}} = Var(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}[\neq \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}]$$

- **Pytanie:** jak policzyć $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots \sigma_n^2)$?
- **Odpowiedź** zaproponował Halbert White. Zastąpił on σ_i^2 kwadratami reszt MNK $\hat{\varepsilon}_i^2$. Estymator wariancji White'a jest dany przez $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{\mathrm{White}} = \mathrm{diag}(\hat{\varepsilon}_1^2, \hat{\varepsilon}_2^2, \ldots, \hat{\varepsilon}_n^2)$.
- Błędy szacunku liczone przy wykorzystaniu macierzy:

$$\widehat{\boldsymbol{\Sigma}_{\widehat{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{\mathrm{White}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

określamy jako **błędy szacunku odporne na heteroskedastyczność**
(HC, heteroskedasticity consistent)

Przykład 6.3. Heteroskedastyczność
metody postępowania

Na podstawie danych o nieruchomościach sprzedawanych w Stockton (plik `stockton5.gdt`), ceny domów ($SPRICE$, USD) objaśniono przez ich powierzchnię ($LIVAREA$, 100 sq. feet).

Oszacowania MNK:

```
współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const      −30637,5         2263,44         −13,54    2,15e-040 ***
  livarea      9466,74         132,089         71,67     0,0000    ***
Statystyka testu White'a: TR^2 = 291,829570, z wartością p = P(Chi-kwadrat(2) > 291,829570) = 0,000000
```

Oszacowania ważonej MNK, gdzie wagi są odwrotnie proporcjonalne do $LIVAREA$

```
           współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const       −8892,47        1827,51         −4,866    1,21e-06  ***
  livarea      8048,75         132,925        60,55     0,0000    ***
```

Oszacowania MNK z odpornymi błędami

```
współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const      −30637,5         4621,64         −6,629    4,09e-011 ***
  livarea      9466,74         315,471        30,01     3,28e-170 ***
```

Zadania

Zadanie 6.1

Odpowiedz na następujące pytania:

a. Czy heteroskedastyczność prowadzi do obciążenia estymatora MNK?

b. Czy heteroskedastyczność prowadzi do niepoprawnych błędów szacunku?

c. Czy w przypadku heteroskedastyczności wyniki testów istotności są miarodajne?

d. Rozważmy następujący model dla gospodarstw domowych
(C – wydatki na książki, Y – dochód, $W = 1$ – wykształcenie wyższe):

$$C_i = \beta_0 + \beta_1 Y_i + \beta_2 W_i + \varepsilon_i$$

- Dlaczego możemy spodziewać się heteroskedastyczności?
- Jaki test na heteroskedastyczność byś zaproponował?
- Co możemy zrobić, aby uwzględnić występowanie heteroskedastyczności?

Zadanie 6.2

Przeprowadź następującą symulację w programie Gretl / R.

Ustal parametry opisujące DGP (=data generating proces) oraz próby:

$N = 1000$:	liczba obserwacji
$\gamma = 10$	parametr dyspersji

a. Stwórz pusty zbiór danych z zakresem N obserwacji i wygeneruj zmienne:

zmienna skali:	$\sigma \sim U(1, \gamma)$
składnik losowy:	$\varepsilon = \sigma u$, gdzie $u \sim N(0,1)$
zmienna objaśniająca:	$x = \sigma v$, gdzie $v \sim N(0,1)$
zmienna objaśniana:	$y = 5 + 3x + \varepsilon$,

b. Dla modelu:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Oszacuj parametry MNK i przeprowadź test White'a
- Stwórz wykres reszt $\hat{\varepsilon}$ względem $|x|$. Oszacuj model z odpornymi błędami szacunku. Porównaj z oszacowaniami MNK.
- Zastosuj ważoną MNK, gdzie wagi są dane przez $w = 1/\sigma$. Porównaj z oszacowaniami MNK.

d. Zamień γ na $\gamma = 1$ i powtórz czynności z punktów **a.-b.**

Zadanie 6.3

W pliku `TaylorRule.gdt` zawarte są dane o poziomie stopy procentowej (*IR*, w %), inflacji rocznej (*INF*, %) oraz indeksu aktywności gospodarczej (*Y*, 100 jeżeli normalny poziom aktywności) dla wybranych krajów OECD.

a. Wybierz kraj, który będziesz analizował i oszacuj parametry następującego modelu:

$$IR_t = \beta_0 + \beta_1 INF_t + \beta_2 Y_t + \varepsilon_t$$

b. Narysuj wykres reszt modelu w czasie

c. Przeprowadź test White'a

d. Oblicz błędy odporne (w Gretlu wybierz HC1) i porównaj z błędami z punktu a.

e. Czy potrafisz wyjaśnić, jakie są źródła heteroskedastyczności w tym modelu?

Zadanie 6.4

W pliku Dane `stockton5.gdt` zawarte są dane dotyczące sprzedaży nieruchomości., m.in. $SPRICE$ ($) – cena domu, $LIVAREA$ (100 sq. feet) - powierzchnia, $POOL = 1$, jeżeli jest basen. Oszacuj model:

$$sprice_i = \beta_0 + \beta_1 livarea_i + \beta_2 pool_i + \varepsilon_i$$

a. Narysuj wykres reszt $\widehat{\varepsilon_i}$ względem powierzchni $livarea_i$
b. Przeprowadź test White'a
c. Zastosuj odporne błędy szacunku (jak się zmieniły wyniki?)
d. Zastosuj ważoną MNK, gdzie wagi to $w_i = 1/livarea_i$
e. Oszacuj model dla zmiennej $\ln sprice$.
Czy ta zmiana specyfikacji rozwiązała problem heteroskedastyczności.
f. [trudniejsze] Przeprowadź test Goldfelda-Quandta przy podziale próby ustalonym przez zmienną $pool$. Wskazówki dla GRETL:

dane	→ sortowanie danych względem *pool*
próba	→ zakres próby (próba A i B)
zmienna	→ statystyki opisowe ($\hat{\sigma}_A^2$ i $\hat{\sigma}_B^2$)
narzędzia	→ wyznaczanie wartości p

Zadanie 6.5

Plik `cps5_small.gdt` zawiera dane o stawce godzinowej, wykształceniu i innych zmiennych zebranych w Current Population Survey (CPS) z 2008 roku.

a. Oszacuj model, w którym stawka godzinowa ($WAGE$) zależą od wykształcenia ($EDUC$) oraz doświadczenia (*EXPER*).
b. Narysuj wykres wartości bezwzględnych reszt $|\widehat{\varepsilon_i}|$ względem płac
c. Przeprowadź test White'a
d. Zastosuj odporne błędy szacunku. Czy są inne niż te z punktu **a.**?
e. Oszacuj model dla zmiennej $\ln WAGE$ i powtórz kroki **a.-d.**
Czy ta zmiana specyfikacji rozwiązała problem heteroskedastyczności.
f. Poszukaj zmiany specyfikacji, która usuwa (lub łagodzi) problem heteroskedastyczności
g. Dla modelu z punktu **a.** zastosuj uogólnioną MNK
(zbuduj model ekonometryczny dla wariancji składnika losowego)

Zadanie 6.6

Rozważmy model ekonometryczny, w którym wydatki na żywność ($food\_exp$, USD) zależą od dochodu ($income$, 100USD). Dane znajdują się w pliku `food.gdt`:

a. Oszacuj parametry dwóch modeli:

Model A: $$food\_exp_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

Model B: $$\ln(food\_exp_i) = \beta_0 + \beta_1 \ln(income_i) + \varepsilon_i$$

b. Przeprowadź test White'a na heteroskedastyczność.

c. Czy wyniki testu White'a są dowodem homoskedastyczności?
Zaproponuj inną wersję testu Breuscha-Pagana dla modelu B.

d. Zaproponuj wagi do ważonej metody MNK w modelu A.
Oszacuj model A uogólnioną MNK i porównaj wyniki z punktem a.

Temat 7

Weryfikacja modelu: autokorelacja

Zuzanna Wosko i Karol Szafranek

- Definicja autokorelacji
- Konsekwencje autokorelacji dla estymatora MNK
- Testy Durbina-Watsona i Breuscha-Pagana
- Uogólniona MNK
- Błędy odporne na autokorelację

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej i zbioru zmiennych objaśniających
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Etapy weryfikacji modelu

1. Oceny parametrów i ich znaki
2. Istotność parametrów
3. Dopasowanie modelu do danych
4. Specyfikacja modelu / postać funkcyjna
5. **Własności składnika losowego**
6. Stabilność parametrów

Definicja autokorelacji

Definicja autokorelacji

Autokorelacja składnika losowego dotyczy następującego założenia MNK:

$$\textbf{A3.}\ Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

Występowanie autokorelacji oznacza, że zachodzi:

$$cov(\varepsilon_t, \varepsilon_s) = \sigma^2\rho_{ts} \neq 0 \quad \text{dla} \quad t \neq s$$

Macierz kowariancji składnika losowego

brak autokorelacji

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

autokorelacja

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1N} \\ \rho_{21} & 1 & \dots & \rho_{2N} \\ \dots & \dots & \dots & \dots \\ \rho_{N1} & \rho_{N2} & \dots & 1 \end{bmatrix}$$

Autokorelacja:
konsekwencje dla estymatora MNK

1. Estymator MNK jest nadal nieobciążony, ale przestaje być najbardziej efektywnym estymatorem liniowym: **istnieje inny estymator liniowy o mniejszej wariancji**.

2. Wzór z Tematu 2 na wariancję estymatora MNK :

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] \overset{A3}{=} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

   przestaje być poprawny, czyli błędy szacunku liczone w standardowy sposób są również niepoprawne.

3. Punkt 2 implikuje, że statystyki testów istotności nie mają rozkładu t-Studenta. Podobnie statystyki testów istotności łącznej nie pochodzą z rozkładu *F* lub χ^2.

Wniosek: występowanie autokorelacji (podobnie jak w przypadku heteroskedastyczności) utrudnia weryfikację modelu.

Autokorelacja:
konsekwencje dla estymatora MNK

- Dlaczego Wzór z Tematu 2 na wariancję estymatora MNK jest niepoprawny:

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] \overset{A3}{=} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Zauważmy, że (por. Temat 2):

$$Var(\boldsymbol{\varepsilon}) \quad = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \neq \sigma^2\mathbf{I}$$
$$\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

- A zatem:

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] =$$
$$= E\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Zauważ: jeżeli $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}$, czyli spełnione jest założenie **A2**, to:

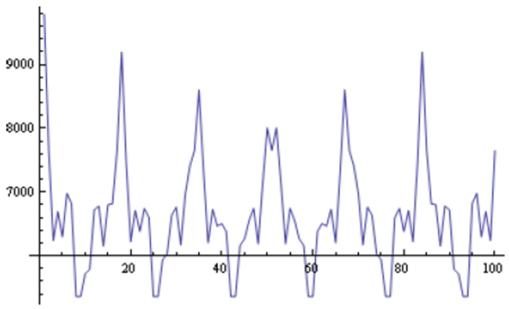
$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Przykład 7.1. Autokorelacja

Jeżeli reszty modelu ekonometrycznego układają się w sposób nielosowy, to jest to oznaka **autokorelacji**

Przez nielosowy układ reszt rozumiemy:

- Długie serie reszt o tym samym znaku
- Zmiana znak w systematyczny sposób (np. zależny od pory roku)
- Inny, stały wzorzec kształtowania się reszt



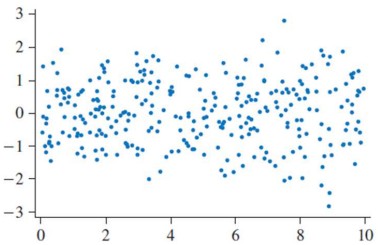
Źródła autokorelacji

- Charakterystyka zmiennej objaśnianej (wysoka inersja)
- Niewłaściwa postać funkcyjna modelu
- Niewłaściwa dynamiczna struktura modelu:
  a. brak opóźnionej zmiennej objaśnianej w roli regresora
  b. regresory powinny być opóźnione
  c. brak zmiennej czasowej
- Pominięcie ważnej zmiennej objaśniającej
- Transformacje szeregów czasowych – interpolacje, wygładzania, agregacje

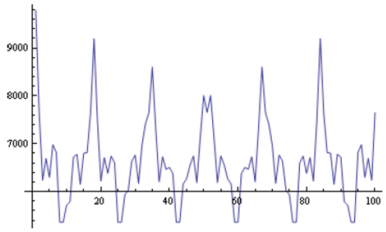
Uwaga: Autokorelacja to problem charakterystyczny dla modeli szeregów czasowych i nie występuje w modelach opartych o dane przekrojowe. Dlaczego?

Wykrywanie autokorelacji

Wykrywanie autokorelacji
Wykres reszt modelu



- Brak widocznego wzorca kształtowania się reszt w modelu.
- Brak dowodów, by odrzucić hipotezę o braku autokorelacji



- Widać powiązania reszt z różnych okresów
- Może występować problem z autokorelacją

Powtórka ze statystyki

Autokorelacja w próbie i populacji

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

współczynniki korelacji w populacji

$$\rho_1 = \frac{Cov(y_t, y_{t-1})}{\sqrt{Var(y_t)Var(y_{t-1})}}$$

współczynnik autokorelacji rzędu 1 w populacji

$$\widehat{\rho_1} = \frac{\sum_{t=2}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Współczynnik autokorelacji rzędu 1 w próbie

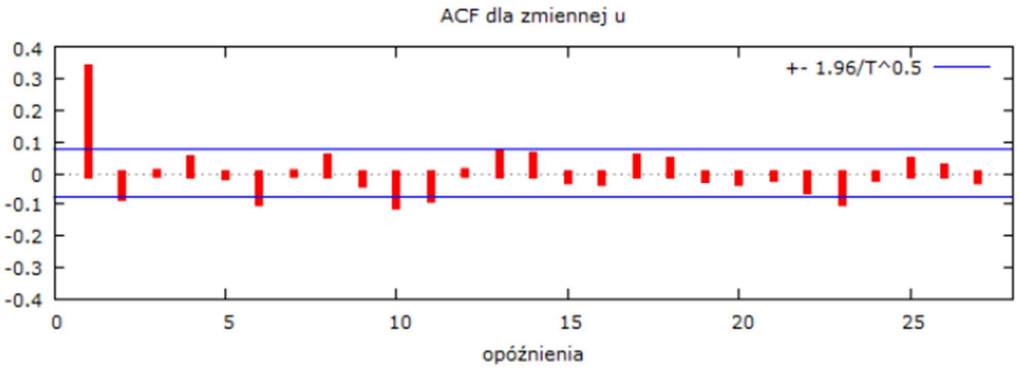
$$\widehat{\rho_k} = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Współczynnik autokorelacji rzędu k w probie dla szeregu y

Wykrywanie autokorelacji

Korelogram

ACF – funkcja autokorelacji (autocorrelation function)



Niebieska 95% przedział ufności (zob. test Bartletta):

$$P\left(-1.96\sqrt{1/T} \leq \widehat{\rho_k} \leq 1.96\sqrt{1/T}\right) = 95\%$$

Wykrywanie autokorelacji
test Durbina-Watsona

Test Durbina-Watsona weryfikuje następujący zespół hipotez:

H_0: $\rho_1 = 0$ (brak autokorelacji)
H_1: $\rho_1 > 0$ (dodatnia autokorelacja)

Statystyka testowa:

$$d = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2} \approx 2(1 - \widehat{\rho_1})$$

Wartości krytyczne z tablic obejmują dwie wartości: d_L (lower) and d_U (upper).

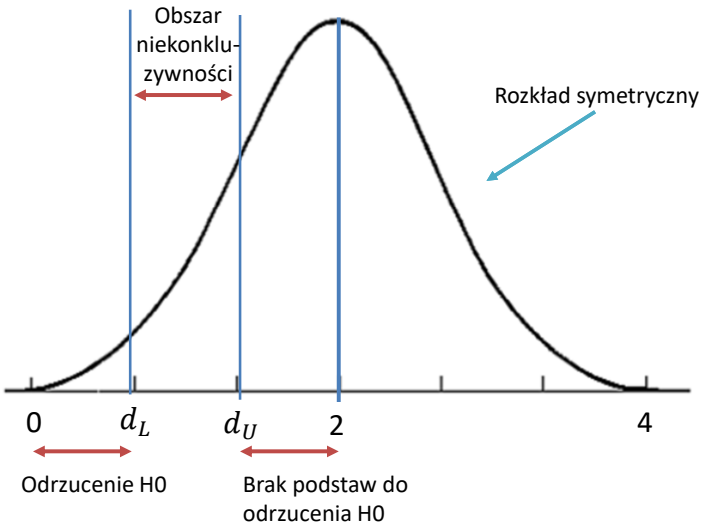
Wnioskowanie na podstawie testu Durbina Watsona:

$0 \le d < d_L$: odrzucamy H_0 na rzecz H_1, czyli autokorelacja dodatnia
$d_L \le d < d_U$: przedział niekonkluzywności
$d_U \le d \le 4$: brak podstaw o odrzucenia H_0, czyli brak autokorelacji

Ważne:
Jeżeli $H_1: \rho_1 > 0$ to wnioskujemy dla $d' = 4 - d$

Wykrywanie autokorelacji
test Durbina-Watsona - ograniczenia

Test Durbina-Watsona nie może być stosowany gdy:
- rozkład reszt nie jest normalny,
- wśród regresorów jest opóźniona zmienna objaśniana,
- nie ma wyrazu wolnego w specyfikacji modelu,
- chcemy sprawdzić wyższe rzędy autokorelacji.



Wykrywanie autokorelacji
test Durbina-Watsona – wykorzystanie tablic

Fragment tablicy rozkładu statystyki testu Durbina Watsona:

	k – liczba oszacowanych współczynników (bez wyrazu wolnego)									
a=5% N	k=1		k=2		k=3		k=4		k=5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.61018	1.40015								
7	0.69955	1.35635	0.46723	1.89636						
8	0.76290	1.33238	0.55907	1.77711	0.36744	2.28664				
9	0.82428	1.31988	0.62910	1.69926	0.45476	2.12816	0.29571	2.58810		
10	0.87913	1.31971	0.69715	1.64134	0.52534	2.01632	0.37602	2.41365	0.24269	2.82165
11	0.92733	1.32409	0.75798	1.60439	0.59477	1.92802	0.44406	2.28327	0.31549	2.64456
12	0.97076	1.33137	0.81221	1.57935	0.65765	1.86397	0.51198	2.17662	0.37956	2.50609
13	1.00973	1.34040	0.86124	1.56212	0.71465	1.81593	0.57446	2.09428	0.44448	2.38967
14	1.04495	1.35027	0.90544	1.55066	0.76666	1.77882	0.63206	2.02955	0.50516	2.29593
15	1.07697	1.36054	0.94554	1.54318	0.81396	1.75014	0.68519	1.97735	0.56197	2.21981
16	1.10617	1.37092	0.98204	1.53860	0.85718	1.72773	0.73400	1.93506	0.61495	2.15672

Wykrywanie autokorelacji
test mnożników Lagrange'a

Test mnożników Lagrange'a weryfikuje hipotezę o występowaniu autokorelacji składnika losowego rzędu P.

Dla modelu $\varepsilon_t = \rho_1\varepsilon_{t-1} + \cdots + \rho_P\varepsilon_{t-P} + \nu_t$ hipoteza zerowa jest następująca:

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_P = 0$$

Etapy przeprowadzania testu mnożników Lagrange'a:

1. Estymacja modelu i obliczenie reszt: $y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$
2. Estymacja parametrów regresji pomocniczej: $\hat{\varepsilon}_t = \mathbf{x}_t'\boldsymbol{\gamma} + \rho_1\hat{\varepsilon}_{t-1} + \cdots + \rho_P\hat{\varepsilon}_{t-P} + v_i$
3. Weryfikacja hipoteza testu LM: $H_0: \rho_1 = \rho_2 = \cdots = \rho_P = 0$
4. Obliczenie statystyki: $LM = nR^2 \sim \chi^2(p)$
5. Podjęcie decyzji: $LM \overset{H_0}{\sim} \chi^2(P)$ dla dużej liczby obserwacji
   $LM \overset{H_0}{\sim} F\big(P, T-(K+1)\big)$ dla małej próby

Przykład 7.2. Testy na autokorelacje

Na podstawie danych z pliku `PhillipsCurve.gdt` oszacowano parametry modelu wyjaśniającego inflację HICP w Polsce (INF_PL, % r/r) przez stopę bezrobocia (U_PL, % sa).

```
Zmienna zależna (Y): Inf_PL

           współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const       1,65904        0,626213         2,649      0,0086   ***
  U_PL        0,168870       0,0460158        3,670      0,0003   ***
```

Wyniki testu Durbina Watsona:

```
Stat. Durbina-Watsona = 0,0135422
wartość p = 1,05471e-015
```

Wyniki testu mnożników Lagrange'a:

```
Statystyka testu: LMF = 342,454698,
z wartością p = P(F(12,250) > 342,455) = 1,39e-147

Statystyka testu: TR^2 = 248,860498,
z wartością p = P(Chi-kwadrat(12) > 248,86) = 2,36e-046
```

Pytanie: jaki jest wniosek dotyczący autokorelacji?

Autokorelacja: Metody postępowania

Autokorelacja: metody postępowania

W przypadku wykrycia autokorelacji istnieją **4 opcje postępowania**:

1. Odporne błędy szacunku
2. Zmiana metody estymacji
3. Zmiana specyfikacji modelu
4. Pozostawić model bez zmian

Autokorelacja: metody postępowania
Odporne błędy szacunku

- Wspomnieliśmy, że w przypadku występowania autokorelacji reszt estymator MNK pozostaje nieobciążony. Problemem jest **niepoprawny wzór na błędy szacunku**.
- Pierwszą metodą postępowania jest pozostawienie oszacowań MNK oraz policzenie "poprawnych" błędów szacunku. Wystarczy wykorzystać wzór (por. początek tego Tematu):

$$\boldsymbol{\Sigma}_{\hat{\beta}} = Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} [\neq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}]$$

- **Pytanie:** jak obliczyć powyższe wyrażenie?
- **Odpowiedź** zaproponowali Newey i West (1987), którzy wykorzystali metody ekonometrii nieparametrycznej:

$$\widehat{\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}}^{NW} = \frac{1}{T}\left(\sum_{t=1}^{T}\hat{\varepsilon}_t^2\mathbf{x}_t'\mathbf{x}_t + \sum_{j=1}^{L}\sum_{t=j+1}^{T}\omega_j\hat{\varepsilon}_t\hat{\varepsilon}_{t-j}\left[\mathbf{x}_t\mathbf{x}_{t-j}' + \mathbf{x}_{t-j}\mathbf{x}_t'\right]\right)$$

gdzie L to maksymalne opóźnienie, zaś wagi wynoszą $\omega_j = 1 - \frac{j}{L}$ (możliwe są inne warianty)

- Błędy szacunku liczone przy wykorzystaniu macierzy:

$$\widehat{\boldsymbol{\Sigma}_{\hat{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}}^{NW}(\mathbf{X}'\mathbf{X})^{-1}$$

określamy jako **błędy szacunku odporne na heteroskedastyczność i autokorelacje** (HAC, heteroskedasticity and autocorrelation consistent)

Autokorelacja: metody postępowania

Zmiana metody estymacji - uogólniona MNK (UMNK)

- Zmiana estymatora pozwala na uzyskanie bardziej efektywnego estymatora niż estymator MNK.
- Przyjmijmy, że w modelu $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, nie jest spełnione założenie **A2**, tj.:

$$Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}$$

- W takim przypadku, możemy wykorzystać estymator UMNK, którego logika jest następująca:
  - Dokonajmy dekompozycji $\boldsymbol{\Sigma}^{-1} = \mathbf{V}'\mathbf{V}$
  - Zdefiniujmy zmienne $\tilde{\boldsymbol{\varepsilon}} = \mathbf{V\varepsilon}$, $\tilde{\mathbf{y}} = \mathbf{Vy}$ oraz $\tilde{\mathbf{X}} = \mathbf{VX}$. Zauważmy, że:

$$Var(\tilde{\boldsymbol{\varepsilon}}) = Var(\mathbf{V\varepsilon}) = \mathbf{V}Var(\boldsymbol{\varepsilon})\mathbf{V}' = \mathbf{V\Sigma V}' = \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}' = \mathbf{V}\mathbf{V}^{-1}\mathbf{V}'^{-1}\mathbf{V}' = \mathbf{I}$$

  - A zatem model $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ spełnia **A2**!

- Estymator UMNK uzyskany na przekształconych danych jest postaci:

$$\hat{\boldsymbol{\beta}}^{UMNK} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

lub po przekształceniach:

$$\hat{\boldsymbol{\beta}}^{UMNK} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

UWAGA: Aby zastosować UMNK, należy oszacować macierz $\boldsymbol{\Sigma}$. Jak to zrobić?

Autokorelacja: metody postępowania

Zmiana metody estymacji - metoda Cochrane'a-Orcutta

Metoda Cochrane-Orcutta jest przykładem UMNK dla następującego modelu:

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$$

Etapy metody

1. Oblicz wartość parametru ρ jako współczynnik autokorelacji w próbie dla reszt
2. Oblicz przetransformowane zmienne: $\tilde{y}_t = y_t - \rho y_{t-1}$ oraz $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \rho\mathbf{x}_{t-1}$
3. Oszacuj MNK parametry modelu: $\tilde{y}_t = \tilde{\mathbf{x}}_t'\boldsymbol{\beta} + \tilde{\varepsilon}_t$
4. Umieść oszacowane parametry w specyfikacji przed transformacją i policz reszty
5. Powtarzaj etapy 1-4 aż do uzyskania zbieżności
   (wartości parametru ρ z kolejnych iteracji będą prawie takie same)

Przykład 7.2. Autokorelacja
metody postępowania

Na podstawie danych z pliku `PhillipsCurve.gdt` oszacowano parametry modelu wyjaśniającego inflację HICP w Polsce (INF_PL, % r/r) przez stopę bezrobocia (U_PL, % sa).

Oszacowania MNK:

współczynnik	błąd standardowy	t-Studenta	wartość p	
const	1,65904	0,626213	2,649	0,0086	***
U_PL	0,168870	0,0460158	3,670	0,0003	***

Oszacowania MNK z odpornymi błędami

współczynnik	błąd standardowy	t-Studenta	wartość p	
const	1,65904	0,905275	1,833	0,0680	*
U_PL	0,168870	0,0735685	2,295	0,0225	**

Oszacowania UMNK (procedura Cochrane-Orcutt)

współczynnik	błąd standardowy	t-Studenta	wartość p
const	1,90589	1,44779	1,316	0,1892
U_PL	−0,0417189	0,0951710	−0,4384	0,6615

Autokorelacja: metody postępowania
Zmiana specyfikacji modelu

Metodą najgłębiej ingerującą w model jest powrót do etapu "specyfikacji" modelu. W szczególności, autokorelację czasami można wyeliminować poprzez rozszerzenie dynamicznej specyfikacji modelu.

W tym celu możemy:

- wprowadzić opóźnioną zmienną objaśnianą do zbioru regresorów,
- zmienić rząd opóźnienia zmiennej objaśnianej,
- wprowadzić opóźnienia regresorów,
- dodać zmienną czasową / zmienne sezonowe

Zadania

Zadanie 7.1

Odpowiedz na następujące pytania:

a. Czy autokorelacja reszt prowadzi do obciążenia estymatora MNK?

b. Czy autokorelacja reszt prowadzi do niepoprawnych błędów szacunku?

c. Czy w przypadku występowania autokorelacji wyniki testów istotności są miarodajne?

d. Rozważmy następujący model dla gospodarstw domowych (π_t – inflacja CPI r/r, $\hat{y}_t$ – luka popytowa, oil_t – ceny ropy naftowej r/r):

$$\pi_t = \beta_0 + \beta_1 \hat{y}_t + \beta_2 oil_t + \varepsilon_t$$

- Dlaczego możemy spodziewać się autokorelacji?
- Jaki test na autokorelację byś zaproponował?
- Co możemy zrobić, aby uwzględnić występowanie autokorelacji?

Zadanie 7.2

Wykorzystując plik z danymi **PhillipsCurve.gdt,** dla wybranego kraju:

a. Oszacuj MNK model objaśniający inflację w zależności od stopy bezrobocia.

b. Sprawdź autokorelację rzędu pierwszego i wyższych rzędów, wykorzystując korelogram i odpowiednie testy.

c. Czy błędy szacunku Neweya-Westa (HAC) są inne niż błędy liczone w tradycyjny sposób?

d. Oszacuj parametry metodą Cochrane'a-Orcutta.

e. Spróbuj zmienić specyfikację modelu, która pozwala zmniejszyć problem autokorelacji.

Zadanie 7.3

W pliku `TaylorRule.gdt` zawarte są dane o poziomie stopy procentowej (*IR*, w %), inflacji rocznej (*INF*, %) oraz indeksu aktywności gospodarczej (*Y*, 100 jeżeli normalny poziom aktywności) dla wybranych krajów OECD.

a. Wybierz kraj, który będziesz analizował i oszacuj parametry następującego modelu:

$$IR_t = \beta_0 + \beta_1 INF_t + \beta_2 Y_t + \varepsilon_t$$

b. Narysuj wykres reszt modelu w czasie
c. Przeprowadź test Durbina-Watsona
d. Przeprowadź test mnożników Lagrange'a
e. Oblicz błędy odporne (w Gretlu wybierz HAC) i porównaj z błędami z punktu a.
f. Czy potrafisz wyjaśnić, jakie mogą być źródła autokorelacji w tym modelu?

Zadanie 7.4

Wykorzystując zbiór `capm5.gdt`, wybierz akcję i oszacuj MNK jednoczynnikowy model Sharpe'a.

Sprawdź autokorelację poprzez:

1. Analizę wykresów reszt
2. Stworzenie korelogramów
3. Przeprowadzenie testów DW i LM
4. Sprawdź, czy $d \approx 2(1-\hat{\rho}_1)$ jest prawdziwe.

Podpowiedź:

Model Sharpe'a:

Stopy zwrotu z akcji A = alpha + beta * stopy zwrotu z indeksu rynkowego + zmienna losowa

Zadanie 7.5

W wyniku estymacji KMNK parametrów modelu ekonometrycznego otrzymano następujące rezultaty:

$$\hat{y}_t = 110 + 1{,}3x_{1t} + 0{,}5x_{2t}$$

Wektor reszt empirycznych jest następujący:

[-4 4 3 -1 6 10 -25 6 10 -1 2 -11 -5 -3 4]

- Czy mamy do czynienia z autokorelacją I rzędu składnika losowego?
- Przeprowadź test DW.

Zadanie 7.6

Na podstawie danych kwartalnych 1993:1 – 2012:2 oszacowano parametry następującego modelu ekonometrycznego MNK (w nawiasach podano wartości statystyki t):

$$\underset{}{\ln \hat{W}_t} = \underset{(13,97)}{5,38} + \underset{(6,78)}{0,94} \ln P_t + \underset{(3,41)}{0,52} \ln XN_t - \underset{(-2,22)}{0,01} \ln U_{t-1}$$

R^2=0,97 DW=1,87 JB=7,83 MAPE=1,04

gdzie:

W– przeciętna miesięczna płaca nominalna brutto w przemyśle w zł,
U – stopa bezrobocia (stan w końcu okresu, %),
P – indeks cen towarów i usług konsumpcyjnych,
XN – wielkość produkcji sprzedanej przemysłu (w mln zł na zatrudnionego)

a. Zweryfikuj model pod względem istotności parametrów
b. Przeprowadź weryfikację merytoryczną oszacowań parametrów
c. Jakie są wnioski na temat autokorelacji z testu DW?

Temat 8

Specyfikacja modelu.

Modele dynamiczne

Jakub Mućk

- Operator opóźnienia
- Model z rozkładem opóźnień (DL)
- Model autoregresyjny (AR)
- Model autoregresyjny z rozkładem opóźnień (AR)
- Mnożnik krótkookresowy i długookresowy
- Funkcja reakcji na impuls

Wprowadzenie

- Dynamiczna natura procesów ekonomicznych:

$$y_t = f\left(x_t, x_{t-1}, x_{t-2}, \ldots\right). \tag{1}$$

- Wysoka persystencja/inercja zmiennych ekonomicznych.
- Niepoprawna struktura dynamiczna modelu prowadzi zazwyczaj do autokorelacji składnika losowego
- Kluczowe założenie: stacjonarne szeregi czasowe, tj. brak wyraźnych trendów, Zdolność powrotu do wartości średniej (*mean reversion*).
- Modee dynamiczne, które uwzględniają persystencję oraz(lub) dynamiczną naturę procesów ekonomicznych: Models that accounts for persistence/ dynamic nature of relationship:
  - modele autoregresywne (*autoregressive models*),
  - modele z rozkładem opóźnień (*distributed lag models*),
  - modele autoregresywne z rozkładem opóźnień (*autoregressive distributed lag models*).

Podstawowe definicje

- Szereg czasowy y_t stanowią obserwacje uporządkowane czasem t, gdzie $t = 1, 2, \ldots, T$.
- Operator opóźnień (*lag operator*) L:

$$L\left(y_t\right) = y_{t-1}. \tag{2}$$

- Operator różnicowania/pierwsze różnice (*Difference operator/first difference*) Δ:

$$\Delta\left(y_t\right) = \left(1 - L\right) y_{t-1} = y_t - y_{t-1}. \tag{3}$$

- Stopa wzrostu (dynamika) mierząca procentową zmianę zmiennej y_t:

$$g = \frac{y_t - y_{t-1}}{y_{t-1}}. \tag{4}$$

- Logarytmiczna stopa wzrostu:

$$\Delta \ln y_t = \ln y_t - \ln y_{t-1} = \ln \frac{y_t}{y_{t-1}} = \ln \frac{y_{t-1} \times (1+g)}{y_{t-1}} \approx g. \tag{5}$$

Model z rozkładem opóźnień

Model z rozkładem opóźnień

- Model z rozkładem opóźnień (*distributed lags model*) rzędu K (oznaczany jako DL(K)):

$$y_t = \mu + \sum_{i=0}^{K} \beta_i x_{t-i} + \varepsilon_t, \tag{6}$$

gdzie
  - y_t –zmienna objaśniana,
  - x_t – zmienne objaśniające,
  - ε_t – składnik losowy.

- **Możnik krótkookresowy (*short-run multiplier*, β^{SR}):**

$$\beta^{SR} = \beta_0. \tag{7}$$

- **Monożnik długookresowy (*long-run multiplier*, β^{LR}):**

$$\beta^{LR} = \beta_0 + \beta_1 + \ldots + \beta_K. \tag{8}$$

- Parametry równania (6) można oszacować metodą najmniejszych kwadratów.

Modele autoregresyjne

AR(1)

- **Model autoregresyjny (*autoregressive model*) pierwszego rzędu (oznaczane jako AR(1))**

$$y_t = \mu + \rho y_{t-1} + \varepsilon_t \tag{9}$$

  gdzie ε_t to składnik losowy oraz $\varepsilon_t \sim \mathcal{N}(0, \sigma)$.
- **Kluczowe założenie:** $|\rho| < 1$
- Parametr ρ mierzy persystencję/inercję szeregu czasowego y_t.
  - Jeżeli ρ jest bliskie 0 to wtedy efekt egzogenicznych zaburzeń (mierzonych przez ε_t) jest absorbowany natychmiast.
  - Jeżeli ρ jest bliskie 1 to wtedy efekt egzogenicznych zaburzeń stopniowo wygasa.
- Wybrane charakterystyki szeregu czasowego y_t w przypadku, gdy jest generowany przez proces AR(1):

$$\mathbb{E}(y_t) = \frac{\mu}{1-\rho}, \tag{10}$$

$$Var(y_t) = \frac{\sigma^2}{1-\rho^2}. \tag{11}$$

- Okres połowicznego wygasania (*half-life*):

$$hl = \frac{\ln(0.5)}{\ln(\rho)}. \tag{12}$$

AR(1) oraz IRF I

- Jak wygląda/przebiega efekt zaburzeń losowych (ε_t) na zmienną objaśnianą?
- Rozważmy uproszczony przypadek ($\mu = 0$) dla modelu AR(1):

$$y_t = \rho y_{t-1} + \varepsilon_t, \tag{13}$$

i załóżmy, że $\varepsilon_0 = 1$ and for $t > 1$, $\varepsilon_t = 0$. Wtedy:

$$\begin{aligned} y_0 &= 0 \times \rho + 1 = 1 \\ y_1 &= y_0 \times \rho + 0 = 1 = \rho \\ y_2 &= y_1 \times \rho + 0 = \rho\rho = \rho^2 \\ &\dots \end{aligned}$$

lub bardziej ogólnie:

$$y_t = \rho^t. \tag{14}$$

- Biorąc pod uwagę fakt, że ε_0 jest równe 0 w powyższym przykładzie, model AR(1) może zostać wyrażony z wykorzystaniem postaci średniej ruchomej (*moving-average*):

$$y_t = \sum_{i=1}^{\infty} \rho^i \varepsilon_{t-i} + \varepsilon_t = \varepsilon_t + \sum_{i=1}^{\infty} \phi_i \varepsilon_{t-i}. \tag{15}$$

AR(1) oraz IRF II

- Postać średniej ruchomej pozwala zilustrować w jaki sposób zmienna objaśniana reaguje w czasie na egzaogeniczne zaburzenia:

$$\frac{\partial \mathbb{E}\left(y_t\right)}{\partial \varepsilon_{t-i}} = \phi_i = \rho^i. \tag{16}$$

- **Funkcja reakcji na impuls (*Impulse response function*)** ilustruje oczekiwaną ewolucję zmiennej objaśnianej w reakcji na jednostkowe zaburzenie egzogeniczne:

$$\left\{1, \phi_1, \phi_2, \dots\right\}. \tag{17}$$

Dla zmiennej objaśnianej opisanej modelem AR(1):

$$\left\{1, \rho, \rho^2, \dots\right\}. \tag{18}$$

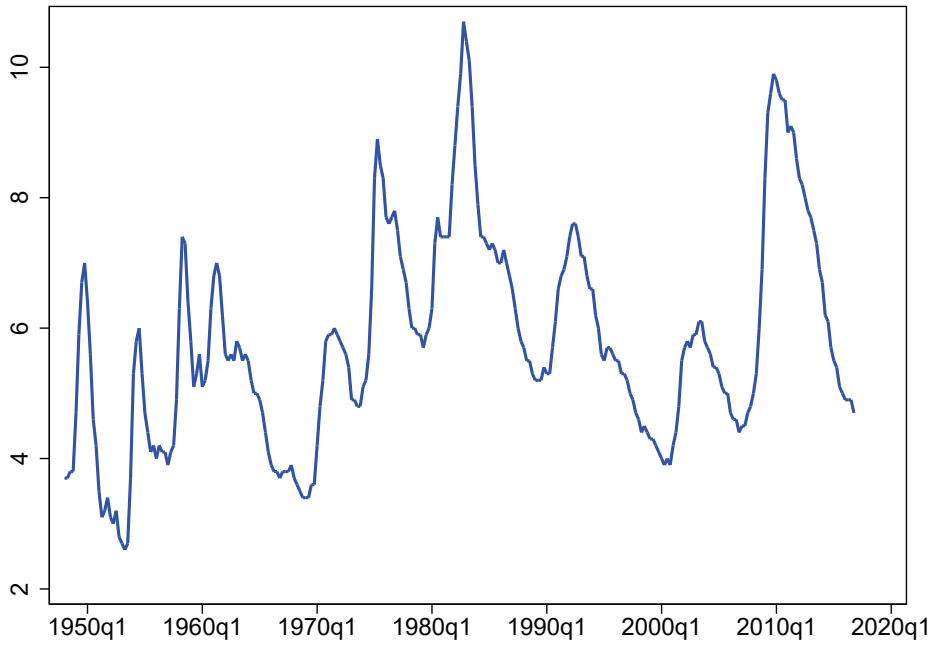
AR(P)

- **Model autoregresyjny P-tego rzędu (oznaczany jako AR(P))**

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_P y_{t-P} + \varepsilon_t. \tag{19}$$

- Parametry powyższego równania można oszacować metodą najmniejszych kwadratów.
- Modele autoregresyjne wyższych rzędów:
  - są wykorzytywane w analize bardziej złożonych własności dynamicznych szeregów czasowych,
  - są wykorzystywane w prognozowaniu.

Przykład empiryczny: stopa bezrobocia U_t w USA



Źródło: FRED

Przykład empiryczny: stopa bezrobocia U_t w USA

- Oszacowania parametrów modelu AR(1)

$$\hat{U}_t = \underset{(0.087)}{0.184} + \underset{(0.014)}{0.969} U_{t-1} \tag{20}$$

wyraźna/ekstremalna persystencja zmiennej.
Ale: autokorelacja reszt (korelacja pomiędzy resztami i pierwszymi opóźnieniami reszt ≈ 0.66).

- Oszacowania parametrów modelu AR(2)

$$\hat{U}_t = \underset{(0.066)}{0.285} + \underset{(0.045)}{1.613} U_{t-1} - \underset{(0.045)}{0.661} U_{t-2} \tag{21}$$

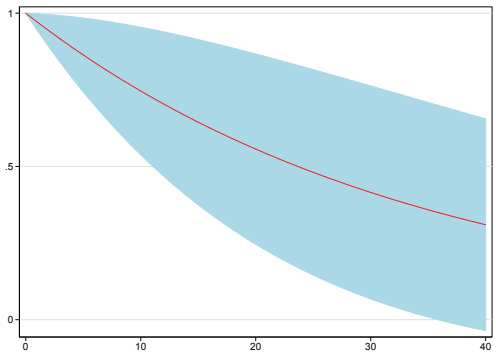
Jak w czasie wygląda reakcja stopy bezrobocia na egozgeniczny szok?

Przykład empiryczny: stopa bezrobocia U_t w USA

Funkcja reakcji na impuls (IRF):

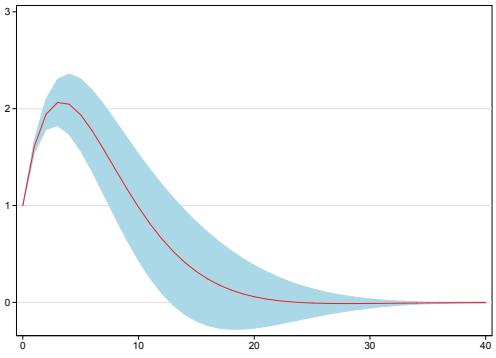
Model AR(1):

$$U_t = \underset{(0.087)}{0.184} + \underset{(0.014)}{0.969} U_{t-1}$$



Model AR(2)

$$U_t = \underset{(0.066)}{0.285} + \underset{(0.045)}{1.613} U_{t-1} - \underset{(0.045)}{0.661} U_{t-1}$$



Modele autoregresyjne z rozkładem opóźnień

Model ADL(1,0) I

- **Model autoregresyjny z rozkładem opóźnień (*autoregressive distributed lag model*) ADL(1,0)**:

$$y_t = \mu + \rho y_{t-1} + \beta_0 x_t + \varepsilon_t, \tag{22}$$

  gdy $|\rho| < 1$.
- Załóżmy, że $y_0 = 0$, $\mu = 0$ oraz $\varepsilon_t = 0$ oraz rozważmy jednostkową zmianę x w okresie 0. Wtedy,

$$\begin{aligned} y_0 &= 0 \times \rho + \beta_0 \times 1 + 0 = \beta_0, \\ y_1 &= \beta_0 \times \rho + \beta_0 \times 0 + 0 = \rho\beta_0, \\ y_2 &= \rho\beta_0 \times \rho + \beta_0 \times 0 + 0 = \rho^2\beta_0, \end{aligned}$$

  lub ogólniej

$$y_t = \rho^t \beta_0.$$

- Mnożnik krótkookresowy: β_0.
- Funkcja reakcji na impuls IRF dla i-tego okresu:

$$\frac{\partial \mathbb{E}(y_t)}{\partial x_{t-i}} = \rho^i \beta_0.$$

Model ADL(1,0) II

- Skumulowana funkcja reakcji na impuls dla i-tego okresu:

$$\sum_{j=0}^{i} \frac{\partial \mathbb{E}\left(y_t\right)}{\partial x_{t-j}} = \sum_{j=0}^{i} \rho^j \beta_0 = \beta_0 + \beta_0\rho + \beta_0\rho^2 + \ldots + \beta_0\rho^j.$$

- Mnożnik długoookresowyr:

$$\sum_{j=0}^{\infty} \frac{\partial \mathbb{E}\left(y_t\right)}{\partial x_{t-j}} = \beta_0 \left(1 + \rho + \rho^2 + \ldots\right) = \frac{\beta_0}{1-\rho}.$$

Modele ADL(P,K)

- **Model autoregresyjny z rozkładem opóźnień ADL(P,K)**:

$$y_t = \mu + \sum_{i=1}^{P} \rho_i y_{t-i} + \sum_{i=0}^{K} \beta_i x_{t-i} + \varepsilon_t. \tag{23}$$

- **Mnożnik krótkookresowy (β^{SR}):**

$$\beta^{SR} = \beta_0. \tag{24}$$

- **Mnożnik długoookresowy (β^{LR}):**

$$\beta^{LR} = \frac{\beta_0 + \beta_1 + \ldots + \beta_K}{1 - \rho_1 - \rho_2 - \ldots - \rho_P} = \frac{\sum_{i=0}^{K} \beta_i}{1 - \sum_{i=1}^{P} \alpha_i}. \tag{25}$$

Wybór specyfikacji modelu AR/DL/ADL

Trade-off pomiędzy:

- Ryzyko pominięcia ważnych opóźnień (kiedy P i/lub K są małe).
- Utrata efektywności (kiedy P i/lub K są duże).

Najpopularniejsze strategie wyboru liczby opóźnień:

- Od ogółu do szczegółu (*from general to specific*).
- Od szczegółu do ogółu (*from specific to general*).

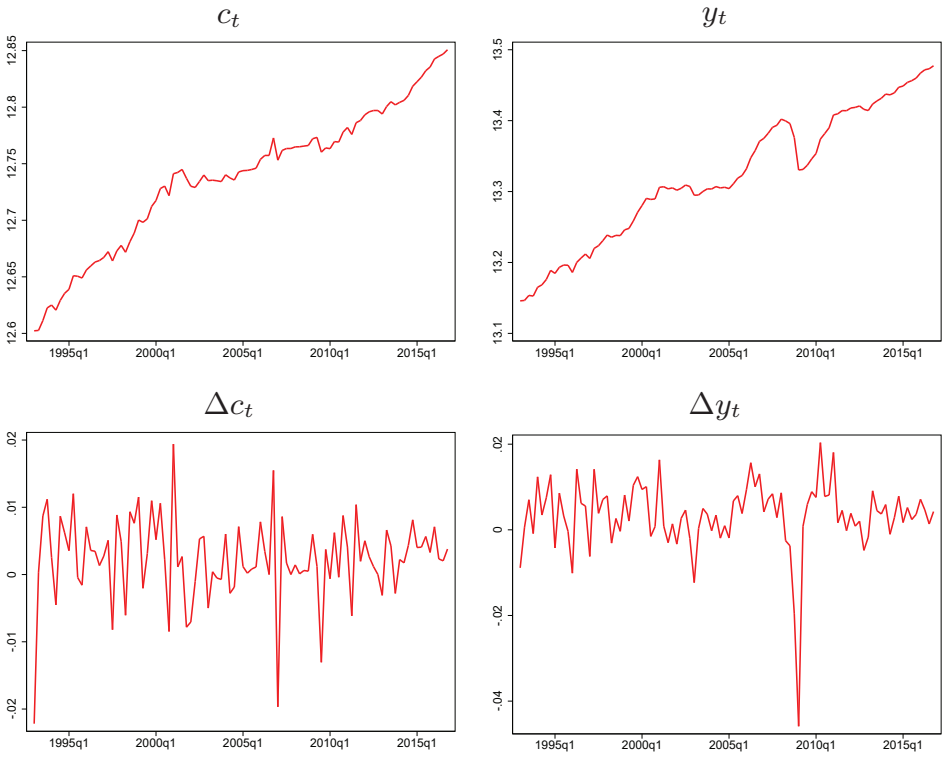
Kryteria selekcji

- Autokorelacja składnika losowego.
- Kryteria informacyjne.
- Istotność oszacowań.

Przykład empiryczny: funkcja konsumpcji w Niemczech

- Dane: szeregi czasowe w okresie od 1993Q1 do 2016Q.
- **Zmienna objaśniana:**
  c_t - logarytm naturalny wydatków konsumpcyjnych (w cenach stałych).
- **Zmienna objaśniająca:**
  y_t - logarytm naturalny realnego PKB (w cenach stałych).

Przykład empiryczny: funkcja konsumpcji w Niemczech



Przykład empiryczny: funkcja konsumpcji w Niemczech

Modele z rozkładem opóźnień DL:

$$\Delta c_t = \alpha_0 + \sum_{i=0}^{K} \beta_i \Delta y_{t-i} + \varepsilon_t, \tag{26}$$

K	0	1	2
μ	0.002	0.002	0.002
	(0.001)	(0.001)	(0.001)
β_0	0.278	0.328	0.295
	(0.092)	(0.096)	(0.087)
β_1		-0.135	-0.200
		(0.096)	(0.091)
β_2			0.142
			(0.087)
β^{LR}	0.278	0.193	0.236
	(0.092)	(0.114)	(0.120)
BIC	-705.547	-696.699	-705.430

Przykład empiryczny: funkcja konsumpcji w Niemczech

Modele z autoregresyjne rozkładem opóźnień ADL:

$$\Delta c_t = \mu + \sum_{j=1}^{P} \rho_j \Delta c_{t-j} + \sum_{i=0}^{K} \beta_i \Delta y_{t-i} + \varepsilon_t, \tag{27}$$

K	0	1	2	0
P	0	0	0	1
μ	0.002	0.002	0.002	0.002
	(0.001)	(0.001)	(0.001)	(0.001)
ρ_1				-0.290
				(0.073)
β_0	0.278	0.328	0.295	0.204
	(0.092)	(0.096)	(0.087)	(0.074)
β_1		-0.135	-0.200	
		(0.096)	(0.091)	
β_2			0.142	
			(0.087)	
β^{LR}	0.278	0.193	0.236	0.205
	(0.092)	(0.114)	(0.120)	(0.068)
BIC	-705.547	-696.699	-705.430	-707.569

Zadania

Zadanie 8.1

Rozważ model ADL(1,1):

$$y_t = \alpha + \rho_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t. \tag{28}$$

Załóż, że $|\rho_1| < 1$.

1. Oblicz analitycznie funkcję reakcji na impuls.
2. Oblicz analitycznie skumulowaną funkcję reakcji na impuls.

Zadanie 8.2

Rozważ model ADL(1,0):

$$y_t = \alpha + \rho_1 y_{t-1} + \beta_0 x_t + \varepsilon_t. \tag{29}$$

Załóż, że $\rho_1 = 1$.

1. Oblicz analitycznie funkcję reakcji na impuls.
2. Oblicz analitycznie skumulowaną funkcję reakcji na impuls.

Zadanie 8.3

Zbiór danych **fernald.gdt** zawierakwartalne szeregi czasowe opisujące stronę podażową gospodarki amerykoańskiej. Szeregi czasowe opisujące zmienne ekonomiczne zostały udostępione przez amerykańskiego ekonomistę Johna Fernalda.

1. Wykorzystanie mocy wytwórczych (*capacity utilization*, oznaczone jako $util_t$) jest zmienną ekonomiczną, która mierzy stopień wykorzystania dostępnych czynników produkcyjnych w bieżącej produkcji. Na podstawie wiedzy ekonomicznej przedyskutuj czy zmienna ta powinna charakteryzować się wysoką czy niską persystencją.
2. Oszacuj parametry modelu AR(1) dla zmiennej $util_t$. Na podstawie uzyskanych oszacowań oblicz i zinterpretuj *half-life*.
3. Przetestuj autokorelację składnika losowego w modelu AR(1).
4. Oszacuj parametry modeli AR(2) oraz AR(4) dla zmiennej $util_t$ oraz przeprowadź testy na autokorelację składnika losowego. Czy wyniki tych testów różnią się od rezultatów dla modelu AR(1).
5. Oblicz i zilustruj funkcję reakcji na impuls na podstawie oszacowań wyżej rozważanych modeli, tj. AR(1), AR(2) i AR(4). Przedyskutuj różnice i podpobieństwa.

Zadanie 8.4

Zbiór danych **HICP.gdt** zawiera szereg czasowy dla zharmonizowanego wskaźnika cen konsumpcyjnych w Polsce (oznaczany jako $HICP_t$).

1. Oblicz pierwsze różnice logarytmu naturalnego $HICP_t$, t.j.,
$$y_t = \Delta \ln HICP_t. \tag{30}$$
W jaki sposób zinterpretujesz uzyskany szereg czasowy?
2. Oszacuj parametry modelu AR(1) dla zmiennej y_t. Oblicz i zilustruj implikowaną funkcję reakcji na impuls dla zmiennej y_t. Na podstawie uzyskanych oszacowań oblicz i zinterpretuj *half-life*.
3. Przetestuj autokorelację składnika losowego w rozważanym powyżej modelu AR(1).
4. Rozważ teraz modele AR(2) oraz AR(12) dla zmiennej y_t. Oszacuj parametry tych modeli oraz zilustruj funkcje reakcji na impuls.
5. Powtórz punkty 2.-4. rozszerzając specyfikację modelu o sezonowe zmienne binarne. Spróbuj wytłumaczyć różnice w porównaniu z wcześniejszymi wynikami.
6. Oblicz zmiany rok-do-roku dla logarytmu naturalnego $HICP_t$, t.j.,
$$x_t = \ln HICP_t - \ln HICP_{t-12}. \tag{31}$$
Naszkicuj x_t i y_t oraz przedyskutuj różnice pomiędzy tymi szeregami czasowymi.
7. Powtórz punkty 2.-5. Dla zmiennej x_t zamiast zmiennej y_t. Czy są różnice w oszacowaniach? Czy te różnice są zgodne analizą wizualną z poprzedniego punktu?

Zadanie 8.5

Zbiór danych **GermanTrade.gdt** zawiera szeregi czasowe realnego eksportu w Niemczech (oznaczane jako EX_t) oraz realnego efektywnego kursu walutowego dla Niemiec (oznaczane jako $REER_t$).

1. Oszacuj parametry dla statycznego równania eksportu:

$$\Delta \ln EX_t = \mu + \beta_0 \Delta \ln REER_t + \varepsilon_t. \tag{32}$$

   Zinterpretuj oszacowania i sprawdź czy składnik losowy charakteryzuje się autokorelacją
2. Rozważ teraz model AR(1) dla $\Delta \ln EX_t$. Przedyskutuj oszacowanie parametru mierzącego persystencję.
3. Rozważ teraz model ADL(1,0) dla $\Delta \ln EX_t$. Zinterpretuj oszacowania w kategoriach mnożników krótko- i długookresowych. Przetestuj autokorelację składnika losowego. Jak uzyskane wyniki różnią się od rezultatów z pierwszego punktu. Naszkicuj funkcję reakcji na impuls dla eksportu na zmianę realnego efektywnego kursu walutowego.
4. Porównaj wcześniejsze wyniki z oszacowaniami parametrów modeli ADL(2,0), ADL(1,1), ADL(4,0) i ADL(1,4). Czy są różnice w oszacowaniach długookresowego efektu parecjacji kursu walutowego?

Zadanie 8.6 I

1. Rozważ zależność pomiędzy pierwszymi różnicami logarytmu naturalnego produktu (oznaczonego jako Δy_t) i pierwszymi różnicami logarytmu naturalnego wykorzystania czynników wytwórczych (oznaczanego jako $\Delta util_t$):

$$\Delta y_t = \mu + \beta_0 \Delta util_t + \varepsilon_t, \tag{33}$$

   Na podstawie uzyskanej dotychczas wiedzy ekonomicznej sformułuj oczekiwania odnośnie znaku parametru β_0.
2. Zbiór danych **fernald.gdt** zawiera szeregi czasowe opisujące stronę podażową gospodarki amerykańskiej. Korzystając z tego zbioru danych oszacuj parametry modelu (33). Przedyskutuj oszacowany efekt zmiany wykorzystania czynników wytwórczych na produkt.
3. Przetestuj autokorelację składnika losowego. Czy wcześniejsze osacowania są wiarygodne?
4. Rozszerz specyfikację modelu (33) do modelu z rozkładem opóźnień rzędu 4, t.j, ADL (0,4) i oszacuj kluczowe parametry. Oblicz i zinterpretuj mnożnik krótko- i długookresowy. Czy uzyskane wyniki różnią się od rezultatów z pierwszego punktu?
5. Rozważ teraz dwa dodatkowe rozszerzenia modelu (33): ADL(1,4) i ADL(4,4). Oblicz i zinterpretuj mnożnik krótko- i długookresowy, a następnie porównaj te wyniki z poprzednim punktem.

Zadanie 8.6 II

6. Rozważ kolejne rozszerzenie podstawowej specyfikacji:

$$\Delta y_t = \mu + \beta_0 \Delta util_t + \gamma_0 \Delta l_t + \varepsilon_t, \tag{34}$$

   gdzie Δl_t to pierwsze różnice logarytmu naturalnego nakładów pracy. Oszacuj parametry modelu (34) i porównaj oszacowania β_0 z rezultatami z pierwszego punktu. Spróbuj wytłumaczyć różnice.
7. Rozważ teraz kolejne rozszerzenia modelu (34) o:
   - 4 opóźnienia $\Delta util_t$;
   - 4 opóźnienia $\Delta util_t$ i jedno Δy_t,
   - 4 opóźnienia $\Delta util_t$ i Δy_t.

   i porównaj krótko- i długookresowe mnożniki z rezultatami z punktów 5 i 6.

Temat 9

Niestacjonarność i kointegracja

Jakub Mućk

- Stacjonarność
- Stopień integracji
- Test Dockeya-Fullera
- Regresja pozorna
- Kointegracja
- Model korekty błędem

Podstawowe definicje

- Szereg czasowy y_t stanowią obserwacje uporządkowane czasem t, gdzie $t = 1, 2, \ldots, T$.
- Operator opóźnień (*lag operator*) L:

$$L\left(y_t\right) = y_{t-1}. \tag{1}$$

- Operator różnicowania/pierwsze różnice (*Difference operator/first difference*) Δ:

$$\Delta\left(y_t\right) = \left(1 - L\right) y_{t-1} = y_t - y_{t-1}. \tag{2}$$

- Stopa wzrostu (dynamika) mierząca procentową zmianę zmiennej y_t:

$$g = \frac{y_t - y_{t-1}}{y_{t-1}}. \tag{3}$$

- Logarytmiczna stopa wzrostu:

$$\Delta \ln y_t = \ln y_t - \ln y_{t-1} = \ln \frac{y_t}{y_{t-1}} = \ln \frac{y_{t-1} \times (1+g)}{y_{t-1}} \approx g. \tag{4}$$

Stacjonarność

Stacjonarność

- **Stacjonarność w szerszym znaczeniu (*weak stationarity*, *wide-sense stationarity*)** występuje jeżeli spełnione są następujące warunki:
  1. stała wartość oczekiwana:
  $$\mathbb{E}(y_t) = \mu \tag{5}$$
  2. stała wariancja :
  $$\text{var}(y_t) = \sigma^2 \tag{6}$$
  3. kowariancja nie zależy od czasu t:
  $$\text{cov}(y_t, y_{t+s}) = \text{cov}(y_t, y_{t-s}) = \gamma_s \tag{7}$$
- Stacjonarne szeregi czasowe wykazują **zdolność powrotu do wartości średniej** (*mean reversion*).
- Niestacjonarne szeregi czasowe charakteryzują się **pierwiastkiem jednostkowym** (*unit root*).
- Intuicja: stacjonarne szeregi czasowe fluktuują systematycznie wokół średniej z próby

Integracja

- Stopień integracji (*integration*) to minimalna liczba różnic, która jest wymagana aby uzyskać szereg stacjonarny.
- Szeregi stacjonarne są zintegrowane w stopniu 0, t.j. $y_t \sim I(0)$.
- Jeżeli szereg czasowy jest niestacjonarny, ale jego pierwsze różnice są stacjonarne to wtedy szereg jest zintegrowany w stopniu 1, t.j., $y_t \sim I(1)$.
- Generalnie,
$$y_t \sim I(d) \iff \Delta^d y_t \sim I(0). \tag{8}$$

Przyrostostacjonarność i trendostacjonarność

- **Przyrostostacjonarność (*difference-stationary*)**:

$$y_t \sim I(1) \iff \Delta y_t \sim I(0). \tag{9}$$

- **Trendostacjonarność (*trend-stationary*)**:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \iff \varepsilon_t \sim I(0). \tag{10}$$

- Wybór pomiędzy przyrostostacjonarnością a trendostacjonarnością jest często arbitralny.
  Jednak wybór ten ma istotne implikacje dla szeregu czasowego y_t.

Przykłady procesów stacjonarnych i niestacjonarnych

Procesy stacjonarne

- Biały szum (*white noise*).
- Proces autoregresyjny (*autoregressive process*).

Procesy niestacjonarne

- Błądzenie losowe (*random walk*).

Biały szum i AR(1)

- **Biały szum (*white noise*):**

$$y_t = \varepsilon_t, \tag{11}$$

gdzie $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

- **Proces autoregresyjny pierwszego rzędu AR(1):**

$$y_t = \rho y_{t-1} + \varepsilon_t, \tag{12}$$

gdzie $|\rho| < 1$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

- Wybrane własności y_t jeżeli jest generowany przez proces AR(1):

$$\mathbb{E}(y_t) = 0, \tag{13}$$

$$Var(y_t) = \frac{\sigma^2}{1-\rho^2}. \tag{14}$$

- Zarówno biały szum, jak i proces AR(1) są stacjonarne.

Proces AR(1) z wyrazem wolnym i trendem czasowym

- Niech μ oznacza wyraz wolny. Wtedy proces AR(1) z wyrazem wolnym:

$$y_t = \mu + \rho y_{t-1} + \varepsilon_t \tag{15}$$

gdzie $|\rho| < 1$ oraz $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

- Alternatywnie równanie (15) można zapisać następująco:

$$(y_t - \mu) = \rho\,(y_{t-1} - \mu) + \varepsilon_t \tag{16}$$

- (Długookresowa) wartość oczekiwana:

$$\mathbb{E}(y_t) = \mu/(1-\rho) \tag{17}$$

- Proces AR(1) może zostać rozszerzony o trend liniowy. Wtedy

$$(y_t - \mu - \delta t) = \rho\,(y_{t-1} - \mu - \delta(t-1)) + \varepsilon_t \tag{18}$$

Błądzenie losowe (*random walk*) I

- **Błądzenie losowe (*random walk*)** jest przykładem procesu **niestacjonarnego**.

$$y_t = y_{t-1} + \varepsilon_t \tag{19}$$

  gdzie $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$
- Wartość oczekiwana (średnia) zależy od długości próby.
- Wykorzystując rekursywną substytucję (*recursive substitution*) można pokazać, że błądzenie losowe jest procesem wędrownym (*wandering*)

$$\begin{aligned} y_1 &= y_0 + \varepsilon_1 \\ y_2 &= y_1 + \varepsilon_2 = y_0 + \varepsilon_1 + \varepsilon_2 = y_0 + \sum_{k=1}^{2} \varepsilon_k \\ &\dots \\ y_t &= y_0 + \sum_{k=1}^{t} \varepsilon_k \end{aligned} \tag{20}$$

- Skumulowana suma szoków/zaburzeń losowych ($\sum_{k=1}^{t} \varepsilon_k$) jest zazwyczaj interpretowana jako **trend stochastyczny** (*stochastic trend*).

Błądzenie losowe (*random walk*) II

- $\mathbb{E}(y_t)$ zależy od wartości początkowej:

$$\mathbb{E}(y_t) = \mathbb{E}(y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t) = y_0 \tag{21}$$

- Ale wariancja zależy od czasu i nie może być ograniczona:

$$\operatorname{var}(y_t) = \operatorname{var}(\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t) = t\sigma^2 \tag{22}$$

  Stąd, założenie o stałej wariancji nie jest spełnione. Dlatego, błądzenie losowe jest procesem niestacjonarnym

Błądzenie losowe z przesunięciem i trendem

- Proces **błądzenia losowego z przesunięciem (*random walk with drift*)**.

$$y_t = \mu + y_{t-1} + \varepsilon_t \tag{23}$$

  gdzie $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.
- Wartość oczekiwana i wariancja:

$$\begin{aligned} \mathbb{E}(y_t) &= t\mu + y_0 + \mathbb{E}(\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t) = t\mu + y_0, \\ \operatorname{var}(y_t) &= \operatorname{var}(\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t) = t\sigma^2. \end{aligned} \tag{24}$$

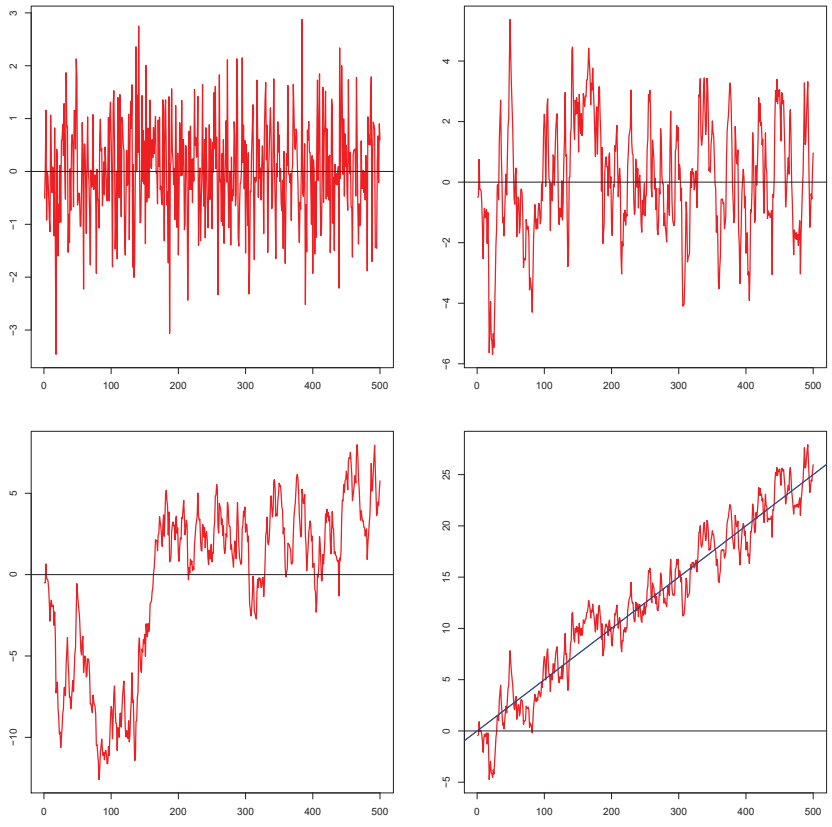
- Proces **błądzenia losowego z trendem (*random walk with trend*)**

$$y_t = \mu + \beta t + y_{t-1} + \varepsilon_t, \tag{25}$$

  where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.
- Uwzględnienie trendu deterministycznego amplikfikuje efekt trendu stochastycznego:

$$\mathbb{E}(y_t) = t\mu + \left(\frac{t^2 + t}{2}\right)\beta + y_0 t. \tag{26}$$

Przykłady: dane symulowane



Test Dickeya-Fullera

Test Dickeya-Fullera (wersja podstawowa) I

- Podstawowe założenie: y_t jest generowany przez proces AR(1):

$$y_t = \rho y_{t-1} + \varepsilon_t \tag{27}$$

- Generalna idea polega na testowaniu czy ρ jest równe jedności czy jest **statystycznie istotnie niższe od jedności**.
- **Hipoteza zerowa postuluje występowanie pierwiastka jednostkowego** (y_t jest niestacjonarne).
- Szacowanie parametru ρ w równaniu (27), obliczenie statystyki t-studenta i klasyczne (standardowe) wnioskowanie na podstawie tej statystki może prowadzić do niewiarygodnych wyników ze względu na ryzyko regresji pozornej (*spurious regression*). Dlatego, y_t jest różnicowany:

$$\Delta y_t = (\rho - 1)\, y_{t-1} + \varepsilon_t = \gamma y_{t-1} + \varepsilon_t \tag{28}$$

  gdzie $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ i $\gamma = (1 - \rho)$.
- Hipoteza zerowa (y_t jest niestacjonarny) i hipoteza alternatywna (y_t jest stacjonarny)

$$\begin{array}{lllll} \mathcal{H}_0: & \rho = 1 & \iff & \mathcal{H}_0: & \gamma = 0 \\ \mathcal{H}_1: & \rho < 1 & \iff & \mathcal{H}_1: & \gamma < 0 \end{array} \tag{29}$$

Test Dickeya-Fullera (wersja podstawowa) II

- Aby formalnie przetestować stacjonarność, szacowane są parametry równania (27) i obliczana jest statystyka t-studenta dla parametry γ. W tym przypadku, **rozkład t-studenta nie może zostać wykorzystany do wnioskowania statystycznego** ponieważ **obliczona statystyka t-studenta nie ma rozkładu t-studenta**. Dlatego wykorzystywany jest **rozkład** τ. Wartości krytyczne statystyki τ są obliczane na podstawie rozkładów uzyskanych numerycznie.
- Hipoteza zerowa może zostać odrzucona jeżeli wartość statytyki testowej τ **jest poniżej wartości krytycznej**.
- Równie testowe(27) można rozszerzyć o komponenty deterministyczne: **wyraz wolny** i **trend liniowy**. Hipoteza zerowa i alternatywna są wtedy takie same jak w podstawowej wersji.

Rozszerzony Test Dickeya-Fullera (*Augmented Dickey—Fuller test*)

- Aby uniknąć ryzyko autokorelacji składnika losowego regresja testowa może zostać rozszerzona o część autoregresyjną do rzędu P włącznie.:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{P} \alpha_i \Delta y_{t-i} + \varepsilon_t. \tag{30}$$

- Hipoteza zerowa w rozszerzonym teście Dickeya-Fullera (ADF) postuluje niestacjonarność szeregu czasowego:

$$\begin{aligned} \mathcal{H}_0 : &\quad \gamma = 0 \\ \mathcal{H}_1 : &\quad \gamma < 0 \end{aligned} \tag{31}$$

- W praktyce, uwzględnianie części autoregresyjnej jest bardzo popularne.

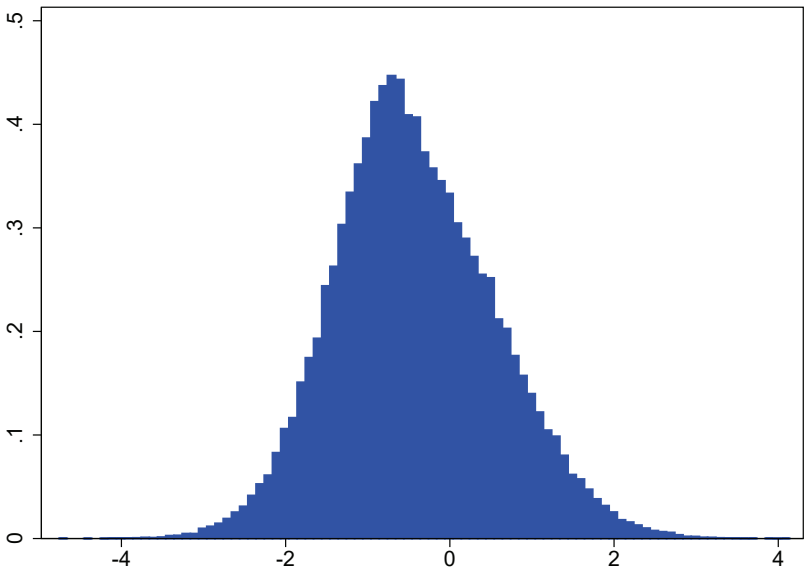
Rozkład statystyki τ testu Dickeya-Fullera

- Rozkład statystyki τ testu Dickeya-Fuller jest nieznany i przybliżany numerycznie.
- W hipotezie zerowej, szereg czasowy y_t charakteryzuje się pierwiastkiem jednostkowym, t.j., $\rho = 1$ i

$$y_t = y_{t-1} + \varepsilon_t \tag{32}$$

  gdzie $\varepsilon_t \in \mathcal{N}(0, \sigma^2)$.
- Założenie:
  - $T = 1000$;
  - $\sigma = 0.1$;
  - Liczba replikacji: 100000.

Ćwiczenie symulacyjne – rozkład statystyki τ testu Dickeya-Fullera



kwantyl	0.001	0.01	0.5	0.1	0.5
wartośc	-3.304	-2.569	-1.938	-1.614	-0.500

Wartości krytyczne dla (rozszerzonego) testu Dickeya-Fullera

- Wartości krytyczne dla (rozszerzonego) testu Dickeya-Fullera różnią się znacząco od standardowych wartości krytycznych w teście t-studenta.
- Pomiędzy dostępnym oprogramowaniem mogą pojawić się nieznaczne różnice w wartościach krytycznych

Regresja	1%	5%	10%
$\Delta y_t = \gamma y_{t-1} + \varepsilon_t$	-2.56	-1.94	-1.62
$\Delta y_t = \mu + \gamma y_{t-1} + \varepsilon_t$	-3.43	-2.86	-2.57
$\Delta y_t = \mu + \delta t + \gamma y_{t-1} + \varepsilon_t$	-3.96	-3.41	-3.13
Wartość statystyki t-studenta	-2.33	-1.65	-1.28

Uwagi: powyższe wartości krytyczne pochodzą z pracy Davidsona i MacKinnona (1993)

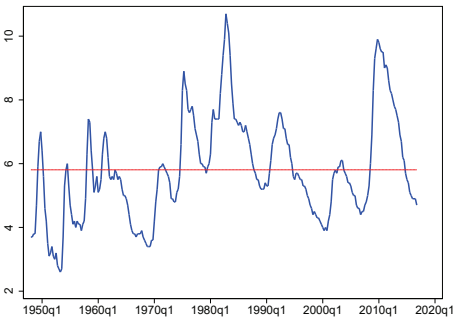
Przykład empiryczny: stopa bezrobocia U_t w USA

Regresja pomocnicza testu Dickeya-Fullera:

$$\Delta U_t = \underset{(0.185)}{0.087} - \underset{(0.0144)}{0.0311}\, U_{t-1}, \qquad (33)$$

Wartość statystyki testowej: $-0.031/0.014 \approx -2.16$
Wartość krytyczna (10% poziom istotności): -2.570
$\Longrightarrow$ nie ma podstaw do odrzucenia hipotezy zerowej $\mathcal{H}_0$

[Ale:] autokorelacja składnika losowego
Korelacja pomiędzy resztami a ich pierwszymi opóźnieniami ≈ 0.65



Regresja pomocnicza testu Dickeya-Fullera:

$$\Delta U_t = \underset{(0.066)}{0.2854} - \underset{(0.011)}{0.049}\, U_{t-1} + \underset{(0.045)}{0.662}\, \Delta U_{t-1}, \qquad (34)$$

Wartość statystyki testowej: $-0.049/0.011 \approx -4.47$
Wartość krytyczna (1% poziom istotności): -3.458
$\Longrightarrow$ są podstawy do odrzucenia $\mathcal{H}_0$ o niestacjonarności nawet na 1% poziomie istotności.

Przykład empiryczny: logarytm naturalny realnego GDP ($\ln GDP_t$)

	statystyka testu ADF	p
$\ln GDP_t$	-2.07	1
$\Delta \ln GDP_t$	-11.19	0

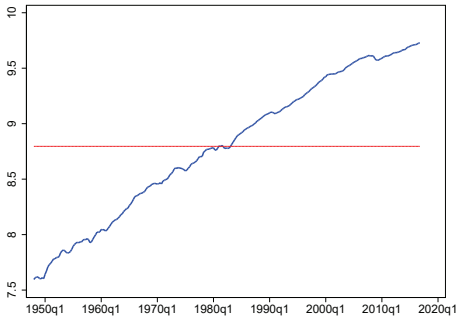
gdzie p to liczba opóźnień w teście Dickeya-Fullera.

Wartości krytyczne		
1%	5%	10%
-3.458	-2.879	-2.570

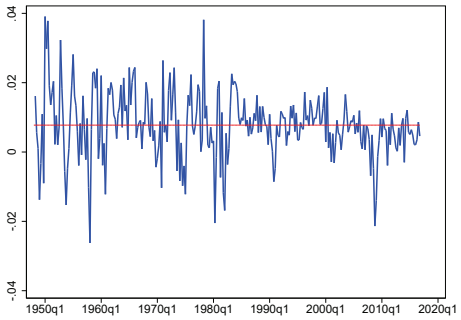
Jaki jest stopień integracji $\ln GDP_t$?

Czy zmienna $\ln GDP_t$ jest przyrostostacjonarna?

$\ln GDP_t$

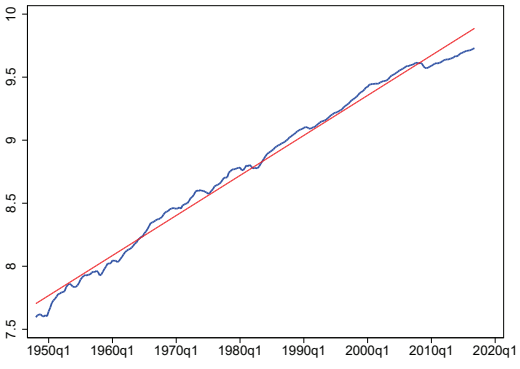


$\Delta \ln GDP_t$



Przykład empiryczny: $\ln GDP_t$ a trendostacjoarność

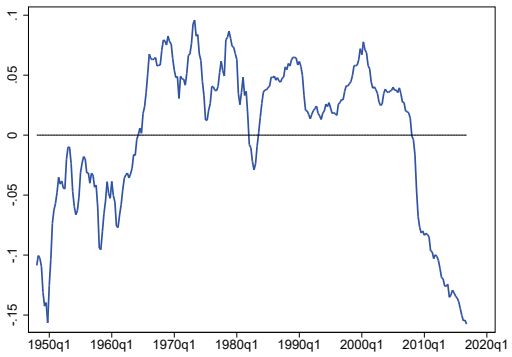
Czy $\ln GDP_t$ jest stacjarny wokół (deterministycznego) trendu?



Wartości statystyki testowej ADF: -1.232

Wartość krytyczna (10% poziom istotności): -3.130.

Reszty z regresji $\ln GDP_t$ względem trendu deteministycznego.



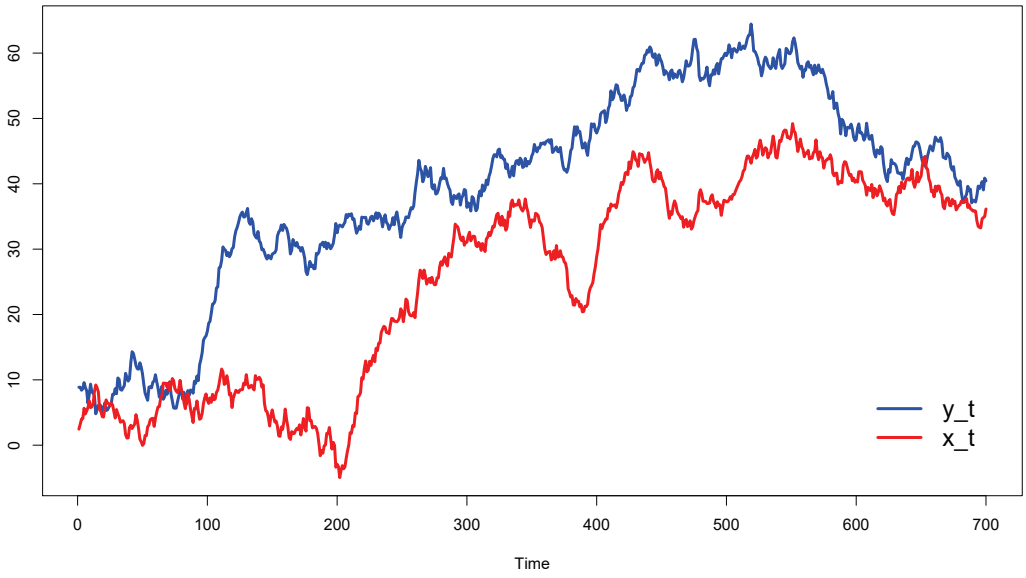
Regresja pozorna (*spurious regressions*)

- Analiza empiryczna przeprowadzona na podstawie niestacjonarnych szeregów czasowych może prowadzić do zjawiska **regresji pozornej *spurious regressions***). Zjawisko to polega na uzyskaniu **statystycznie istotnych oszacowań** na podstawie danych **pomiędzy którymi nie ma związku**.
- Aby zilustrować problem regresji pozornej warto przeanalizować wysymulowane dane (y_t i x_t):

$$\begin{array}{lrcl} \text{DGP}_1: & y_t & = & y_{t-1} + \varepsilon_t \\ \text{DGP}_2: & x_t & = & x_{t-1} + \eta_t \end{array} \tag{35}$$

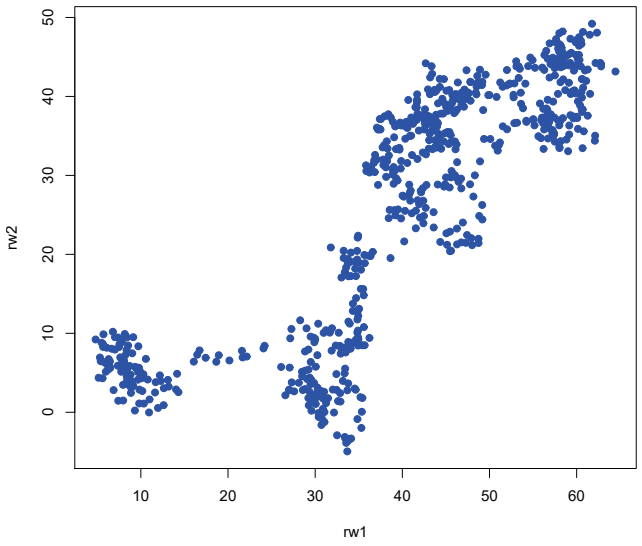
  gdzie $\eta_t \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$ I $\varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$.
- Szeregi czasowe y_t i x_t wysymulowano niezależnie, a więc nie ma zależności pomiędzy tymi zmiennymi.

Rysunek: Wysysmulowane szeregi czasowe y_t i x_t



Mimo braku prawdziwej realcji pomiędzy szeregami widoczny jest rosnący trend w obu przypadkach.

Rysunek: Scatter plot dla wysymulowanych szeregów y_t i x_t



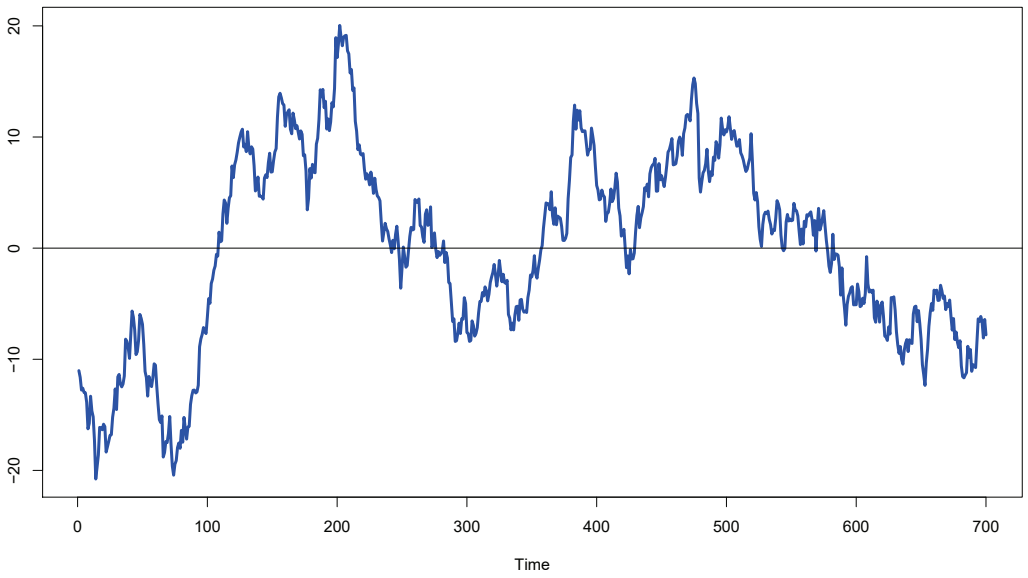
- Prosta regresja y_t względem x_t (błędy szacunku w nawiasach):

$$y_t = \underset{(0.62048)}{17.818} + \underset{(0.02062)}{0.842} x_t \tag{36}$$

- t-statistic for x_t : 40.82
- R^2 wynosi około 0.705
- Ale wiadomo, że zarówno y_t, jak i x_t uzyskano niezależenie i **nie ma między nimi prawdziwej zależności**. Dlatego, powyższe oszacowania są niewiarygodne lub **pozorne**.
- Jeżeli w regresji wykorzystywane są szeregi niestacjonarne to wtedy estymator metody najmniejszych kwadratów nie ma standardowych własności. W rezultacie, np. statystyka testu t jest niewiarygodna.
- Reszty w regresji pozornej charakteryzują się zazwyczaj wysoką autokorelacją.

Zobacz reszty

Rysunek: Reszty z regresji y_t względem x_t



Statystyka DW: 0.22
Statystyka LM (autokorelacja pierwszego rzędu): 682.958[0.0000]

Powrót do przykładu

Kointegracja

Kointegracja

- **Kointegracja (*Cointegration*)** jest szczególnym przypadkiem zależności pomiędzy zmiennymi niestacjonarnymi.
- Kluczowe założenie: y_t i x_t są **zintegrowane w stopniu pierwszym** a reszty e_t, t.j,:

$$e_t = y_t - \beta_0 - \beta_1 x_t \tag{37}$$

są **stacjonarne**. Wtedy, zmienne x_t i y_t są **skointegrowane**.

- Intuicja(1): jeżeli zmienne są skointegrowane to podążają za tym samym trendem stochastycznym.
- Intuicja(2): jeżeli zmienne są skointegrowane to wtedy istnieje długookresowa zależność (*long-run relationship*) pomiędzy nimi.

Testowanie kointegracji

- **Pierwszy krok:** testowanie stacjonarności. Jeżeli zmienne x_t i y_t są zintgrowane w stopniu pierwszym to przechodzimy do kolejnego kroku
- **Drugi krok:** szacujemy parametry relacji długookresowej (dla zmiennych niestacjonarnych),uzyskujemy reszty (e_t) i testujemy ich stacjonarność (bez wyrazu wolnego):

$$e_t = y_t - \beta_0 - \beta_1 x_t \tag{38}$$

- Hipoteza zerowa i alternatywna::

$$\begin{array}{lllll} \mathcal{H}_0: & e_t \sim \mathrm{I}(1) & \Longleftrightarrow & \mathcal{H}_0: & x_t \text{ and } y_t \textbf{ nie} \text{ są skointegrowane} \\ \mathcal{H}_1: & e_t \sim \mathrm{I}(0) & \Longleftrightarrow & \mathcal{H}_1: & x_t \text{ and } y_t \text{ są skointegrowane} \end{array} \tag{39}$$

- Jednak używamy innych wartości krytyczny niż w teście ADF ponieważ testujemy kointegrację (a nie tylko stacjonarność szeregu czasowego, tj. reszt):

Tablica: Wartości krytyczne

Długookresowa relacja	1%	5%	10%
$y_t = \beta_1 x_t + e_t$	-3.39	-2.76	-2.45
$y_t = \beta_0 + \beta_1 x_t + e_t$	-3.96	-3.37	-3.07
$y_t = \beta_0 + \delta t + \beta_1 x_t + e_t$	-3.98	-3.42	-3.13

Uwagi: wartości krytyczne pochodzą z pracy Hamiltona (1994)

Model korekty błędem (*error correction model*)

- Jeżeli x_t oraz y_t są zintegrowane i

$$e_t = y_t - \beta_0 - \beta_1 x_t \tag{40}$$

to reszty, e_t, mierzą odchylenie od długookresowej relacji

- **Długookresowa elastyczność (*long-run elasticities*)** jest równaβ_1 w (40).
- **[Model korekty błędem (*Error correction model*)]**. Model ten pozwala uwzględnić odchylenie od długookresowej równowagi w krótkookresowej dynamice analizowane zjawiska. Odchylenie od długookresowej równowagi jest mierzone opóźnionym resztami z długookresowego równania, t.j., e_{t-1}.

$$\Delta y_t = \mu + \delta e_{t-1} + \sum_{i=1}^{P} \rho_i \Delta y_{t-i} + \sum_{i=0}^{K} \beta_i \Delta x_{t-i} + \varepsilon_t, \tag{41}$$

gdzie parametr δ mierzy tempo powrotu do długookresowej równowagi $\delta \in (-1, 0)$.

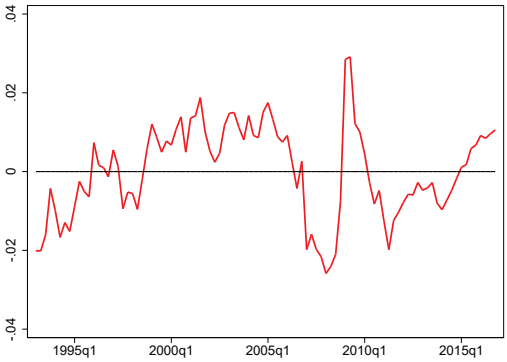
- Half-life:

$$hl = \frac{\ln(0.5)}{\ln(1+\delta)}. \tag{42}$$

Przykład empiryczny: funkcja konsumpcji w Niemczech

Zarówno c_t i y_t są zintegrowane w stopniu pierwszym.

Reszty z regresji dla zmiennych niestacjonarnych



Długookresowa relacja:

$$\hat{c}_t = \underset{(0.174)}{3.984} + \underset{(0.013)}{0.657} y_t \qquad (43)$$

Jaka jest interpretacja długookresowej elastyczności ?

Statystyka testu ADF -3.52.

Przykład empiryczny: funkcja konsumpcji w Niemczech

Model korekty błędem (specyfikacja uproszczona):

$$\Delta\hat{c}_t = \underset{(0.050)}{-0.153} ec_{t-1} + \underset{(0.066)}{0.283}\Delta y_t \qquad (44)$$

gdzie ec_{t-1} to opóźnione reszty z regresji dla zmiennych niestacjoarnych.

Oszacowanie parametru δ mierzące tempo dostosowania do długoookresowej równowagi jest statystycznie istotne i negatywne.

half-life: ≈ 4.13 kwartały ($4.13 \approx \ln(0.5)/\ln(1-0.153)$).

Zadania

Zadanie 9.1 Realny efektywny kurs polskiego złotego

Zbiór danych **REER.gdt** zwiera szereg czasowy realnego efektywnego kursu polskiego złotego deflowanego CPI (oznaczany jako $REER_t$). Aprecjacja polskiej waluty jest tożsama wzrostowi wskaźnika $REER_t$.

1. Na podstawie wiedzy ekonomicznej przedyskutuj czy $REER_t$ powinien być zmienną stacjonarną czy nie.
2. Przeanalizuj wykres logarytmu naturalnego $REER_t$. Spróbuj wytłumaczyć czy szereg ten jest stacjonarny czy nie.
3. Przeprowadź test ADF dla logarytmu naturalnego $REER_t$. Zinterpretuj uzyskany wynik.
4. Określ stopień zintegrowania logarytmu naturalnego $REER_t$.
5. Czy logarytm naturalny $REER_t$ jest trendostacjonarny?
6. Ogranicz próbę do okresu 2000:1-2019:3. Potwórz punkty 2.-4. Przedyskutuj jak zmiana próby zmieniła wnioski nt. Stacjonarności logarytmu naturalnego $REER_t$. Spróbuj wytłumaczyć źródło tych różnic.

Zadanie 9.2 Popyt na dobra importowe w Niemczech

Zbiór danych **GermanTrade.gdt** zawiera szeregi czasowe opisujące realny import w Niemczech (oznaczany jako IM_t), popyt krajowy w Niemczech (oznaczany jako DD_t) oraz realny niemiecki eksport (oznaczany jako EX_t).

1. Przetestuj kointegrację pomiędzy logarytmem naturalnym IM_t oraz logarytmem naturalnym DD_t.
2. Przetestuj kointegrację pomiędzy logarytmem naturalnym IM_t, logarytmem naturalnym DD_t oraz logarytmem naturalnym eksportu EX_t. Spróbuj wytłumaczyć różnicę we wnioskach pomiędzy obecnym a poprzednim punktem.
3. Zinterpretuj długookresowe elastyczności oszacowane w poprzednim punkcie.
4. Oszacuj parametry modelu korekty błędem wyjaśniającego pierwsze różnice logarytmu naturalnego realnego importu ($\Delta \ln IM_t$). Zinterpretuj krótkookresowe elastyczności.
5. Zinterpretuj oszacowanie przy tzw. elemencie korekty błędem, t.j., $\hat{\delta}$. Oblicz i zinterpretuj half-life.

Zadanie 9.3 Środowiskowa krzywa Kuznetsa

Zbiór danych **C0_Emission.gdt** zawiera szeregi czasowe opisujące emisję dwutlenku węgla CO^2 (oznaczane jako $CO2_t$) i PKB w Stanach Zjednoczonych (oznaczane jako GDP_t).

1. Przetestuj kointegrację między logarytmem naturalnym $CO2_t$ i logarytmem naturanym GDP_t.
2. Wykorzystaj zmienną POP_t, która mierzy populację w Stanach Zjednoczonych, aby znormalizować analizowane zmienne, t.j., $CO2_t$ i GDP_t, przez populację. Przetestuj kointegrację pomiędzy znormalizowanymi zmiennymi.
3. Zgodnie z hipotezą środowiskowej krzywej Kuznetsa (*environmental Kuznets curve hypothesis*) relacja pomiędzy jakością środowiska a poziomem rozwoju jest paraboliczna, tj. odwrotnie U-kształtna. Zaproponuj w jaki sposób ta hipoteza może zostać przetestowana w analizowanym przypadku.
4. Korzystając z metod kointegracyjnych przetestuj środowiskową krzywą Kuznetsa dla Stanów Zjednoczonych.

Zadanie 9.4 Zagregowana funkcja produkcji Cobba-Douglasa w USA

Fundamentalnym i szeroko stosowanym założeniem we współczesnej makroekonomii jest zagregowana funkcja produkcji postaci Cobba-Douglasa, którą można zapisać następująco:

$$y_t = a_t + \alpha k_t + \beta l_t, \tag{45}$$

gdzie y_t to logarytm naturalny produktu, k_t to logarytm naturalny dostępnego kapitału i l_t to logarytm naturalny nakładów pracy. Z kolei, a_t to tzw. (logarytm naturalny) łącznej produktywności czynników (*total factor productivity*,TFP) i zmienna ta jest nieobserwowalna.

1. Zbiór danych **fernald.gdt** zawiera kwartalne szeregi czasowe opisujące stronę podażową gospodarki amerykańskiej. Zbadaj rząd integracji y_t, k_t i l_t.
2. W kolejnym kroku załóż, że a_t jest stałe w czasie i przetestuj kointegrację pomiędzy porduktem a czynnikami wytwórczymi? Czy uzyskane wyniki potwierdzają zasadność wykorzystywania zagregowanej funkcji produkcji Cobba-Douglasa?
3. Zmodyfikuj założenie odnośnie a_t. Załóż teraz, że a_t zwiększa się nieprzerwanie, a tempo zmian a_t jest stałe w czasie. Korzystając z tego założenie przetestuj kointegrację pomiędzy y_t, k_t i l_t.
4. Jeśli $\alpha + \beta = 1$ to wtedy funkcja produkcji charakteryzuje się stałymi korzyściami skali (CRS, *constant returns to scale*). Czy i jak można wykorzystać standardowe metody estymacji i wnioskowania aby przetestować hipotezę CRS? Odpowiedź uzasadnij pamiętając o wynikach z poprzednich podpunktów.

Zadanie 9.5 Rozkład statystyki testu Dickeya-Fullera

Rozważ następujący process generujący dane (*data generating proces*):

$$y_t = y_{t-1} + \varepsilon_t \tag{46}$$

gdzie ε_t to składnik losowy i $\varepsilon_t \sim \mathcal{N}(0, 1)$.

1. Załóż, że $t = 1, \ldots, 100$. Korzystając z generowanie liczb losowych wysymuluj ε_t. Załóż dodatkowo, że $y_0 = 0$ i oblicz rekursywnie y_t.
2. Przeprowadź test Dickeya-Fullera dla zmiennej y_t. Czy szereg jest stacjonarny?
3. Powtórz punkty 1.-2. wielokrotnie (≥ 1000). Naszkicuj uzyskany rozkład statystyki testu Dickeya-Fullera. Czy rozkład ten jest zbliżony do rozkładu normalnego?

Temat 10

Prognoza ekonometryczna

Michał Rubaszek

- Niestabilność parametrów w modelu ekonometrycznym
- Test Chowa stabilności parametrów
- Prognoza ex-ante
- Źródła błędów prognozy ex-ante
- Miary jakości prognozy ex-post (ME, MAE, RMSE)

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Stages of the verification process

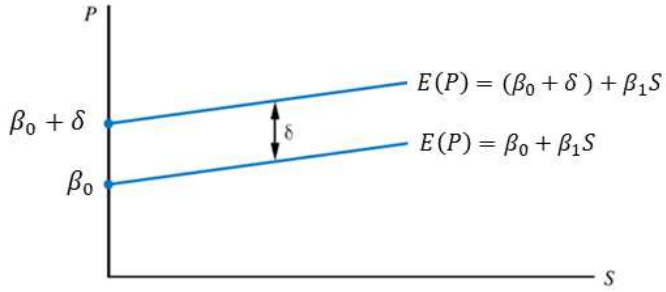
1. Interpretacja oszacowań parametrów
2. Istotność parametrów
3. Dopasowanie do danych
4. Postać funkcyjna
5. Własności składnika losowego
6. **Stabilność parametrów**

Stabilność parametrów: przesunięcie stałej

Rozważmy model, w którym cena nieruchomości (P) zależy od:

- S: powierzchni [m2]
- L: lokalizacji, zmienna binarna z wartością 1, jeżeli centrum miasta

Załóżmy, że lokalizacja zwiększa cenę o δ, co można zilustrować jak poniżej:



- Jeżeli $\delta \neq 0$, to zależność między powierzchnią S i ceną P nie jest stabilna w próbie

Stabilność parametrów: przesunięcie stałej

- Formalny zapis dla modelu:

$$P_i = \beta_0 + \beta_1 S_i + \varepsilon_i$$

- Naszym celem jest ustalenie, czy stała (β_0) jest taka sama w dwóch lokalizacjach:
  - $L = 1$, centrum miasta
  - $L = 0$, pozostałe dzielnice

- Na to pytanie możemy odpowiedzieć szacując parametry modelu:

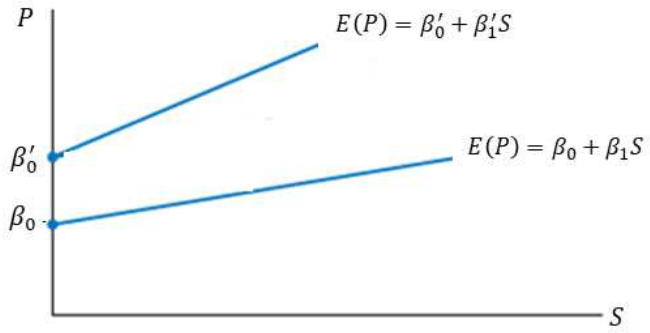
$$P_i = \beta_0 + \beta_1 S_i + \delta L_i + \varepsilon_i$$

oraz weryfikując hipotezę:

$$H_0: \delta = 0$$

Stabilność parametrów: przesunięcie stałej i nachylenia

- Skomplikujmy trochę bardziej nasze rozważania i załóżmy, że związek między ceną nieruchomości i jej powierzchnią zależy od lokalizacji w dwóch aspektach, tj. przez zmianę:
  - stałej: β_0
  - nachylenia: β_1
- Ilustracja jest następująca:



Stabilność parametrów: przesunięcie stałej i nachylenia

- Formalny zapis jest następujący:

$$P_i = \begin{cases} \beta_0 + \beta_1 S_i + \varepsilon_i & \text{dla } L_i = 0 \\ \beta_0' + \beta_1' S_i + \varepsilon_i & \text{dla } L_i = 1 \end{cases}$$

- Możemy to zapisać jako następujący model ekonometryczny:

$$P_i = \beta_0 + \beta_1 S_i + \delta L_i + \gamma (L_i \times S_i) + \varepsilon_i$$

- Zauważmy, że:

$$\beta_0' = \beta_0 + \delta \quad \text{oraz} \quad \beta_1' = \beta_1 + \gamma$$

Test na stabilność parametrów

- A zatem, dla modelu:

$$P_i = \begin{cases} \beta_0 + \beta_1 S_i + \varepsilon_i & \text{dla } L_i = 0 \\ \beta_0' + \beta_1' S_i + \varepsilon_i & \text{dla } L_i = 1 \end{cases}$$

warunkiem stabilności parametrów jest: $\beta_0 = \beta_0' \wedge \beta_1 = \beta_1'$

- Powyższą hipotezę możemy zweryfikować poprzez estymację modelu:

$$P_i = \beta_0 + \beta_1 S_i + \delta L_i + \gamma(S_i \times L_i) + \varepsilon_i$$

oraz weryfikację hipotezy:

$$H_0\colon \delta = 0 \wedge \gamma = 0$$

- Statystyka testu jest następująca (por. Temat 3)

$$F = \frac{(SSE_R - SSE_U)/2}{SSE_U/(N-4)} \sim F_{(2,N-4)}$$

Test na stabilność parametrów

Test Chowa

Rozważmy model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

Możemy sprawdzić, czy parametry modelu są stabilne, za pomocą następujących kroków:

1. Podziel próbę na dwie podpróby A i B
2. Stwórz zmienną D_i, który przyjmuje wartości 0 i 1 odpowiednio w podpróbach A i B
3. Oszacuj rozszerzony model
$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + (\mathbf{x}_i' \times D_i)\gamma + \varepsilon_i$$
4. Zweryfikuj hipotezę:
$$H_0\colon \gamma = 0$$
o statystyce:
$$F = \frac{(SSE_R - SSE_U)/(K+1)}{SSE_U/(N-2(K+1))} \sim F_{(K+1,N-2(K+1))}$$
5. Podejmij decyzję: odrzucamy H_0, jeżeli $F \geq F_c$

Ważne. Wyniki testu Chowa zależą od podziału próby z punktu 1

Test na stabilność parametrów

Test Chowa – alternatywne podejście

Rozważmy model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

Możemy sprawdzić, czy parametry modelu są stabilne w próbie, za pomocą następujących kroków:

1. Podziel próbę na dwie podpróby A i B
2. Oszacuj model na podstawie dwóch podprób ($\boldsymbol{\beta}_A$ i $\boldsymbol{\beta}_B$). Oblicz reszty e_{Ui}.
3. Oszacuj model wykorzystując wszystkie obserwacje. Oblicz reszty e_{Ri}.
4. Oblicz $SSE_U = \sum_i e_{Ui}^2$ oraz $SSE_R = \sum_i e_{Ri}^2$
5. Zweryfikuj hipotezę

   $$H_0\colon \boldsymbol{\beta}_A = \boldsymbol{\beta}_B$$

   o statystyce:

   $$F = \frac{(SSE_R - SSE_U)/(K+1)}{SSE_U/(N-2(K+1))} \sim F_{(K,N-2(K+1))}$$
6. Podejmij decyzję: odrzucamy H_0, jeżeli $F \geq F_c$

Przykład 10.1.

Test Chowa

Na podstawie danych z pliku `utown.gdt` sprawdzono, czy parametry modelu wyjaśniającego ceny nieruchomości ($price$, `1000USD`) przez powierzchnię ($sqft$, `100 sq. feet`) są takie same dla dwóch lokalizacji, blisko uniwersytetu ($utown = 1$) lub daleko od uniwersytetu.

```
Pomocnicze równanie regresji dla testu Chowa
Estymacja KMNK, wykorzystane obserwacje 1-1000
Zmienna zależna (Y): price

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       23,0625         6,23919          3,696    0,0002    ***
  sqft         7,6642         0,24660         31,08     8,69e-149 ***
  utown       28,1235         8,51058          3,305    0,0010    ***
  ut_sqft      1,2795         0,33547          3,814    0,0001    ***

Test Chowa na strukturalne różnice poziomów ze względu na zmienną: utown
  F(2, 996) = 1919,86 z wartością p 0,0000
```

Pytanie: jaka jest postać regresji dla obydwu lokalizacji?

Prognoza ekonometryczna

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. Weryfikacja
6. **Zastosowanie**

Prognoza ekonometryczna: wprowadzenie

The ultimate goal of a positive science is to develop a theory or hypothesis that yields valid and meaningful predictions about phenomena not yet observed. Theory is judged by its predictive power.

A hypothesis can't be tested by its assumptions. What is important is specifying the conditions under which the hypothesis works. What matters is its predictive power.

Milton Friedman, 1953. *The Methodology of Positive Economics*.
in *Essays in Positive Economics*: University of Chicago Press.

Czym jest prognoza ekonometryczna?

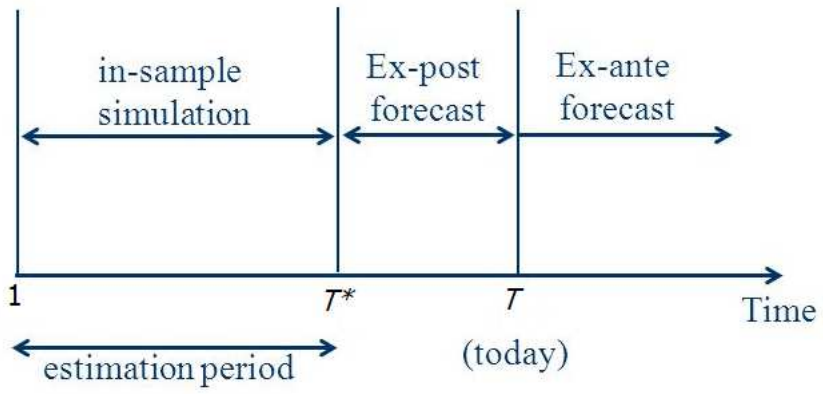
Definicja prognozy ekonometrycznej

Wnioskowanie statystyczne na temat obserwacji y_τ na podstawie modelu ekonometrycznego oszacowanego na podstawie danych z okresu $i = 1,2,\dots,N$, gdzie $\tau \notin \{1,2,\dots,N\}$.

Innymi słowy, jest to analiza dotycząca wartości zmiennej y poza próbą na podsatwie której został oszacowany model.

Prognoza ex-ante vs ex-post

- **Prognoza ex-ante:** jest prawdziwym wnioskowaniem poza próbą
  *Dotyczy obserwacji, dla których **nie znamy realizacji***
- **Prognoza ex-post:** ma na celu sprawdzenie jakości modelu
  *Dotyczy obserwacji, dla których **znamy realizację***



Dlaczego potrzebujemy prognoz?

Trafne prognozy przydatne dla:

1. **Gospodarstwa domowe**
   np. prognozy cen nieruchomości pomocne przy podejmowaniu decyzji inwestycyjnych

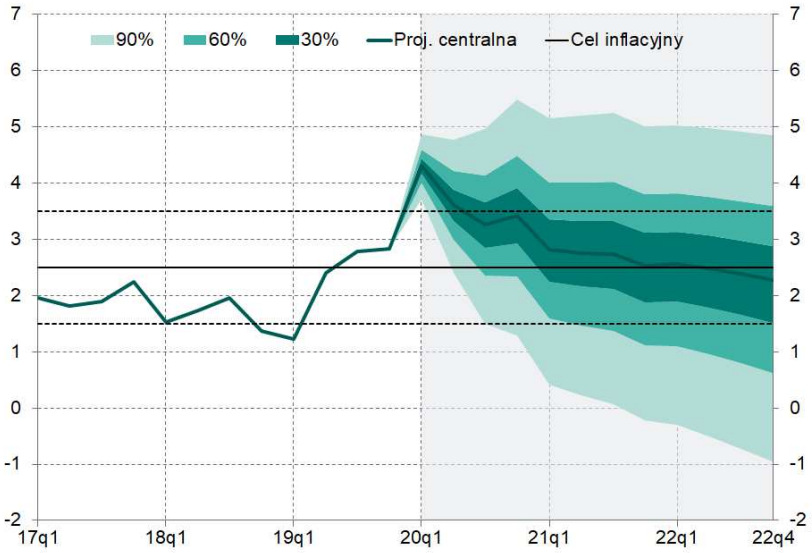
2. **Firmy**
   np. dokładne prognozy kosztów i przychodów pomagają podejmować lepsze decyzje

3. **Banki**
   np. umiejętność oszacowania prawdopodobieństwa niewykonania zobowiązania zwiększa zyski

4. **Decydenci z zakresu polityki gospodarczej**
   np. dokładne prognozy PKB, inflacji pomocne w prowadzeniu lepszej polityki monetarnej / fiskalnej

Dokładne prognozy prowadzą do lepszych decyzji!

Przykład prognozy inflacji w Polsce



Źródło: Narodowy Bank Polski, Raport o Inflacji

Prognozowanie - wprowadzenie

Rodzaje prognoz

- Ilościowa / oparta o model statystyczny
- Jakościowa / oparta o wiedzę ekspercką
- Mieszana / wykorzystująca wiedzę ekspertów oraz modele statystyczne

Ogólna charakterystyka prognoz:

- Prognozowanie opiera się na założeniu, że zależności w próbie są prawdziwe poza próbą.
  Zastanów się, czy nie wystąpiły zmiany strukturalne

- Prognozy są zawsze błędne
  Jednak niektóre metody / modele mogą dostarczać trafniejszych prognoz niż inne
  ***George Box**: All models are wrong, but some are useful*

- Prognozy szeregów czasowych są zwykle dokładniejsze dla krótszych horyzontów prognozy
  Nie należy porównywać prognoz dla różnych horyzontów

Prognoza ex-ante

Prognoza ex-ante forecast: założenia

Etapy liczenia prognozy ex-ante

1. Ustalenie celu i horyzontu prognozy
2. Zbudowanie oraz weryfikacja modelu ekonometrycznego
3. Ustalenie wartości zmiennych objaśniających w horyzoncie prognozy
4. Obliczenie prognozy punktowej
5. Obliczenie przedziału ufności dla prognozy

Warunki, które powinny być spełnione, aby móc wykorzystać model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

do obliczenia prognozy dla okresu $\tau > N$:

1. Model musi być kompleksowo i pozytywnie zweryfikowany
   (spełnienie założeń MNK, poprawna postać funkcyjna, stabilność parametrów)
2. Dysponujemy wiarygodną wartością zmiennej objaśniającej dla okresu prognozy: x_τ
3. Nie oczekujemy zmian w relacji między y i x w horyzoncie prognozy

Punktowa prognoza ex-ante

Dla modelu $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ oraz okresu prognozy $\tau > N$

- Wartość **prognozy punktowej**

$$y_\tau^P = \hat{\beta}_0 + \hat{\beta}_1 x_\tau$$

gdzie $\hat{\beta}_i$ jest oszacowaniem MNK dla parametru β_i

- Jeżeli nie znamy wartości x_τ, wtedy wzór przyjmuje postać:

$$y_\tau^P = \hat{\beta}_0 + \hat{\beta}_1 x_\tau^P$$

- Dla regresji z wieloma zmiennymi objaśniającymi $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ powyższy wzór to:

$$y_\tau^P = (\mathbf{x}_\tau^P)'\widehat{\boldsymbol{\beta}}$$

gdzie $\mathbf{x}_i = [1\ x_{1i}\ x_{2i}\ \ldots x_{Ki}]'$ oraz $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \ldots \beta_K]'$, zaś $\widehat{\boldsymbol{\beta}}$ jest oszacowaniem MNK dla $\boldsymbol{\beta}$

Przykład 10.2.

Punktowa prognoza *ex-ante*

Zbudowano model objaśniający ceny nieruchomości ($price$, 1000USD) przez powierzchnię ($sqft$, 100 sq. feet)

```
Model: Estymacja KMNK, wykorzystane obserwacje 1-1000. Zmienna zależna (Y): price

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       30,9203         9,33699          3,312    0,0010    ***
  sqft         8,59732        0,367919        23,37     1,13e-096 ***

Średn.aryt.zm.zależnej  247,6557   Odch.stand.zm.zależnej  42,19273
Suma kwadratów reszt     1149512   Błąd standardowy reszt  33,93841
```

Obliczono prognozę ceny mieszkania o powierzchni $sqft = 40$. Uzyskano wyniki:

```
Dla 95% przedziału ufności, t(998, 0,025) = 1,962
            price      prognoza  błąd ex ante   95% przedział ufności
1001                   374,813      34,3886      307,331 -   442,295
```

Wzór: $Price_{1001}^P = 30.9203 + 8.5973 \times 40 = 374{,}813$

Pytanie: jak obliczono przedział ufności?

Źródła błędu prognozy ex-ante

Błąd prognozy ex-ante

$$e_\tau^P = y_\tau - y_\tau^P$$

Źródła błędu prognozy ex-ante:

1. Błąd losowy: $\varepsilon_\tau \neq 0$
2. Błąd estymacji: $\widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$
3. Błąd zmiennych egzogenicznych: $\mathbf{x}_\tau^P \neq \mathbf{x}_\tau$
4. Błąd specyfikacji modelu
5. Zmiana strukturalna w okresie prognozy

- Jeśli znamy prawdziwy model (optymalna prognoza), nie możemy uniknąć błędu (1)
- Dla modelu jednowymiarowego możemy obliczyć błędy (1) i (2) → kolejne slajdy
- W przypadku modeli wielowymiarowych możemy również obliczyć błąd (3)
- Błędy (4) i (5) można zminimalizować, jeśli poświęcimy czas na zbudowanie dobrego modelu.

Błąd prognozy ex-ante a rozmiar modelu

Zmieniając specyfikację modelu wpływamy na błędy estymacji i specyfikacji, czyli tzw.:
variance / bias trade-off:

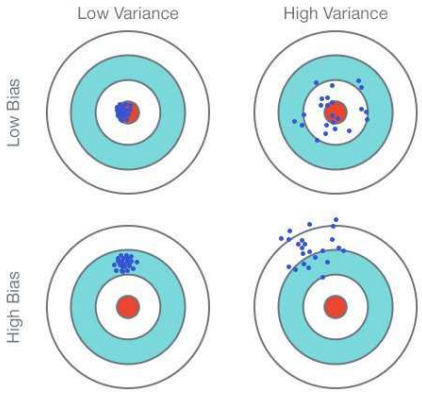
Duże / skomplikowane modele

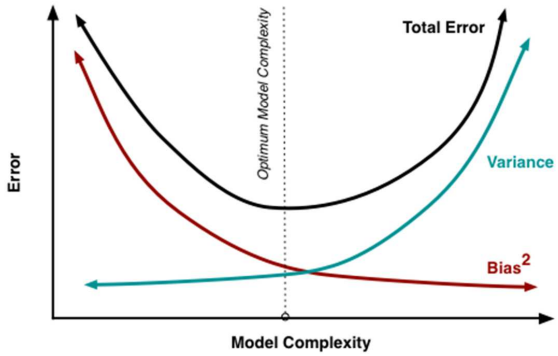
- Wiele parametrów → wysoki błąd estymacji (high variance)
- Wiele zmiennych objaśniających → dobra specyfikacja (low bias)

Małe / proste modele

- Niewiele parametrów → niski błąd estymacji (low variance)
- Niewiele zmiennych objaśniających → potencjalny błąd specyfikacji (high bias)

Który efekt dominuje? Nie wiemy i musimy to sprawdzić





Kwantyfikacja błędu prognozy ex-ante

Dekompozycja błędu prognozy na błąd losowy oraz błąd estymacji

$$e_\tau^P = y_\tau - y_\tau^P = (\mathbf{x}_\tau'\boldsymbol{\beta} + \varepsilon_\tau) - \mathbf{x}_\tau'\widehat{\boldsymbol{\beta}} = \underbrace{\varepsilon_\tau}_{\text{błąd losowy}} + \underbrace{\mathbf{x}_\tau'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}_{\text{błąd estymacji}}$$

Wzór na wariancję błędu prognozy ex-ante:

$$Var(e_\tau^P) = Var(\varepsilon_\tau) + \mathbf{x}_\tau' Var(\widehat{\boldsymbol{\beta}})\mathbf{x}_\tau + \underbrace{2Cov(\varepsilon_\tau, \mathbf{x}_\tau(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}))}_{0}$$

podstawiając oszacowania:

- $\widehat{Var(\varepsilon_\tau)} = s^2$
- $\widehat{Var(\widehat{\boldsymbol{\beta}})} = s^2(\mathbf{X}'\mathbf{X})^{-1}$

otrzymujemy :

$$Var(e_\tau^P) = s^2 + s^2\mathbf{x}_\tau'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\tau = s^2[1 + \mathbf{x}_\tau'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\tau]$$

Zauważ, że w przypadku modelu z jedną zmienną objaśniającą wzór upraszcza się do:

$$Var(e_\tau^P) = s^2\left[1 + \frac{1}{N} + \frac{(x_\tau - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

Kwantyfikacja błędu prognozy ex-ante

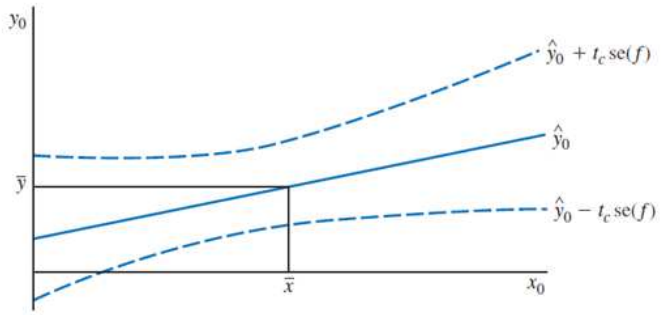
Błąd standardowy prognozy ex-ante:

$$S_\tau^P = \sqrt{Var(e_\tau^P)} = s\sqrt{1 + \mathbf{x}_\tau'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\tau}$$

oraz dla modelu z jedną zmienną objaśniającą :

$$S_\tau^P = S\sqrt{1 + \frac{1}{N} + \frac{(x_\tau - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Powyższa wartość jest najmniejsza dla $x_\tau = \bar{x}$



Prognoza przedziałowa

Jeżeli składnik losowy ma rozkład normalny, tj. $\varepsilon_t \sim N(0, \sigma^2)$, to:

$$u_\tau^P = \frac{e_\tau^P}{S_\tau^P} \sim t_{N-(K+1)}$$

Implikuje to:

$$P(-t_\alpha^* \leq u_\tau^P \leq t_\alpha^*) = 1 - \alpha$$

zaś po podstawieniu $e_\tau^P = y_\tau - y_\tau^P$ otrzymujemy:

$$P(y_\tau^P - t_\alpha^* S_\tau^P \leq y_\tau \leq y_\tau^P - t_\alpha^* S_\tau^P) = 1 - \alpha$$

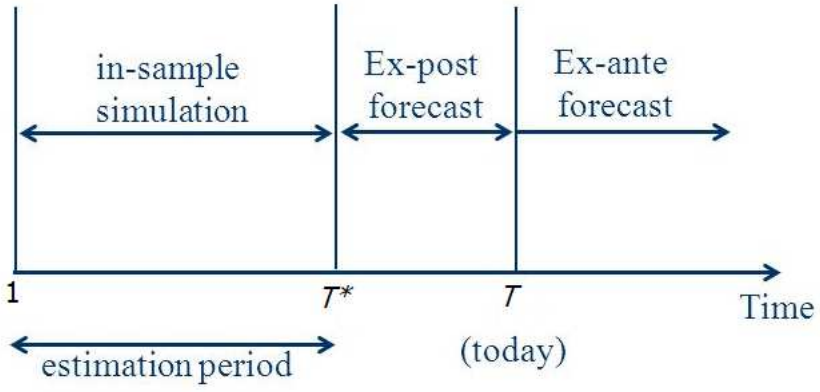
Sprawdź powyższy wzór dla przykładu 10.2., tj. modelu cen nieruchomości:

```
Dla 95% przedziału ufności, t(998, 0,025) = 1,962
            price      prognoza  błąd ex ante   95% przedział ufności
1001                   374,813      34,3886      307,331 -   442,295
```

Prognoza ex-post

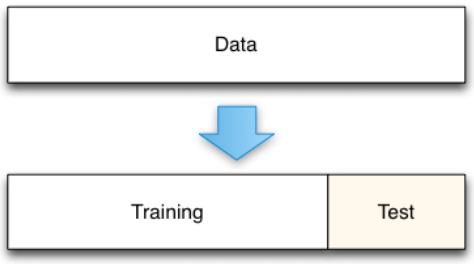
Prognoza ex-post

- Zwykle pracujemy z modelami, które dobrze działały w przeszłości
- Przy ewaluacji prognozy ex post zadajemy pytanie:
  jak dobre byłby prognozy z modelu, gdyby był używany w przeszłości
- Innymi słowy, porównujemy prognozy y_τ^P do realizacji y_τ
- Robimy to, aby móc ocenić jakość modelu

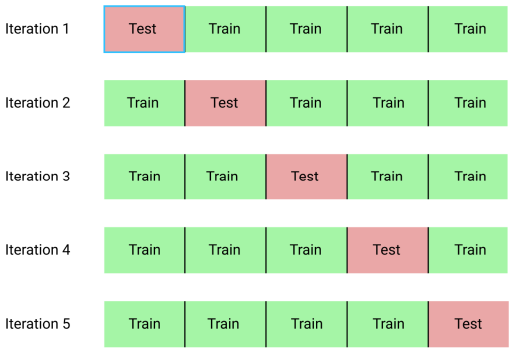


Prognoza ex-post

- Dla modeli opartych o dane przekrojowe podział próby na obserwacje wykorzystywane do estymacji parametrów (training sample) oraz weryfikacji własności prognostycznych (testing sample) jest arbitralny

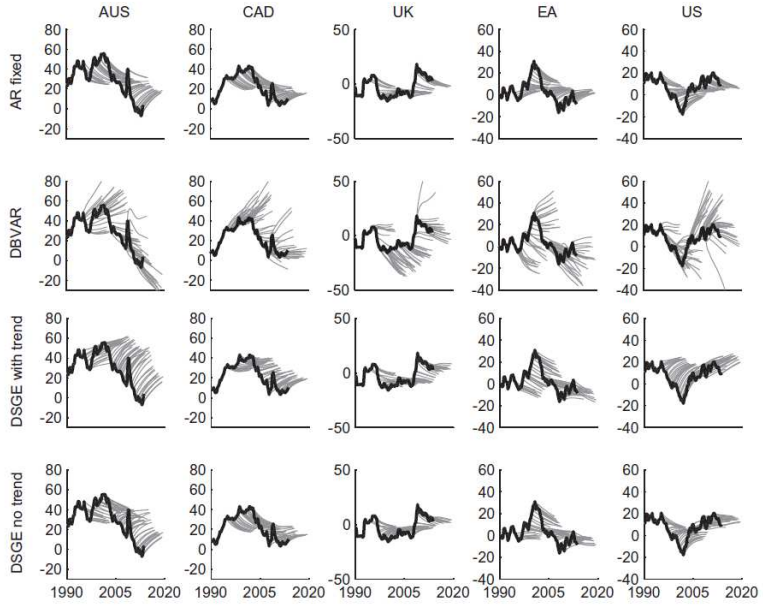


- Z tego powodu często stosowana jest walidacja krzyżowa (k-fold cross validation). W tym podejściu estymujemy model k razy, gdzie każda obserwacja jest $k-1$ razy wykorzystywana do estymacji modelu, zaś jednokrotnie przy weryfikacji prognoz



Prognoza ex-post: ilustracja

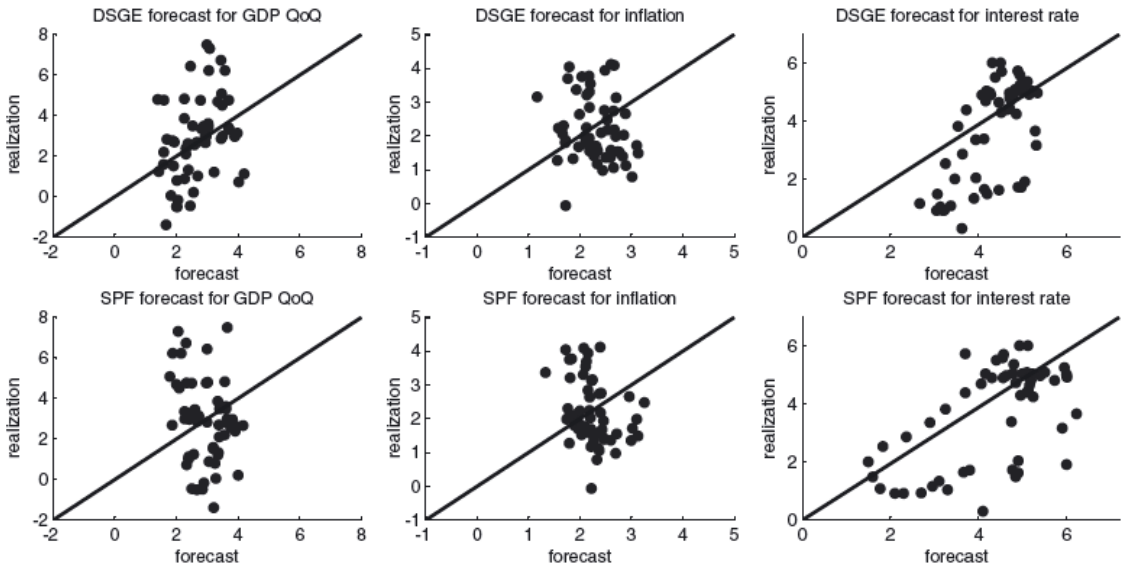
Fig. 3. Sequential real exchange rate forecasts.



Źródło: Ca' Zorzi M. & Kolasa M. & Rubaszek M., 2017. Exchange rate forecasting with DSGE models, Journal of International Economics

Prognoza ex-post: ilustracja

FIG. 2. Actuals and Four-Quarter-Ahead Forecasts.



Źródło: M. Kolasa & M. Rubaszek & P. Skrzypczyński, 2012. Putting the New Keynesian DSGE Model to the Real-Time Forecasting Test, Journal of Money, Credit and Banking

Miary jakości prognozy ex-post

Średni błąd (ME - Mean Error):

$$ME = \frac{1}{M}\sum_{\tau=N+1}^{N+M}(y_\tau - y_\tau^P)$$

Średni błąd absolutny (MAE - Mean Absolute Error):

$$MAE = \frac{1}{M}\sum_{\tau=N+1}^{N+M}|y_\tau - y_\tau^P|$$

Pierwiastek błędu średnionkwadratowego (RMSFE - Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{M}\sum_{\tau=N+1}^{N+M}(y_\tau - y_\tau^P)^2}$$

Średni absolutny błąd procentowy (MAPE - Mean Absolute Percentage Error):

$$MPAE = \frac{1}{M}\sum_{\tau=N+1}^{N+M}\left|\frac{y_\tau - y_\tau^P}{y_\tau}\right|$$

Miary jakości prognozy ex-post

Dekompozycja błędu średniokwadratowego:

$$\frac{1}{M}\sum_{\tau=N+1}^{N+M}(y_\tau - y_\tau^P)^2 = U_1^2 + U_2^2 + U_3^2$$

Obciążenie (bias):

$$U_1^2 = (\bar{y} - \bar{y}^P)^2$$

Elastyczność (variance):

$$U_2^2 = (S_R - S_P)^2$$

Niezgodność kierunku (correlation):

$$U_3^2 = 2S_RS_P(1-r)$$

gdzie $S_R = \sqrt{M^{-1}\sum(y_\tau - \bar{y})^2}$, $S_P = \sqrt{M^{-1}\sum(y_\tau^P - \bar{y}^P)^2}$ oraz $r = cor(y_\tau, y_\tau^P)$.

Współczynnik U Theila:

$$U = \sqrt{\frac{\sum(y_\tau - y_\tau^P)^2}{\sum y_\tau^2}}$$

Zauważ, że:

$$U^2 = \frac{M(U_1^2 + U_2^2 + U_3^2)}{\sum y_\tau^2}$$

Przykład 10.3. Jakość prognozy ex-post

Korzystając z pliku `utown.gdt` oszacowano parametry modelu $price_i = \beta_0 + \beta_1 sqft_i + \varepsilon_i$. Do estymacji parametrów wykorzystano 900 obserwacji (training sample), zaś wartości prognoz obliczono dla obserwacji 901-1000 (testing sample). Uzyskano następujące statystyki:

```
Miary dokładności prognoz ex post

  Średni błąd predykcji              ME =    33,016
  Pierwiastek błędu średniokwadr.  RMSE =    36,354
  Średni błąd absolutny             MAE =    33,308
  Średni błąd procentowy            MPE =    11,776
  Średni absolutny błąd procentowy MAPE =    11,898
  Współczynnik Theila (w procentach)  I =    0,81733
  Udział obciążoności predykc. I1^2/MSE =    0,82478
  Udział niedost.elastyczności I2^2/MSE =    0,00083237
  Udział niezgodności kierunku I3^2/MSE =    0,17439
```

Podaj interpretację wyników

Zadania

Zadanie 10.1

Na podstawie 5 rocznych obserwacji dla wydatków pewnego gospodarstwa domowego [tys PLN]: $y = [10\ 8\ 10\ 16\ 26]'$ oszacuj parametry modelu trendu $y_t = \beta_0 + \beta_1 t + \varepsilon_t$. Następnie, dla okresów prognozy $\tau = 6$, $\tau = 8$ i $\tau = 10$ oblicz:

a. Punktową prognozę ex-ante
b. Błąd prognozy ex-ante
c. 95% oraz 90% przedział ufności dla prognozy

Wartości rozkładu t-Studenta są następujące: $t_{3,5\%} = 3.18$ oraz $t_{3,10\%} = 2.35$

Dodatkowe informacje [spróbuj je obliczyć samemu, wykorzystując wiedzę z poprzednich spotkań]:

$$\hat{y}_t = 2 + 4t \qquad S^2 = \frac{52}{3} \qquad (X'X)^{-1} = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{bmatrix}$$

Zadanie 10.2

Na podstawie obserwacji dotyczących wydatków pięciu gospodarstw domowych, dane w tys PLN, $y = [8\ 3\ 5\ 4\ 10]'$ oraz ich dochodów, $x = [14\ 2\ 6\ 8\ 10]'$, oszacuj parametry modelu:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Jaka jest prognoza wydatków dla kojennych gospodarstw ($\tau = 6, 7, 8$), wiedząc, że $x_6 = 8$, $x_7 = 15$ oraz $x_8 = 10$. Oblicz:

a. Punktową prognozę ex-ante
b. Błąd prognozy ex-ante
c. 95% oraz 90% przedział ufności dla prognozy

Wartości rozkładu t-Studenta są następujące: $t_{3,5\%} = 3.18$ oraz $t_{3,10\%} = 2.35$

Dodatkowe informacje [spróbuj je obliczyć samemu, wykorzystując wiedzę z poprzednich spotkań]:

$$\hat{y}_t = 2 + 0.5x_t \qquad S^2 = \frac{14}{3} \qquad (X'X)^{-1} = \begin{bmatrix} 1.0 & -0.1 \\ -0.1 & 1/80 \end{bmatrix}$$

Zadanie 10.3

Korzystając z danych zawartych w pliku `utown.gdt` oszacuj parametry modelu

$$price_i = \beta_0 + \beta_1 sqft_i + \varepsilon_i$$

a. Użyj testu Chowa, aby sprawdzić, czy zależność między ceną i powierzchnią zależy od tego, czy dom ma basen (zmienna $pool$) lub kominek (zmienna $fireplace$)

b. Zinterpretuj wyniki regresji pomocniczej testu Chowa.
Czy spodziewałeś się zmiany stałej, nachylenia, czy obydwu parametrów?

c. Powtórz punkt a. dla logarytmicznej specyfikacji modelu

$$\log(price_i) = \beta_0 + \beta_1 \log(sqft_i) + \varepsilon_i$$

d. Sprawdź występowanie zmiany strukturalnej modelu z punktu c. wykorzystując lokalizację ($utown$) jako zmienną, która dzieli próbę na dwie części. Jak interpretujemy przesunięcie stałej i nachylenia w tym przypadku?

Zadanie 10.4

Korzystając z danych zawartych w pliku `PhillipsCurve.gdt` wybierz kraj oraz oszacuj model, w którym inflacja (π) zależy od stopy bezrobocia (u):

$$\pi_t = \beta_0 + \beta_1 u_t + \epsilon_i$$

a. Użyj testu Chowa, aby sprawdzić, czy zależność jest stabilna w czasie

b. Oblicz prognozę ex-ante (punktową i przedziałową) dla $T+1$ przy założeniu, że $u_{T+1}^P = u_T$

c. Zbuduj model dynamiczny (ADL) dla relacji między π_t and u_t.

d. Użyj modelu z punktu c., aby obliczyć prognozę dla okresu $T+1$

e. Porównaj wyniki z punktów a. i e.. Która prognoza jest bardziej precyzyjna?

Zadanie 10.5

Korzystając z danych zawartych w pliku `CPS5.gdt` oszacuj dwa modele:

M1: $wage_i = \beta_0 + \beta_1 exper_i + \varepsilon_i$

M2: $\ln(wage_i) = \beta_0 + \beta_1 exper_i + \varepsilon_i$

a. Użyj testu Chowa, aby sprawdzić, czy zależność między płacą i doświadczeniem zależy od tego, czy pracownik jest kobietą (*female*) oraz czy należy do związku zawodowego (*union*)
b. Korzystając z obu modeli, oblicz prognozę ex ante (punktową i przedziałową) dla osoby z 16 i 20-letnim doświadczeniem.
c. Oblicz taką samą prognozę wykorzystując oddzielny model dla mężczyzn i kobiet. Która prognoza jest bardziej dokładna?

Zadanie 10.6

Dla regresji $\hat{y}_i = 2 + 1.0x_i$ oszacowanej na podstawie 5 obserwacji. Wiemy, że w próbie ewaluacji prognozy ex-post (testing sample) realizacje wynoszą:

τ	6	7	8	9	10
x_τ	6	4	5	6	8
y_τ	10	10	6	10	8

a. Oblicz ME / MAE / RMSE / MAPE
b. Zinterpretuj wyniki z punktu a.
c. Oblicz współczynnik $U - Theil$ i jego składowe U_1^2, U_2^2 i U_3^2.
d. Jakie jest główne źródło błędu średniokwadratowego (MSE)

Zadanie 10.7

Korzystając z danych zawartych w pliku `CPS5.gdt` oszacuj model $wage_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$ dla podpróby mężczyzn (variable $female$=0). Oszacowania, które powinny się pojawić są następujące:

```
Model: Estymacja KMNK, wykorzystane obserwacje 1-5424
Zmienna zależna (Y): wage

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       −9,50978       0,907687         −10,48    1,93e-025 ***
  educ         2,44739       0,0635463         38,51    4,49e-287 ***
```

a. Na podsatwie pozostałych obserwacji oblicz miary dokładności prognozy ex post: ME/MAE/RMSE/MAPE
b. Zinterpretuj wyniki dla ME. Czy są zgodne z oczekiwaniami?
c. Zdekomponuj MSE na U_1^2, U_2^2 and U_3^2. Jakie jest główne źródło błędu prognozy?
d. Powtórz tego rodzaju analizę dla zmiennych: *married*, *asian*, *black*, *union*

Uwaga: aby wykonać tę analizę, musisz wykonać dwa kroki w Gretl:

- Dane → Sortowanie danych przekrojowych → female
- Próba → Zakres próby → 1:5424

Temat 11

Modele zmiennej jakościowej

Zuzanna Wośko i Bartłomiej Wiśnicki

- Rodzaje zmiennych jakościowych
- Liniowy model prawdopodobieństwa
- Modele logitowe i probitowe
- Efekty krańcowe w modelu logitowym
- Tablica trafności
- Pseudo-R^2 i zliczeniowy R^2
- Krzywa ROC i miara AUC

Ogólnie o zmiennych jakościowych

Jak zmienne jakościowe pojawiają się w modelach (po co?)

Jako…

- Jakościowy regresor po prawej stronie równania:
  na przykład płeć (1-kobieta, 0-mężczyzna), status spłaty kredytu (1-terminowo, 0-przeterminowany)
- Jakościowa zmienna zależna (modele dyskretnego wyboru):
  na przykład modelowanie defaultów firm, 1-default firmy, 0-brak defaultu
- Zmienna sezonowa:
  na przykład 1 – dla danego kwartału roku, 0 - dla kwartałów pozostałych
- Jakościowy regresor po prawej stronie równania objaśniający nadzwyczajne wydarzenia / obserwacje nietypowe / zmienne interakcyjne (zmiana współczynników nachylenia)

Zmienna sezonowa (dane kwartalne):

$$S3_t^{\,T} = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad \dots]$$

Trzeci kwartał każdego roku

Zmienna opisująca zdarzenie nadzwyczajne:

$$Z_t^{\,T} = [0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad \dots]$$

Zdarzenie nadzwyczajne

Rodzaje zmiennych jakościowych

- Binarne (dychotomiczne):
  na przykład płeć (kobieta=1, mężczyzna=0)
- Wielomianowe:
  sposób dojazdu do pracy –> subway=1, bus=2, tram=3, bike=4, taxi=5
- Wielomianowe uporządkowane (uporządkowanego wyboru):
  credit ratings (S&P) –> AAA = 1, AA+ = 2, AA = 3, etc.
- Licznikowe (count data):
  liczba wizyt u lekarza grupy wybranych osób – 23 wizyty, 12 wizyt etc.

Liniowy Model Prawdopodobieństwa (LMP)

Modelowanie zmiennej jakościowej (binarnej)

Załóżmy następującą regresję:

$$y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i$$

gdzie y_i^* zdefiniowano jako:

$$y_i^* = \begin{cases} 1 & \text{jeżeli } y_i > 0, \\ 0 & wpp \end{cases}$$

y_i^* jest zmienną nieobserwowalną (latent variable). To co obserwujemy, to tylko binarna zmienna y_i.

y_i^* możemy utożsamiać ze „skłonnością" jednostki *i* do $y_i = 1$.

Specyfikacja LMP

Jeśli nie przejmujemy się tym, że zmienna objaśniana jest zmienną binarną, specyfikacja modelu wygląda jak w przykładzie poniżej:

$$y_i = \alpha + \beta_0 x_{1i} + \beta_1 x_{2i} + \beta_2 x_{3i} + e_t \qquad t=1, \dots, N$$

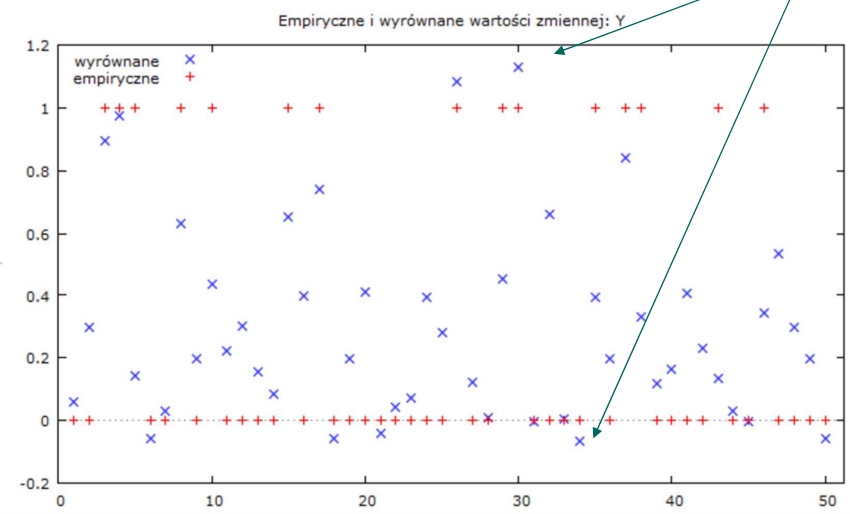
y_i - zmienna jakościowa – czy student mieszka z rodzicami (y=1) czy nie (y=0)

x_{1i} – rok studiów (od 1 do 5)
x_{2i} – dochód rodziny studenta
x_{3i} – płeć studenta (1-kobieta, 0-mężczyzna)

Niestety, kiedy używamy MNK wówczas oszacowane wartości *y* mogą znaleźć się nie tylko w przedziale [0, 1] ale również poza tym przedziałem.
Ponadto, pojawia się często heteroskedastyczność składnika losowego (lepiej użyć wtedy UMNK zamiast KMNK).
Interpretacja parametrów:
Zmiana prawdopodobieństwa p_i na skutek jednostkowej zmiany x_j *(ceteris paribus)*.

Specyfikacja LMP

Oszacowane wartości y są poza przedziałem [0, 1]



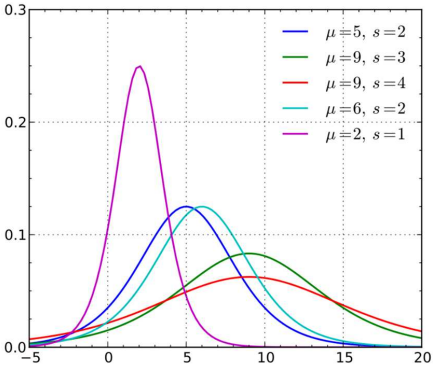
Modele logitowe i probitowe

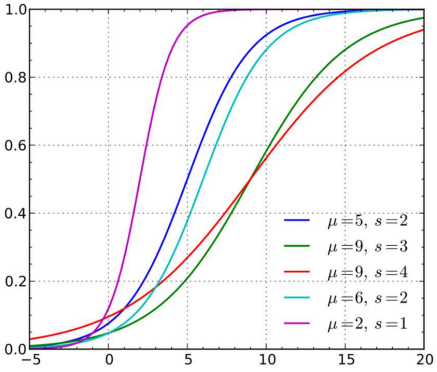
Model logitowy/probitowy

Problem z LMP: szacowane wartości zmiennych objaśniających mogą znajdować się poza przedziałem [0, 1]

Idea: Transformacja zmiennej zależnej. Dwie popularne opcje:
- Dystrybuanta rozkładu logistycznego: model logitowy
- Dystrybuanta rozkładu normalnego: model probitowy

Poniżej funkcja gęstości rozkładu logistycznego (wykres lewy) oraz dystrybuanty rozkładu logistycznego (wykres prawy).





Transformacja logitowa

Dystrybuanta rozkładu logistycznego:

$$F_{\text{logistic}}[Z_i] = \frac{e^{Z_i}}{1+e^{Z_i}} \qquad Z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

$$F_{logistic}\left[\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right] = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}} \qquad Z_i = ln\frac{F(Z_i)}{1-F(Z_i)}$$

$$ln\frac{P_i}{1-P_i} = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} \rightarrow \text{Model logitowy}$$

Iloraz szans (odds ratio)

logit

$$F^{-1}[P_i] = ln\frac{P_i}{1-P_i}$$

Dokonaliśmy monotonicznej transformacji prawdopodobieństwa z przedziału (0; 1) do przedziału (-∞; + ∞) poprzez transformację logistyczną.

Transformacja probitowa

Dystrybuanta rozkładu normalnego:

$$F_{normal}[Z_i] = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{s^2}{2}\right) ds$$

gdzie $Z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$

Jest zmienną losową o standaryzowanym rozkładzie normalnym

$$F_{normal}^{-1}[P_i] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Estymacja Metodą Największej Wiarygodności

Szukamy wektora parametrów beta, który daje największe prawdopodobieństwo uzyskania empirycznych wartości *y*. Dlatego budujemy funkcję wiarygodności, która jest prawdopodobieństwem uzyskania konkretnego wyniku dla obserwacji zmiennej *y* (czyli iloczynem prawdopodobieństwa uzyskania wyniku = 1 dla konkretnej obserwacji (obserwacje są niezależne)

$$y_i \in \{0,1\} \Rightarrow P(y_i = 1) = p_i \qquad P(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$L = \Pi_{i=1}^{n} P(y_i) = \Pi_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Celem jest znalezienie parametrów modelu (β_j), które maksymalizymalizują funkcję wiarygodności. Ze względów praktycznych logarytmizujemy funkcję L

$\ln(L) = \sum_{i=1}^{n} \ln(p_i) + (1-y_i)\ln(1-p_i)$ *gdzie* $p_i = \frac{e^{z_i}}{1+e^{Z\_i}}$

$$\ln(L) = \sum_{i=1}^{n} y_i Z_i - \ln(1+e^{Z_i})$$

$$Z_i = \beta_0 + \sum_{j=1} \beta_j x_{ij}$$

Specyfikacja modelu logitowego/probitowego

$$\ln(L) = \sum_{i=1}^{n} y_i Z_i - \ln(1 + e^{Z_i}) \qquad Z_i = \beta_0 + \sum\nolimits_{j=1} \beta_j x_{ij}$$

Funkcja *L* osiąga maksimum gdy dla każdego *j = 0, 1, 2, …, K*

$$\frac{\partial lnL}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i - \frac{e^{Z_i}}{1 + e^{Z_i}} \right] x_{ij} = \sum_{i=1}^{n} [y_i - p_i] x_{ij} = 0$$

Suma wartości odchyleń zmiennej zależnej od p_i, czyli prawdopodobieństwa że y_i wyniesie 1 jest równa zeru.

Rozwiązujemy *k+1* równań celem znalezienia wartości β_j.
Równania nie są liniowe: stosujemy metody numeryczne

Estymacja Metodą Największej Wiarygodności

Szukamy wektora parametrów beta, który daje największe prawdopodobieństwo uzyskania empirycznych wartości *y*. Dlatego budujemy funkcję wiarygodności, która jest funkcją gęstości wektora składnika losowego ***e***.

Jest ona równa iloczynowi funkcji gęstości wszystkich $e_i \sim N(0,\sigma^2)$.
Obserwacje są niezależne, dlatego łączne prawdopodobieństwo jest iloczynem prawdopodobieństw pojedynczych obserwacji.

Zmienne losowe są nieobserwowalne. Ale wiemy, że zmienna zależna jest funkcją *e*,
Zatem jest również losowa oraz związek między tymi funkcjami gęstości może być zaprezentowany jako:

$$f(y_i) = \left| \frac{de_i}{dy_i} \right| f(e_i) = f(e_i) \quad =1 \qquad y_i \sim IN(\boldsymbol{x}\beta, \sigma^2 \boldsymbol{I})$$

Funkcja wiarygodności (łączna funkcja gęstości obserwacji) ma następującą formułę:

$$\ell_y = f(y_1, y_2, \ldots, y_n) = \prod\nolimits_{i=1}^{N} f(y_i)$$

$$L = ln\, \ell_y = \sum_{i=1}^{n} ln\, f\,(y_i)$$

Ale maksymalizujemy logarytmy, co nie zmienia rezultatów optymalizacji

Specyfikacja modelu logitowego/probitowego

$$y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i$$

To co możemy prognozować w przypadku *y*, to prawdopodobieństwo *y* równego 1 dla danej *i*-tej obserwacji:

$$P_i = P(y_i = 1) = P\left[e_i > -(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})\right] = 1 - F\left[-(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})\right]$$

Jeżeli rozkład *e* jest symetryczny, wówczas *1-F(-Z)=F(Z),* zatem prawdopodobieństwo „1" dla *i*-tej obserwacji jest równe:

$$P_i = F\left[\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right]$$

Przykład 11.1

(Cieślak, „Prognozowanie gospodarcze", *przykład 4.5, s. 138)*

W miesiącach kwiecień-czerwiec przeprowadzono w Kanadzie badanie 5315 kobiet, które do momentu badania przez 60 miesięcy były mężatkami. Badanie dotyczyło bezdzietności. Zmienna Y była równa 1 gdy kobieta była bezdzietna, Y = 0 gdy kobieta miała jedno lub więcej dzieci.

Jako zmienne objaśniające wybrano:

X1 – miejsce zamieszkania (x_{1j} = 1 dla miasta and x_{1j} = 0 dla wsi),

X2 – poziom wykształcenia (x_{2j} = 1, gdy kobieta skończyła co najwyżej szkołę średnią, i x_{2j} = 0 w innym wypadku),

X3 - religijność (x_{3j} = 1, gdy kobieta regularnie uczęszcza do kościoła, i x_{3j} = 0 w innym wypadku),

X4 - wiek (x_{4j} = 1, gdy kobieta miała 25-34 lat, i x_{4j} = 0 w innym wypadku).

Oszacowany model ma postać:

$$ln\frac{P_i}{1-P_i} = 0{,}0001 + 0{,}021x_{1j} - 0{,}034x_{2j} - 0{,}020x_{3j} + 0{,}006x_{4j}.$$

Parametry przy zmiennych 1-3 były statystycznie istotne; parametr przy zmiennej x_4 był nieistotny. Zinterpretuj rezultaty.

Przykład 11.1

(Cieślak, „Prognozowanie gospodarcze”, *przykład 4.5, s. 138)*

$$ln\frac{P_i}{1-P_i} = 0{,}0001 + 0{,}021x_{1j} - 0{,}034x_{2j} - 0{,}020x_{3j} + 0{,}006x_{4j}.$$

Odpowiedź:

- Znaki ocen parametrów oznaczają kierunek zależności miedzy daną cechą a ryzykiem względnym. **Nie należy więc ich interpretować w sposób ilościowy!**
- **Iloraz szans to tzw. ryzyko względne,** czyli w przypadku tego zadania stosunek prawdopodobieństwa bezdzietności do prawdopodobieństwa posiadania dzieci.
- Wpływ na bezdzietność:
  - X1: Dodatni, jeśli kobieta mieszka w mieście, to względne ryzyko bezdzietności jest większe, i dokładnie rośnie $e^{0{,}021}$ razy (czyli około 1,02 razy = zwiększa się o 2%).
  - X2: Ujemny, jeśli kobieta nie ma wyższego wykształcenia, to względne ryzyko spada $e^{-0{,}034}$ = 0,97 razy (czyli zmniejsza się o około 3%).
  - X3: Ujemny, jeśli kobieta jest religijna, względne prawdopodobieństwo bezdzietności spada $e^{-0{,}020}$ = 0, 98 razy (czyli zmniejsza się o około 2%).
  - X4: Dodatni, jeśli kobieta jest religijna, względne prawdopodobieństwo bezdzietności spada $e^{-0{,}020}$ = 0, 98 razy (czyli zmniejsza się o około 2%). (zmienna nieistotna statystycznie)

Efekty krańcowe w modelu logitowym/probitowym

Nie interpretujemy bezpośrednio oszacowanych parametrów (jedynie znaki mogą podpowiadać kierunek zależności). Trzeba obliczyć pochodną funkcji y po danej zmiennej egzogenicznej x. Efekt krańcowy dla modelu logitowego:

$$\frac{\partial P_i}{\partial x_{ij}} = P_i(1-P_i)\beta_j = F\left(\beta_0 + \sum_{j=1}^{k}\beta_j\, x_{ij}\right)\left[1 - F\left(\beta_0 + \sum_{j=1}^{k}\beta_j\, x_{ij}\right)\right]\beta_j =$$

$$= \frac{\exp(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij})}{[1+exp(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij})]^2}\,\beta_j$$

Dla modelu probitowego:

$$\frac{\partial P_i}{\partial x_{ij}} = f(\beta_0 + \sum_{j=1}^{k}\beta_j\, x_{ij})\beta_j$$

Gdzie f to funkcja rozkładu prawdopodobieństwa standaryzowanego rozkładu normalnego.

- Ponieważ mamy do czynienia z modelem nieliniowym, efekt krańcowy zmienia się zależnie od wartości wszystkich zmiennych X. W praktyce podaje się efekt krańcowy dla średnich wartości zmiennych X.
- **Dla zmian logarytmu ilorazu szans w modelu logitowym efekt jest stały i równy wartości parametru beta.**

Przykład 11.2

(Cieślak, „Prognozowanie gospodarcze”, *przykład 4.5, s. 138)*

Pytanie:

Jakie jest prawdopodobieństwo bezdzietności u kobiety niereligijnej ze wsi, z wyższym wykształceniem, w wieku 30 lat?

Odpowiedź:

$$ln\frac{P_i}{1-P_i} = 0{,}0001 + 0{,}021 \cdot 0 - 0{,}034 \cdot 0 - 0{,}020 \cdot 0 + 0{,}006 \cdot 1 = 0{,}0061$$

$$\frac{P_i}{1-P_i} = e^{0{,}0061}$$

$$P_i = (1-P_i)e^{0{,}0061}$$

$$P_i = \frac{e^{0{,}0061}}{1+e^{0{,}0061}} = 0{,}5015$$

Na podstawie wyników estymacji zaproponowanego modelu można stwierdzić, iż przewidywane ryzyko bezdzietności u kobiety niereligijnej ze wsi, z wyższym wykształceniem, w wieku 30 lat wynosi 50,15%.

Weryfikacja modelu logitowego/probitowego: Pseudo R-kwadrat

Z powodu nieliniowości modelu logitowego inna miara dopasowania modelu jest stosowana – pseudo R-kwadrat.

Pseudo-R^2 McFaddena:

$$pseudoR^2 = 1 - \frac{lnL_{FM}}{lnL_{RM}}$$

ln L – logarytm funkcji wiarygodności

FM - pełny model, RM - model zredukowany (tylko ze stałą)

Celem tej miary jest porównanie alternatywnych modeli, nie stricte sama ocena, jak model jest dopasowany do danych.

Prognozy z modelu

Po oszacowaniu prawdopodobieństw należy określić, czy dany poziom prawdopodobieństwa oznacza, że Y jest równe "0" czy "1". Na przykład, jeśli prawdopodobieństwo jest równe 0.6 czy to oznacza "0" czy "1"?

Wszystko zależy od założonego progu odcięcia (cutoff):

- standardowa reguła - próg 0.5 (50%)
- zasada optymalnej wartości progowej (Cramer 1999) dla niezbilansowanej próby gdzie udział jedynek (Y = 1) w próbie stanowi próg δ. Ten próg jest równy średniej wartości zmiennej Y (można ją odczytać z wydruku Gretla). Np. 0.3 (30%).

Przykład:

Oszacowane prawdopodobieństwo	Prognoza (zasada standardowa)	Prognoza (delta Cramera = 0.40)
0.41	0	1
0.72	1	1
0.21	0	0
0.92	1	1
0.55	1	1
0.13	0	0

Wyniki estymacji modelu

```
Model 1: Estymacja Logit, wykorzystane obserwacje 1-500
Zmienna zależna (Y): Y
Błędy standardowe na bazie Hessian

           współczynnik  błąd standardowy     z     efekt krańcowy
  ---------------------------------------------------------------
  const      -2.42710        0.422593       -5.743
  X1          0.143078       0.0681603       2.099     0.0356672
  X2          0.0175878      0.00288492      6.096     0.00438440
  X3         -0.432675       0.190060       -2.277    -0.107445
  X4          0.124280       0.189998        0.6541    0.0309738

Średn.aryt.zm.zależnej  0.476000   Odch.stand.zm.zależnej  0.499924
McFadden R-kwadrat      0.080311   Skorygowany R-kwadrat   0.065860
Logarytm wiarygodności -318.2100   Kryt. inform. Akaike'a  646.4200
Kryt. bayes. Schwarza   667.4930   Kryt. Hannana-Quinna    654.6890

Liczba przypadków 'poprawnej predykcji' = 377 (75.4%)
f(beta'x) do średnich niezależnych zmiennych = 0.248
Test ilorazu wiarygodności: Chi-kwadrat(4) = 55.5747 [0.0000]

               Przewidywane
                 0     1
  Empiryczne 0  210    52
             1   71   167

Wyłączając stałą, największa wartość p jest dla zmiennej 5 (X4)
```

Delta Cramera

Zliczeniowy R^2 (dla odcięcia 0,5)

Test ilorazu wiarygodności (**likelihood ratio test**) wykorzystuje się do testowania zerowej hipotezy, że podzbiór współczynników β jest równy 0. Liczba bet w pełnym modelu jest równa p, a liczba bet w modelu zredukowanym wynosi r. (Pamiętaj, że model zredukowany to model gdzie wektor β w zerowej hipotezie jest równy 0.) Dlatego liczba bet testowanych w hipotezie zerowej jest równa $p-r$.

$$LR = -2\left[L(M_{int\,ercept}) - L(M_{full})\right]$$

Prognozy z modelu

Zliczeniowy R^2

Jest to udział przypadków trafionych prognoz do łącznej liczby obserwacji N.

Przypadki trafionych prognoz:

empiryczny Y = 1 a prognozowany Y = 1

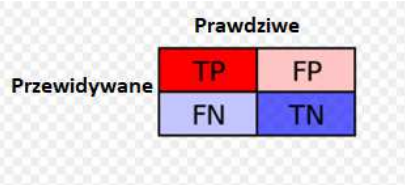
empiryczny Y = 0 a prognozowany Y = 0

		Przewidywane	
		0	1
Empiryczne	0	210	52
	1	71	167

Zliecznіowy R^2 = (210 + 167) / (210 + 52 + 72 + 167) = 377/500 = 75,4%

Krzywa ROC

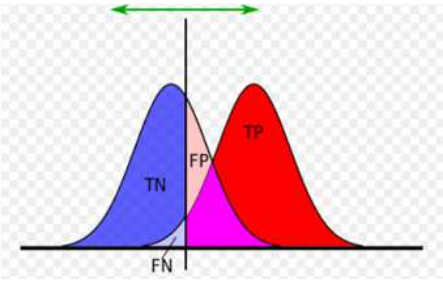
Tablicę trafności możemy podzielić zgodnie z diagramem

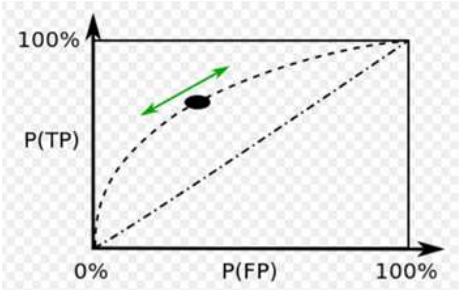


TP – true positive
FP – false positive – błąd I-szego rodzaju
FN – false negative – błąd II-giego rodzaju
TN – true negative

Krzywa ROC (receiver operating characteristic) – krzywa pokazująca kombinacje błędów α i β dla różnych wartości progu odcięcia (w rzeczywistości zależność pomiędzy prawdopodobieństwem TP a FP).

P(FP) = FPR = FP / (FP+TN); P(TP) = TPR = TP / (TP+FN)



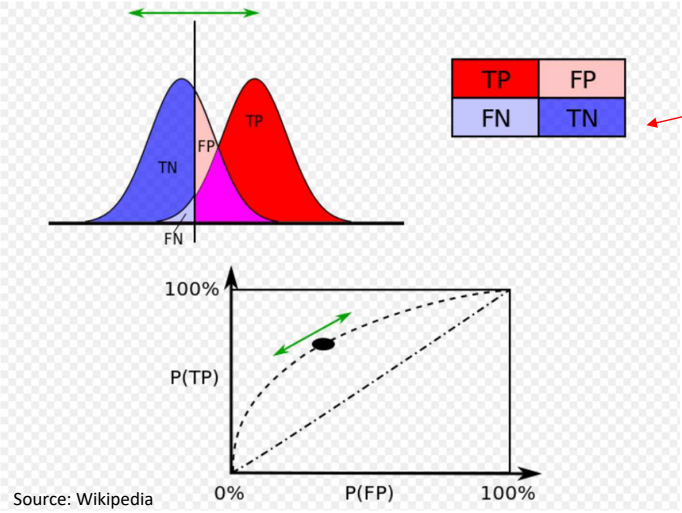


AUC

AUC (A) – Area Under the Curve, proporcja powierzchni pod krzywą w relacji do całkowitej powierzchni (równej 1). A ∈ $[0{,}5; 1]$ może być interpretowana jako miara dopasowania modelu.

A=0.5 -> model losowy

A=1.0 -> perfekcyjna dyskryminacja



TP	FP
FN	TN

Tablica kontyngencji

P(FP) = FPR = FP / (FP+TN)

P(TP) = TPR = TP / (TP+FN)

Inne modele zmiennej jakościowej: (ordered choice model)

Zmienna zależna przyjmuje więcej niż 2 wartości, np. 0,1,2 itd. Nie mają charakteru ilościowego, jednak charakteryzuje je uporządkowanie (możemy je porównać i ustawić w kolejności). To są często odpowiedzi z kwestionariuszy (np. jak oceniasz zajęcia z ekonometrii). Mogą to być cechy charakteryzujące gospodarkę, ale zawsze w pewnej hierarchii względem siebie, np. reżimy kursowe.

Analogicznie jak dla zmiennej binarnej modelujemy nieobserwowalną zmienną ciągłą y^*. Kluczowe jest zatem (oprócz parametrów równania) oszacowanie punktów odcięcia (κ_j) wartości y^* na przedziały odpowiadające różnym wartościom zmiennej zależnej.

$$y_i = \begin{cases} 0 & \text{dla } y_i{}^* \leq \kappa_1 \\ 1 & \text{dla } \kappa_1 < y_i{}^* \leq \kappa_2 \\ 2 & \text{dla } \kappa_2 < y_i{}^* \leq \kappa_3 \\ \vdots & \vdots \\ M & \text{dla } \kappa_M < y_i{}^* \end{cases}$$

Zamiast nieobserwowanej zmiennej podstawiamy $y_i^* = x_i'\beta + \varepsilon_i$

Parametry modelu (wraz z punktami odcięcia) szacujemy metodą MNW.

Przykład 11.3
Wielopoziomowy uporządkowany logit

```
Model 1: Estymacja Wielopoziomowy uporządkowany Logit, wykorzystane
obserwacje 1-4642
Zmienna zależna (Y): smoke
Błędy standardowe na bazie Hessian

             współczynnik   błąd standardowy      z        wartość p
  ------------------------------------------------------------------
  married     -1.02774         0.100585        -10.22     1.65e-024 ***
  hisp        -1.04228         0.268535         -3.881    0.0001    ***
  foreign1    -0.679424        0.255121         -2.663    0.0077    ***
  alcohol      1.45779         0.165146          8.827    1.07e-018 ***
  mage         0.0106108       0.00785502        1.351    0.1767
  edu         -0.129759        0.0183238        -7.081    1.43e-012 ***
  fedu        -0.0470623       0.0113488        -4.147    3.37e-05  ***
  race         0.654595        0.114191          5.732    9.90e-09  ***

  cut1        -0.484520        0.242133         -2.001    0.0454    **
  cut2        -0.133709        0.242415         -0.5516   0.5812
  cut3         0.719808        0.245098          2.937    0.0033    ***

Średn.aryt.zm.zależnej  0.399612   Odch.stand.zm.zależnej  0.898863
Logarytm wiarygodności -2928.722   Kryt. inform. Akaike'a  5879.445
Kryt. bayes. Schwarza   5950.316   Kryt. Hannana-Quinna    5904.378

Liczba przypadków 'poprawnej predykcji' = 3765 (81.1%)
Test ilorazu wiarygodności: Chi-kwadrat(8) = 978.207 [0.0000]
```

Współczynniki interpretujemy jakościowo, tj. według znaku.

progi

Zadania

Zadanie 11.1

Wykorzystując zbiór **coke.gdt** oszacuj model zmiennej jakościowej określającej wybór napoju typu cola:

(1 – coke, 0 – pepsi)

1. Zastosuj model typu LMP
2. Zastosuj model logitowy
3. Zastosuj model probitowy
4. Oceń wykresy empirycznych i oszacowanych wartości y.

Zadanie 11.2

Wykorzystując zbiór danych **coke.gdt** oszacuj model logitowy.

1. Zinterpretuj oszacowania parametrów
2. Zinterpretuj efekty krańcowe
3. Co znaczą wyniki testu *z* oraz Chi-kwadrat?
4. Zinterpretuj zliczeniowy R-kwadrat.
5. Ile wynosi wartość progowa prawdopodobieństwa (cutoff value) według zasady Cramera?

Zadanie 11.3

Na próbie 750 kobiet oszacowano model logitowy. Zmienna objaśniana to zmienna binarna i jest równa 1 jeśli kobieta była zatrudniona w 2010. Skłonność kobiet do pracy jest zależna od wieku (age), liczby lat doświadczenia w ostatnim miejscu pracy (experience), liczby posiadanych dzieci poniżej 6 roku życia (children) oraz dochodów gospodarstwa domowego w dolarach (income). Rezultaty estymacji są następujące:

	coefficient	Standard error	z	p-value
const	3.217	0.564	5.701	3.5e-08
age	-0.107	0.013	-8.049	3.4e-15
experience	0.130	0.013	9.866	2.9e-22
children	-1.324	0.191	-6.920	1.6e-11
income	2.622e-05	7.315e-06	3.584	6.5e-04

a) Zapisz oszacowaną postać modelu.

b) Czy zmienne objaśniające są istotne?

c) Oblicz prawdopodobieństwo, że kobieta w wieku lat 35, z doświadczeniem 2-letnim w ostatnim miejscu pracy, mająca 3 dzieci w wieku poniżej 6 lat była zatrudniona, jeśli roczne dochody jej gospodarstwa domowego wynosiły 50 000 dolarów.

Zadanie 11.3, cd.

Na próbie 750 kobiet oszacowano model logitowy. Zmienna objaśniana to zmienna binarna i jest równa 1 jeśli kobieta była zatrudniona w 2010. Skłonność kobiet do pracy jest zależna od wieku (age), liczby lat doświadczenia w ostatnim miejscu pracy (experience), liczby posiadanych dzieci poniżej 6 roku życia (children) oraz dochodów gospodarstwa domowego w dolarach (income). Rezultaty estymacji są następujące:

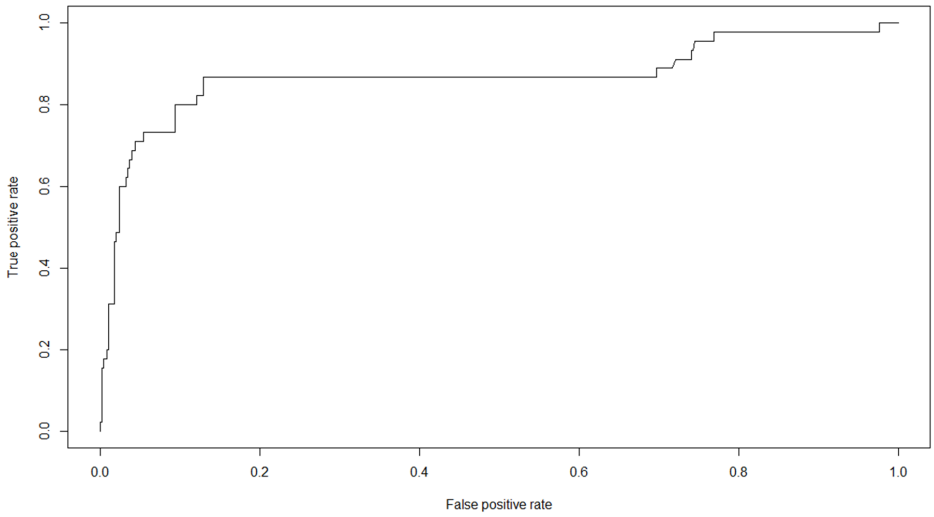
	coefficient	Standard error	z	P-value
const	3.217	0.564	5.701	3.5e-08
age	-0.107	0.013	-8.049	3.4e-15
experience	0.130	0.013	9.866	2.9e-22
children	-1.324	0.191	-6.920	1.6e-11
income	2.622e-05	7.315e-06	3.584	6.5e-04

d) Oblicz efekt krańcowy dla zmiennej experience (średnie wartości regresorów w próbie to: age = 41, experience = 5, children=1.1, income=40 000. Zinterpretuj rezultaty.

e) Zinterpretuj ocenę parametru przy zmiennej „children”.

Zadanie 11.4

Zinterpretuj rezultaty z modelu probitowego binarnej zmiennej objaśnianej:



AUC = 0.8689377

Zadanie 11.5

Wykorzystując dane **Bundesliga.gdt** oszacuj model logitowy w 3 wersjach:

1. Zmienną zależną jest „Win" (jeśli drużyna gospodarzy wygrywa: Win = 1, w przeciwnym wypadku Win = 0)
2. Zmienna zależna to „Lose" (jeśli drużyna gospodarzy przegrywa: Lose = 1, w przeciwnym wypadku Lose = 0)
3. Zmienna zależna to „Draw" (jeśli remis to Draw = 1, w przeciwnym wypadku Draw = 0)

Wykorzystując odpowiednie regresory z listy zmiennych w pliku z danymi.

Pytania:

a) Które zmienne są statystycznie istotne a które nie?
b) Który model lepiej prognozuje wynik?
c) Jak bardzo prawdopodobieństwo zwycięstwa zmienia się kiedy drużyną gospodarzy jest Bayern Munich? I co się dzieje z prawdopodobieństwem zwycięstwa kiedy Bayern Munich jest drużyną gości?

Temat 12

Endogeniczność w modelu ekonometrycznym

Katarzyna Bech-Wysocka i Michał Rubaszek

- Losowe zmienne objaśniające
- Powrót do statystyki: metody asymptotyczne
- Prawo iterowanych oczekiwań
- Egzogeniczność i endogeniczność zmiennych objaśnianych
- Skutki endogeniczności dla estymatora MNK
- Przyczyny endogeniczności zmiennej objaśnianej

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej i zbioru zmiennych objaśniających
3. Zebranie danych
4. **Estymacja: wybór metody**
5. Weryfikacja
6. Zastosowanie

Przypomnienie: Założenia KMNK

Założenia KMNK (przypomnienie z Temat 2)

A1. Prawdziwy model jest następujący: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

A2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ oraz $E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$

A3. $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

A4. $\mathbf{X}$ jest nielosową macierzą, której rząd wynosi $rank(X) = (K+1) < N$

A5. $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$

W Temat 12-14 skupimy się na **A2**.

- **A2** oznacza, że jedynie zmienna objaśniana (y) jest losowa, zaś wartości regresorów (x_k) są znane.
- Założenie o nielosowości regresorów może być prawdziwe w niektórych modelach, np. w modelu trendu $y_t = \beta_0 + \beta_1 t + \varepsilon_t$. W większości przypadków jest to tylko uproszczenie rzeczywistości w celu zastosowania prostszych metod analizy.
- W Tematach 12-14 będziemy zakładać, że zarówno zmienna objaśniana jak i zmienne objaśniające są zmiennymi losowymi i zobaczymy jak to wpływa na własności estymatora MNK.

Podstawy statystyki

Kontynuacja bloku z Temat 1

Zgodność estymatora

Zbieżność według prawdopodobieństwa (Convergence in Probability)

Ciąg zmiennych losowych X_n **zbiega według prawdopodobieństwa** do zmiennej losowej X, jeżeli dla dowolnego $\varepsilon > 0$:

$$\lim_{n\to\infty} \Pr(|X_n - X| > \varepsilon) = 0.$$

Zapisujemy to jako

$$X_n \xrightarrow{p} X \quad \text{lub} \quad \operatorname{plim}_{n\to\infty} X_n = X.$$

Przykład: Jeżeli $X_n = X + Y_n$, gdzie $Y_n \sim \mathcal{N}(\frac{1}{n}, \frac{\sigma^2}{n})$, to $X_n \xrightarrow{p} X$

Uwaga: Jeżeli $P(X = \mathbf{c}) = 1$, to mówimy o zbieżności według prawdopodobieństwa do wektora $\mathbf{c}$

Zgodność estymatora

Estymator $\widehat{\boldsymbol{\beta}_n}$ jest **zgodny**, jeżeli wraz ze wzrostem liczby obserwacji n, jego wartość zbiega według prawdopodobieństwa do prawdziwej wartości parametru $\boldsymbol{\beta}$:

$$\operatorname{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}_n} = \boldsymbol{\beta}.$$

Nieobciążoność a zgodność estymatora

Przypadek 1: Estymator nieobciążony, ale **niezgodny**:

Wyobraźmy sobie, że na podstawie n obserwacji $X_i \sim iid\ \mathcal{N}(\mu, \sigma^2)$ chcemy oszacować parametr μ. Niech estymatorem będzie pierwsza obserwacja (niezależnie od liczebności próby), tj.:

$$\widehat{\mu_n} = X_1$$

Zauważmy, że $\mathrm{E}(\widehat{\mu_n}) = E(X_1) = \mu$, a zatem jest to estymator nieobciążony. Jednakże, nie jest to estymator zbieżny, ponieważ wraz ze wzrostem obserwacji $\widehat{\mu_n}$ nie zbiega do μ:

$$\lim_{n\to\infty} \Pr(|\widehat{\mu_n} - \mu| > \varepsilon) = \lim_{n\to\infty} \Pr(|X_1 - \mu| > \varepsilon) \neq 0$$

Przypadek 2: Estymator obciążony, ale **zgodny:**

Niech, dla poprzedniego zbioru danych, estymatorem będzie wyrażenie:

$$\widehat{\mu_n} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}$$

Zauważmy, że $E(\widehat{\mu_n}) = \mu + \frac{1}{n}$, a zatem jest to estymator obciążony. Jednakże, jest to estymator zgodny. W szczególności, $\lim_{n\to\infty} E(\widehat{\mu_n}) = \mu$ oraz $\lim_{n\to\infty} Var(\widehat{\mu_n}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$ implikuje:

$$\lim_{n\to\infty} \Pr(|\widehat{\mu_n} - \mu| > \varepsilon) = 0$$

Słabe prawo wielkich liczb

Słabe Prawo Wielkich Liczb (WLLN, Weak Law of Large Numbers)

WLLN wskazuje, że dla zmiennych IID średnia z próby zbiega według prawdopodobieństwa do wartości oczekiwanej. W szczególności, dla zmiennych IID X_i, gdzie $E(X_i) = \mu$ oraz $E(|X_i|) < \infty$:

$$\operatorname{plim}_{n\to\infty} \widehat{\mu_n} = \operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

WLLN dostarcza warunków koniecznych do upewnienia się, że momenty w próbie dążą do odpowiednich momentów w populacji.

Przykład: Rozważmy wielokrotny rzut monetą, dla której szanse na wyrzucenie orła lub reszki są identyczne. Jeżeli rzucimy monetą 10 razy, prawdopodobnie proporcje orłów i reszek nie będą takie same. Prawo wielkich liczb mówi, że ze wzrostem liczby rzutów monetą proporcja wyrzuconych orłów będzie dążyć do 0.5.

Zbieżność według rozkładu

Zbieżność według rozkładu (Convergence in Distribution)

Ciąg zmiennych losowych X_n **zbiega według rozkładu** do zmiennej losowej X, jeżeli:

$$\lim_{n\to\infty} \Pr(X_n \leq x) = F_X(x),$$

gdzie $F_X(x)$ jest dystrybuantą dla zmiennej losowej X. Zapisujemy to jako

$$X_n \xrightarrow{d} X.$$

- Rozkład $F_X(x)$ nazywamy **rozkładem asymptotycznym**.
- Jeżeli nie znamy dokładnego rozkładu estymatora (np. gdy składnik losowy w modelu ekonometrycznym nie ma rozkładu normalnego), możemy zastanawiać się, jak wygląda rozkład asymptotyczny. Jest on dobrym przybliżeniem rozkładu estymatora uzyskanego na podstawie dużej próby.

Centralne twierdzenie graniczne

Najważniejszym zastosowaniem zbieżności według rozkładu jest **Centralne Twierdzenie Graniczne** (CLT, Central Limit Theorem). Jest ono bardzo pomocne przy wyprowadzeniu wzorów dla asymptotycznego rozkładu estymatora MNK i statystyk testowych.

Centralne Twierdzenie Graniczne (Central Limit Theorem)

Jeżeli X_i są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, takiej samej wartości oczekiwanej $\mu = E(x_i)$ oraz (skończonej) wariancji $\sigma^2 = Var(X_i)$ to zmienna losowa

$$Z_n = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$$

gdzie $\bar{x}_n = \frac{1}{n}\sum x_i$, zbiega wraz z liczebnością próby n do rozkładu $N(0,1)$:

$$Z_n \xrightarrow{d} N(0,1).$$

Rozkłady asymptotyczne

- CLT gwarantuje, że nawet jeśli składniki losowe nie mają rozkładu normalnego (ale są iid z wartością oczekiwaną 0 i wariancją σ^2) to i tak dla estymatora MNK:

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}_n} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(0, \sigma^2 \left(\operatorname{plim} \frac{1}{n} X'X\right)^{-1}\right)$$

  gdzie $\widehat{\boldsymbol{\beta}_n}$ to estymator MNK oparty n obserwacji

- Oznacza to, że asymptotyczny rozkład estymatora MNK jest rozkładem normalnym, nawet jeżeli nie jest spełnione założenie o normalności rozkładu składników losowych
- A zatem, przy dużej liczebności próby statystyki testów istotności będą miały rozkłady asymptotyczne rozkłady normalne czy χ^2. Zauważmy, że:

$$t_n \xrightarrow{d} \mathcal{N}(0,1)$$
$$F_{m,n} \xrightarrow{d} \chi^2_m/m$$

Przydatne zależności

Twierdzenie Słuckiego (Slutsky Theorem)

Niech dane będą dwa ciągi, dla których $X_n \xrightarrow{p} X$ oraz $Y_n \xrightarrow{p} Y$, oraz ciągła funkcja g.
Implikacjami twierdzenia Słuckiego są zależności:

$$\operatorname{plim} g(X_n) = \mathrm{g}(\operatorname{plim} X_n) = \mathrm{g}(X)$$
$$\operatorname{plim} X_n + Y_n = \operatorname{plim} X_n + \operatorname{plim} Y_n = X + Y$$
$$\operatorname{plim} X_n Y_n = \operatorname{plim} X_n \times \operatorname{plim} Y_n = XY$$

Prawo iterowanych oczekiwań (LIE, Law of Iterated Expectations):

LIE mówi, że dla dwóch zmiennych losowych X i Y zachodzi zależność:

$$E(Y) = E[E(Y|X)].$$

Implikuje to, że:

$$E(XY) = E[E(XY|X)] = E[XE(Y|X)] = E_X[XE(Y|X)]$$

Model ekonometryczny

Losowe zmienne objaśniające

Założenia – nowy zbiór założeń

Nielosowe regresory:	**Losowe regresory:**
A1: Prawdziwy jest model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$	**A1:** Prawdziwy jest model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$
A2: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ oraz $E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$	**A2:** $E(\boldsymbol{\varepsilon}\|\mathbf{X}) = \mathbf{0}$
A3: $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$	**A3:** $Var(\boldsymbol{\varepsilon}\|\mathbf{X}) = \sigma^2\mathbf{I}$
A4: $\mathbf{X}$ jest nielosową macierzą i ma pełny rząd kolumnowy.	**A4:** $\mathbf{X}$ ma pełny rząd kolumnowy
A5: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$	**A5:** $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$

Założenie **A2** w modelu z losowymi regresorami implikuje (korzystamy z LIE):

B1. $E(\boldsymbol{\varepsilon}) = E(E(\boldsymbol{\varepsilon}|\mathbf{X})) = \mathbf{0}$
B2. $E(\mathbf{X}'\boldsymbol{\varepsilon}) = E(E(\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X})) = E(\mathbf{X}'E(\boldsymbol{\varepsilon}|\mathbf{X})) = \mathbf{0}$

A zatem w obu przypadkach, przy spełnieniu A1-A4, MNK jest BLUE.
Nasze zainteresowanie: co się dzieje, gdy **A2** nie jest spełnione.

Losowe regresory:
skutki niespełnienia założenia A2 - obciążenie

- Załóżmy, że A2 nie jest spełnione, a zatem: $E(\boldsymbol{\varepsilon}|\mathbf{X}) \neq \mathbf{0}$
- Przypomnienie, wzór na estymator MNK:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.$$

- Sprawdźmy, czy jest to estymator nieobciążony:

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}) = \\ &= E(\boldsymbol{\beta}) + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] \overset{LIE}{=} \\ &= \boldsymbol{\beta} + E\big[E\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}\big)\big] = \\ &= \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}|\mathbf{X})] \neq \boldsymbol{\beta} \end{aligned}$$

Niespełnienie założenia A2 prowadzi do obciążenia estymatora MNK

Egzogeniczność

- W ekonometrii zamiast założenia **A2**, $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$, często rozpatrujemy występowanie korelacji między składnikiem losowym ε_i oraz regresorami x_{ki} dla $k = 1,2,\dots,K$.

- Występują dwie możliwości:
  - egzogeniczny regresor: $cor(\varepsilon_i, x_{ki}) = 0$
  - endogeniczny regresor: $cor(\varepsilon_i, x_{ki}) \neq 0$

Jeżeli spełnione są założenia:

A1: Prawdziwy jest model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

A2: Wszystkie regresory są egzogeniczne

A3: $Var(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$

A4: X ma pełny rząd kolumnowy

A5: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$

To estymator MNK jest BLUE

Losowe regresory:
skutki niespełnienia założenia A2 - zgodność

Żeby sprawdzić, czy estymator MNK jest zgodny musimy sprawdzić, czy $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. Zauważmy, że:

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} = \operatorname*{plim}_{n\to\infty} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \operatorname*{plim}_{n\to\infty} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

$$\operatorname*{plim}_{n\to\infty} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \times \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right) \qquad \text{[implikacje twierdzenie Słuckiego]}$$

WLLN implikuje natomiast, że:

$$\frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i') \xrightarrow{p} E(\mathbf{x}_i\mathbf{x}_i')$$

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\boldsymbol{\varepsilon}_i) \xrightarrow{p} E(\mathbf{x}_i\boldsymbol{\varepsilon}_i)$$

Egzogeniczne regresory: $E(\mathbf{x}_i\boldsymbol{\varepsilon}_i) = \mathbf{0}$, zaś estymator $\widehat{\boldsymbol{\beta}}$ jest zgodny
Endogeniczne regresory: $E(\mathbf{x}_i\boldsymbol{\varepsilon}_i) \neq \mathbf{0}$, zaś estymator $\widehat{\boldsymbol{\beta}}$ nie jest zgodny

Przykład 12.1. Endogeniczność regresorów a zgodność estymatora MNK

W pakiecie R (plik `T12.R`) wygenerowano szeregi z następującego DGP (data generating proces):

$$x_i \sim \mathcal{N}(0,\sigma_x^2); \quad \varepsilon_i \sim \mathcal{N}(0,\sigma_\varepsilon^2);\ cor(x_i,\varepsilon_i) = \rho;\ \ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

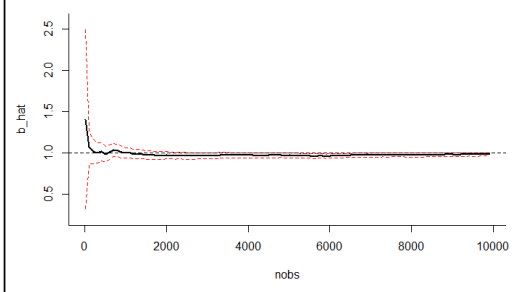
Założono następujące wartości parametrów:

M1: $\beta_0 = 1; \beta_1 = 1; \sigma_x = 1; \sigma_\varepsilon = 1; \rho = 0.0$ [egzogeniczny regresor]

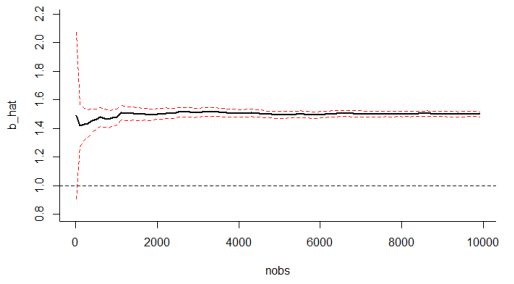
M2: $\beta_0 = 1; \beta_1 = 1; \sigma_x = 1; \sigma_\varepsilon = 1; \rho = 0.5$ [endogeniczny regresor]

Następnie, na podstawie rosnącej liczby obserwacji oszacowano parametr β_1

Model 1:



Model 2:



Pytanie: dlaczego oszacowania z modelu 2 nie zbiegają do $\beta_1 = 1$?

Przyczyny endogeniczności regresorów

Przyczyny endogeniczności regresora x_k mogą być następujące:

1. **Błąd pomiaru** zmiennej x_k (nie posiadamy dokładnych danych)
2. **Pominięcie zmiennej** objaśniającej z (omitted variable), która jest skorelowana zarówno ze zmienną x_k jak i zmienną objaśnianą y
3. **Symultaniczność**: występuje sprzężenie zwrotne między zmiennymi x_k i y, czyli gdy zmiana y prowadzi do zmian wartości x_k
4. **Autokorelacja składnika losowego w modelach autoregresyjnych**

Endogeniczność: błąd pomiaru

- Załóżmy, że prawdziwa zależność jest dana przez model (z egzogenicznym regresorem):

$$y_i = \beta_0 + \beta_1 x_i^* + \varepsilon_i.$$

- Estymacja powyższego modelu nie jest możliwa, ponieważ nie znamy wartości x^*. Obserwujemy jedynie zmienną x, która jest nieprecyzyjną miarą (tj. proxy) dla x^*:

$$x_i = x_i^* + u_i,$$

gdzie u ma wartość oczekiwaną 0, wariancję σ_u^2 i jest nieskorelowana z x^* oraz ε.

- Po połączeniu obydwu równań uzyskujemy model:

$$y_i = \beta_0 + \beta_1 x_i + v_i,$$

w którym $v_i = (\varepsilon_i - \beta_1 u_i)$. W modelu tym zachodzi zależność:

$$cov(x_i, v_i) = cov\big((x_i^* + u_i), (\varepsilon_i - \beta_1 u_i)\big) = -\beta_1 \sigma_u^2 \neq 0,$$

- **A zatem x_i jest endogenicznym regresorem, zaś estymator MNK nie jest zgodny!**

Pominięcie zmiennej objaśniającej

- Załóżmy, że prawdziwy jest następujący model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,$$

w którym $cov(x, z) = \gamma_{xz} \neq 0$ oraz $\beta_2 \neq 0$.

- Specyfikacja modelu empirycznego pomija istotną zmienną z (która np. nie jest obserwowalna):

$$y_i = \beta_0 + \beta_1 x_i + \nu_i,$$

- Zauważmy, że zmienna z jest zawarta w składniku losowym, tj.

$$\nu_i = \beta_2 z_i + \varepsilon_i$$

- W takim przypadku, zachodzi zależność:

$$cov(x_i, \nu_i) = cov(x_i, \beta_2 z_i + \varepsilon_i) = \beta_2 \gamma_{xz} \neq 0.$$

- **A zatem x_i jest endogenicznym regresorem, zaś estymator MNK nie jest zgodny!**

Przykład 12.2. Endogeniczność w modelu AR(1) z autokorelacją składnika losowego

Zbiór danych `endo_ar.gdt` zawiera sztucznie wygenerowany szereg czasowy dla y_t, dla którego DGP jest następujący

$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$, gdzie $u_t \sim N(0,1)$ *i* $\varepsilon_0 = 0$

$y_t = 0.2y_{t-1} + \varepsilon_t$, gdzie $y_0 = 1$.

Endogeniczność w tym modelu wynika z tego, że:

$$cov(y_{t-1}, \varepsilon_t) = cov(\delta y_{t-2} + \varepsilon_{t-1}, 0.5\varepsilon_{t-1} + u_t) = 0.5var(\varepsilon_{t-1}) \neq 0$$

Oszacowania parametrów modelu AR(1), $y_t = \delta y_{t-1} + \varepsilon_t$, ilustrują obciążenie estymatora MNK

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 2-100 (N = 99)
Zmienna zależna (Y): y

           współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const      0,0357354       0,0961049        0,3718    0,7108
  y_1        0,640545        0,0785393        8,156     1,25e-012 ***
```

Przykład 12.3: Symultaniczność

W zbiorze danych `wine_aus.gdt` znajdują się następujące zmienne opisujące rynek konsumpcji i produkcji wina w Australii w latach 1955-1974:

- *Q* – spożycie wina na osobę (w litrach)
- *Pw* – cena wina (relatywnie do CPI)
- *Pb* – cena piwa (relatywnie do CPI)
- *A* – średnie wydatki na reklamę wina ($/osobę)
- *Y* – średni dochód ($/osobę)
- *S* – koszty magazynowania (indeks).

Rynek ten można scharakteryzować następującym modelem:

popyt: $\ln(Q_t) = \beta_0 + \beta_1 \ln(Pw_t) + \beta_2 \ln(Pb_t) + \beta_3 \ln(Y_t) + \beta_4 \ln(A_t) + \varepsilon_t$

podaż: $\ln(Q_t) = \alpha_0 + \alpha_1 \ln(Pw_t) + \alpha_2 \ln(S_t) + u_t$

Q_t oraz Pw_t to dwie zmienne endogeniczne w tym systemie.

Potraktujmy ten system jako dwa osobne równania i spróbujmy oszacować rozdzielnie równanie popytu i równanie podaży stosując MNK.

Przykład 12.3: Symultaniczność

Równanie popytu:

Zmienna zależna (Y): l_Q

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−23,6512	3,91278	−6,045	2,24e-05	***
l_P_w	1,15826	0,289826	3,996	0,0012	***
l_P_b	−0,274827	0,607658	−0,4523	0,6575	
l_Y	3,21206	0,714004	4,499	0,0004	***
l_A	−0,602985	0,449740	−1,341	0,2000	

Równanie podaży

Zmienna zależna (Y): l_Q

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	−15,5671	0,848016	−18,36	1,21e-012	***
l_P_w	2,14495	0,238546	8,992	7,18e-08	***
l_S	1,38255	0,154509	8,948	7,69e-08	***

Pytania:

- Które regresory mogą być endogeniczne w danych modelach?
- Czy znaki oszacowanych parametrów pokrywają się z teorią ekonomii?

Przykład 12.4. Endogeniczność regresorów

Dlaczego regresory w poniższych modelach są endogeniczne?

A. Lepsze wykształcenie powoduje wyższe zarobki.

B. Znajomość ekonometrii prowadzi do większego zainteresowania pakietem Gretl.

Odpowiedzi:

A. **Ukryta zmienna:** poziom wykształcenia oraz zarobki mogą być kształtowane przez wspólny czynnik (np. inteligencja), którego nie uwzględniliśmy w modelu. Zmienna wykształcenie jest skorelowana ze składnikiem losowym, który zawiera informację dotyczącą czynników wpływających na y, które nie zostały w modelu uwzględnione.

B. **Symultaniczność.** W tym przykładzie istotnym pytaniem jest to, co jest przyczyną i skutkiem. Zależność prawdopodobnie jest obustronna.

Zadania

Zadanie 12.1

W trakcie zajęć dowiedzieliśmy się jak poważne są konsekwencje błędnego pomiaru regresora. Czy błąd pomiaru zmiennej objaśniającej y jest równie problematyczny?

Rozważmy model:

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

w którym badamy wpływ zmiennej x na y^*. Niestety, obserwujemy tylko zmienną y (tj. jej proxy):

$$y_i = y_i^* + u_i,$$

gdzie u to składnik losowy ze średnią 0 i wariancją σ_u^2, nieskorelowany z x oraz ε.

a. Zapisz model, który możemy oszacować na podstawie dostępnych danych
b. Sprawdź czy estymator MNK $\widehat{\beta_1}$ jest zgodny
c. Oblicz wariancję estymatora MNK w tym modelu. Porównaj ją do wariancji estymatora MNK w modelu, w którym zmienna y^* jest obserwowalna.
d. Jakie są konsekwencje błędu pomiaru zmiennej objaśnianej?

Zadanie 12.2

Przeprowadźmy symulację, która pozwoli nam określić wpływ błędu pomiaru w praktyce.

W pliku `BladPomiaru.gdt` znajdują się następujące zmienne:

- $xstar$ - „prawdziwe" wartości x^*, które zostały wylosowane z rozkładu $\mathcal{N}(0,1)$
- $ystar$ – „prawdziwe" wartości, które zostały wykosowane z DGP:

$$y_i^* = 25 + 0.6x_i^* + \varepsilon_i, \text{ gdzie } \varepsilon_i \sim \mathcal{N}(0,1)$$

a. Stwórz zmienne, które są obserwowane z błędem (przyjmij $\sigma_v = 1$ oraz $\sigma_\omega = 1$):

$$y_i = y_i^* + v_i, \text{ gdzie } v_i \sim \mathcal{N}(0, \sigma_v^2)$$
$$x_i = x_i^* + \omega_i, \text{ gdzie } \omega_i \sim \mathcal{N}(0, \sigma_\omega^2)$$

b. Oszacuj trzy modele:
M1: $y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i$
M2: $y_i = \beta_0 + \beta_1 x_i^* + \varepsilon_i$
M3: $y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$
Porównaj oszacowania i błędy szacunku. Jakie są wnioski przeprowadzonej analizy?
c. Skróć próbę do 100 obserwacji i powtórz polecenia z punktu b.
d. Porównaj oszacowania i błędy szacunku modelu **M2** dla różnych wartości $\sigma_v \in \{1, 2, 4, 8\}$
e. Porównaj oszacowania i błędy szacunku modelu **M3** dla różnych wartości $\sigma_\omega \in \{1, 2, 4, 8\}$

Zadanie 12.3

Przeprowadź dyskusję na temat tego, co może powodować endogeniczność regresorów w poniższych modelach?

a. **Wpływ reklamy na sprzedaż**. Firma musi zdecydować, ile zainwestować w reklamę w celu zwiększenia sprzedaży swoich produktów. Dyrektor chce wiedzieć o ile wzrośnie sprzedaż z każdym dodatkowym dolarem wydanym na reklamę, tj.

$$sales_i = \beta_0 + \beta_1 advert_i + \varepsilon_i.$$

b. **Posiadanie samochodu a ocena z ekonometrii.** Rozważana jest następująca regresja opisująca wpływ posiadania samochodu na ocenę z egzaminu z ekonometrii:

$$grade_i = \beta_0 + \beta_1 car_i + \varepsilon_i.$$

c. **Zależność między ceną nieruchomości a wysokością czynszu.** Celem analizy jest ustalenie jak czynsz zależy od wartości nieruchomości. Dane opisują średnią cenę oraz stawkę czynszu za metra kwadratowy w różnych lokalizacjach Polski. Specyfikacja modelu jest następująca:

$$HouseRent_i = \beta_0 + \beta_1 HousePrice_i + \varepsilon_i.$$

Zadanie 12.4

Pewien naukowiec wierzy, że poziom aktywności w szarej strefie gospodarki (y) jest dodatnio powiązany z wysokością podatków (x) oraz ujemnie powiązany z wydatkami rządu na zwalczanie szarej strefy (z). Dane o y, x oraz z obejmują 30 krajów rozwijających się i 30 krajów rozwiniętych.

Uzyskano następujące oszacowania MNK.

Kraje rozwinięte:	**Kraje rozwijające się:**
M1: $\widehat{\ln(y_i)} = \underset{(0.863)}{-1.137} + \underset{(0.154)}{0.699}\ln(x_i) - \underset{(0.162)}{0.646}\ln(z_i)$	M1: $\widehat{\ln(y_i)} = \underset{(0.873)}{-1.122} + \underset{(0.137)}{0.806}\ln(x_i) - \underset{(0.117)}{0.091}\ln(z_i)$
M2: $\widehat{\ln(y_i)} = \underset{(1.069)}{-1.065} + \underset{(0.120)}{0.201}\ln(x_i)$	M2: $\widehat{\ln(y_i)} = \underset{(0.858)}{-1.024} + \underset{(0.090)}{0.727}\ln(x_i)$
M3: $\widehat{\ln(y_i)} = \underset{(0.896)}{-1.230} - \underset{(0.124)}{0.053}\ln(z_i)$	M3: $\widehat{\ln(y_i)} = \underset{(0.835)}{-2.824} - \underset{(0.116)}{0.427}\ln(z_i)$

Dodatkwo, wiadomo, że x oraz z są dodatnio skorelowane w obu próbach.

Jakie wnioski można wyciągnąć na podstawie analizy powyższych wyników?

Zadanie 12.5

Zbiór danych `GermanTrade.gdt` zawiera szeregi czasowe opisujące wolumen eksportu w Niemczech (oznaczane jako EX_t) oraz poziom realnego efektywnego kursu walutowego dla Niemiec (oznaczane jako $REER_t$).

a. Dla zmiennej $\Delta \ln EX_t$ zbuduj modele AR(1), AR(2) oraz AR(4).
W każdym zbadaj autokorelację składnika losowego.
Czy któryś z modeli nie powinien być szacowany MNK? Dlaczego?

b. Uwzględnij zmienną $\Delta \ln REER_t$ i jej opóźnienia poprzez zbudowanie modelu ADL dla eksportu. Rozpatrz modele ADL(1,0), ADL(2,0), ADL(1,1), ADL(4,0) oraz ADL(1,4).
Czy zastosowanie MNK w tych modelach dostarczy zgodnych oszacowań?

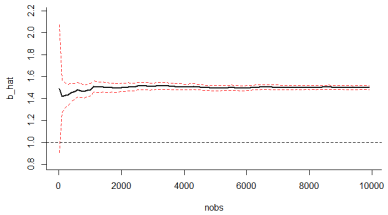
Zadanie 12.6

W przykładzie 12.1, w którym korzystano z pliku `T12.R`, wygenerowano szeregi z DGP

$$x_i \sim \mathcal{N}(0, \sigma_x^2); \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2); \; cor(x_i, \varepsilon_i) = \rho; \; y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Dla parametrów z **M2** ($\beta_0 = 1; \beta_1 = 1; \sigma_x = 1; \sigma_\varepsilon = 1; \rho = 0.5$) uzyskano wynik:

$$\text{plim}\, \widehat{\beta_1} = 1.5 \neq \beta_1$$



a. Czy potrafisz wyjaśnić, dlaczego $\text{plim}\, \widehat{\beta_1} = 1.5 \neq \beta_1$.

b. Korzystając z pliku `T12.R` przeprowadź własne symulacje dla różnych wartości $\rho \in \{0.25, 0.5, 0.75\}$. Jakie są wartości $\text{plim}\, \widehat{\beta_1}$? Czy widzisz zależność między ρ a $\text{plim}\, \widehat{\beta_1}$?

c. Przeprowadź własne symulacje dla różnych wartości $\sigma_x \in \{1,2,4\}$. Jakie są wartości $\text{plim}\, \widehat{\beta_1}$? Czy widzisz zależność między σ_x a $\text{plim}\, \widehat{\beta_1}$

d. Czy teraz potrafisz wyprowadzić, dlaczego w przykładzie 12.1 $\text{plim}\, \widehat{\beta_1} = 1.5$?
Obszerna odpowiedź na to pytanie jest w Temat 13.

Temat 13

Endogeniczność.

Metoda zmiennych instrumentalnych

Katarzyna Bech-Wysocka i Michał Rubaszek

- Zmienne instrumentalne
- Cechy zmiennej instrumentalnej: relevance i exogeneity
- Prawo iterowanych oczekiwań
- Estymator metody zmiennych instrumentalnych
- Dobór instrumentów
- Podwójna MNK

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. **Estymacja: wybór metody**
5. Weryfikacja
6. Zastosowanie

Zmienne instrumentalne

- Wiemy, że jeżeli regresory modelu $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ są endogeniczne, czyli $E(\mathbf{X}'\boldsymbol{\varepsilon}) \neq \mathbf{0}$, to estymator MNK parametru $\boldsymbol{\beta}$ jest niezgodny i obciążony.
- Wynika to z faktu, że wzór na estymator MNK jest równoznaczny z warunkiem $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$. Dowód:

$$\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{X}'\left(\mathbf{y} - \mathbf{X}'\hat{\boldsymbol{\beta}}\right) = \mathbf{X}'\left(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) = \mathbf{y} - \mathbf{y} = \mathbf{0}$$

Jak zatem otrzymać zgodne estymatory parametrów modelu z endogenicznymi regresorami?

Tworzymy zbiór dodatkowych zmiennych $\mathbf{Z} = [\mathbf{1}\ \mathbf{z}_1\ \mathbf{z}_2\ \ldots \mathbf{z}_L]$ spełniających dwa warunki:

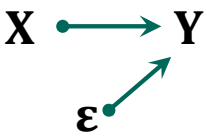
Skorelowania (relevance): zmienne $\mathbf{z}_l$ są skorelowane z $\mathbf{x}_k$

Egzogeniczności (exogeneity): zmienne $\mathbf{z}_l$ nie są skorelowane z $\boldsymbol{\varepsilon}$, tj. $E(\mathbf{Z}'\boldsymbol{\varepsilon}) = \mathbf{0}$.

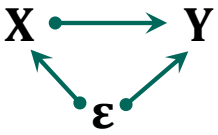
- Zmienne $\mathbf{z}_l$ mogą zostać wykorzystane przy stworzeniu zgodnego estymatora dla $\boldsymbol{\beta}$
- Zmienne te określamy jako **zmienne instrumentalne / instrumenty**
- Metoda estymacji określana jest jako **Metoda Zmiennych Instrumentalnych (MZI)**

Metoda zmiennych instrumentalnych

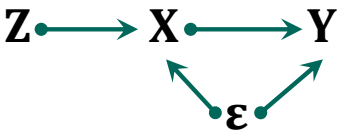
Schemat przyczynowości



- Klasyczne założenia modelu regresji liniowej są spełnione
- Estymator MNK jest zgodny i nieobciążony



- Korelacja między regresorem i składnikiem losowym
- Estymator MNK nie jest zgodny



- Zmienne instrumentalne $\mathbf{Z}$ wpływają na $\mathbf{Y}$ jedynie poprzez zmienną $\mathbf{X}$
- Można uzyskać zgodne estymatory metodą zmiennych instrumentalnych

Metoda zmiennych instrumentalnych

Intuicja metody zmiennych instrumentalnych

Przy wyznaczaniu wzoru na estymator parametru $\boldsymbol{\beta}$

Nieprawdziwy warunek estymatora MNK: $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$

Zamieniamy na prawdziwy warunek: $\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$

- Poszukujemy zatem wektora $\widehat{\boldsymbol{\beta}}^{\mathrm{MZI}}$, dla którego:

$$\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{MZI}}\right) = \mathbf{0}.$$

- Załóżmy, że $\mathbf{Z}'\mathbf{X}$ jest macierzą (kwadratową) odwracalną [warunek ten jest spełniony, gdy mamy dokładnie tyle samo zmiennych instrumentalnych co regresorów - na razie skupimy się na takim przypadku]. W takim przypadku, rozwiązaniem powyższego układu jest:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{MZI}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

- $\widehat{\boldsymbol{\beta}}^{\mathrm{MZI}}$ jest zazwyczaj obciążony w małych próbach, ale możemy udowodnić, że jest zgodny.

Estymator MZI: zgodność

Żeby sprawdzić, czy estymator MZI jest zgodny musimy sprawdzić, czy $\widehat{\boldsymbol{\beta}}^{\text{MZI}} \xrightarrow{p} \boldsymbol{\beta}$. Zauważmy, że:

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}}^{\text{MZI}} = \operatorname*{plim}_{n\to\infty} (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} = \operatorname*{plim}_{n\to\infty} (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}$$

$$\operatorname*{plim}_{n\to\infty} (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon} = \operatorname*{plim}_{n\to\infty} \left(\tfrac{1}{n}\mathbf{Z}'\mathbf{X}\right)^{-1} \times \operatorname*{plim}_{n\to\infty} \left(\tfrac{1}{n}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \qquad \text{[implikacje twierdzenie Słuckiego]}$$

WLLN implikuje natomiast, że:

$\frac{1}{n}\mathbf{Z}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i\mathbf{x}_i{}') \xrightarrow{p} E(\mathbf{z}_i\mathbf{x}_i{}') \neq 0$ [**warunek skorelowania** instrumentów, relevance]

$\frac{1}{n}\mathbf{Z}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i\boldsymbol{\varepsilon}_i) \xrightarrow{p} E(\mathbf{z}_i\boldsymbol{\varepsilon}_i) = 0$ [**warunek egzogeniczności** instrumentów, exogeneity]

A zatem:

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}}^{\text{MZI}} = \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \mathbf{0} = \boldsymbol{\beta}$$

zaś **estymator MZI jest zgodny**

Metoda zmiennych instrumentalnych
Testy istotności parametrów

- W MZI wzór na estymator wariancji składnika losowego σ^2 jest taki sam jak dla MNK:

$$s^2_{MZI} = \widehat{\sigma^2} = \frac{SSE}{N-(K+1)} = \frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{MZI}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{MZI}})}{N-(K+1)}$$

  gdzie SSE to suma kwadratów reszt.
- Z kolei asymptotyczna wariancja estymatora MZI wynosi:

$$\boldsymbol{\Sigma}_{\widehat{\beta}^{MZI}} = Var\left(\widehat{\boldsymbol{\beta}}^{\text{MZI}}\right) = E\left[\left(\widehat{\boldsymbol{\beta}}^{\text{MZI}}-\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}^{\text{MZI}}-\boldsymbol{\beta}\right)'\right] = \sigma^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}$$

- Podstawiając za nieznaną wartość σ^2 oszacowanie s^2_{MZI}, uzyskamy wzór na estymator wariancji:

$$\widehat{\boldsymbol{\Sigma}_{\widehat{\beta}^{MZI}}} = \widehat{Var\left(\widehat{\boldsymbol{\beta}}^{\text{MZI}}\right)} = s^2_{MZI}(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}$$

- Oszacowania $\widehat{\boldsymbol{\beta}}^{\text{MZI}}$ oraz macierz $\widehat{\boldsymbol{\Sigma}_{\widehat{\beta}^{MZI}}}$ możemy wykorzystać do weryfikacji hipotez nt. wartości parametrów tak jak to robiliśmy do tej poryw MNK.

Wybór instrumentów

- W praktyce najtrudniejszym elementem MZI jest znalezienie odpowiednich instrumentów. Wiemy, że dobre instrumenty są:
  - egzogeniczne, czyli nieskorelowane ze składnikiem losowym
  - silnie skorelowane z endogenicznymi regresorami
    [najlepszym instrumentem dla egzogenicznego regresora x_k jest zmienna x_k]
- Przypomnijmy, że asymptotyczna wariancja estymatora $\widehat{\boldsymbol{\beta}}^{\text{MZI}}$ to:

$$Var\left(\widehat{\boldsymbol{\beta}}^{\text{MZI}}\right) = \sigma^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}$$

- Jeżeli korelacja między regresorami i instrumentami jest niskie, to:
  - elementy macierzy $\mathbf{Z}'\mathbf{X}$ są niskie
  - elementy macierzy $(\mathbf{Z}'\mathbf{X})^{-1}$ są duże
  - $Var\left(\widehat{\boldsymbol{\beta}}^{\text{MZI}}\right)$ jest wysoka, a zatem precyzja oszacowań niska.

Im silniejsza korelacja między instrumentem i endogenicznym regresorem, tym mniejsza jest wariancja estymatora MZI

Liczba instrumentów $L > K$

- Do tej pory omawialiśmy sytuację, gdy liczba instrumentów jest równa liczbie regresorów, tj. $L = K$, czyli gdy model był **dokładnie zidentyfikowany** (exact identyfication)
- Możemy jednak wykorzystać większą liczbę instrumentów niż liczba regresorów, tj. $L > K$, czy gdy model jest **nadmiernie zidentyfikowany** (over-identification).
- W takim przypadku możemy przeprowadzić regresję MNK dla $\mathbf{X}$ względem $\mathbf{Z}$:

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{v},$$

dla której oszacowanie parametrów wynosi:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

i opisuje wagi przypisane poszczególnym instrumentów dla kolejnych regresorów.

- Po podstawieniu uzyskujemy kombinację liniową instrumentów $\mathbf{Z}$:

$$\widehat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\gamma}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X},$$

która może być wykorzystana w MZI w **dokładnie zidentyfikowanym** modelu:

$$\widehat{\boldsymbol{\beta}}^{\text{MZI}} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{y}$$

Zauważ, że $\widehat{\boldsymbol{\beta}}^{\text{MZI}} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{y} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y}$, czyli wzór na estymator MNK w regresji $\mathbf{y}$ na $\widehat{\mathbf{X}}$.

Liczba instrumentów $L > K$

- Procedura opisana na poprzednim slajdzie opisywana jest jako **Podwójna Metoda Najmniejszych Kwadratów 2MNK** (Two-stage least squares)
- Kiedy mamy więcej lub tyle samo instrumentów co regresorów ($L \geq K$) estymator $\widehat{\boldsymbol{\beta}}^{\text{MZI}}$ można uzyskać w dwóch krokach:

Krok 1:

- Oszacuj metodą MNK parametry modelu, w którym każdy regresor jest objaśniany przez wszystkie instrumenty:

$$\mathbf{x}_k = \mathbf{Z}\boldsymbol{\gamma}_k + \boldsymbol{v}_k \text{ dla } k = 1,2,\dots,K$$

Uwaga: dla egzogenicznych regresorów regresja zamienia się w tożsamość, tj. $\mathbf{x}_k = \mathbf{x}_k$.

- Oblicz wartości teoretyczne modeli i zapisz jako $\widehat{\mathbf{X}} = [1 \ \widehat{\mathbf{x}_1} \ \widehat{\mathbf{x}_2} \ \dots \ \widehat{\mathbf{x}_K}]$.

Krok 2:

- Przeprowadź regresję MNK $\mathbf{y}$ na $\widehat{\mathbf{X}}$, aby uzyskać wartość $\widehat{\boldsymbol{\beta}}^{\text{MZI}} = \widehat{\boldsymbol{\beta}}^{\text{2MNK}}$
- ... ale błędy szacunku oblicz, korzystając ze

$$\widehat{\boldsymbol{\Sigma}_{\widehat{\beta}^{2MNK}}} = Var\left(\widehat{\boldsymbol{\beta}^{2MNK}}\right) = s^2_{2MNK}\left(\widehat{\mathbf{X}}'\mathbf{X}\right)^{-1}\left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)\left(\mathbf{X}'\widehat{\mathbf{X}}\right)^{-1}$$

Przykład 13.1. Podwójna MNK

Rozważmy model popytu na ryby:

$$fish_i = \beta_0 + \beta_1 fish\_price_i + \varepsilon_i.$$

Jest to typowy przykład symultaniczności w modelu ekonometrycznym: popyt wpływa na cenę, a cena na popyt. Wybieramy zatem 2MNK jako metodę estymacji, w celu wyeliminowania konsekwencji endogeniczności regresora $fish\_price_i$.

Naszym wyborem zmiennej instrumentalnej jest siła wiatru. Powód:

- pogoda wpływa na liczbę kutrów wypływających w morze, a zatem na podaż i cenę (relevance)
- siła wiatru nie ma bezpośredniego wpływu na popyt (exogeneity)

Krok 1: Szacujemy MNK parametry modelu:

$$fish\_price_i = \alpha_0 + \alpha_1 wind_i + u_i$$

i zapisujemy wartości $\widehat{fish\_price_i}$.

Krok 2: Szacujemy MNK:

$$fish_i = \beta_0 + \beta_1 \widehat{fish\_price_i} + \varepsilon_i,$$

a otrzymane oszacowania pochodzą z 2MNK/MZI.

Przykład 13.2. Problem pominiętych zmiennych

Załóżmy, że prawdziwy jest model, w którym płace (w) zależą od wykształcenia (s) oraz inteligencji (q):

$$w_i = \beta_0 + \beta_1 s_i + \beta_2 q_i + \varepsilon_i.$$

Niestety, zmienna q jest nieobserwowalna, więc nie możemy jest włączyć do modelu. A zatem model empiryczny jest następujący:

$$w_i = \beta_0 + \beta_1 s_i + e_i$$

Zauważmy, że inteligencja „ukryła się " w składniku losowym modelu empirycznego: $e_i = \beta_2 q_i + \varepsilon_i$. Jeżeli inteligencja wpływa na wykształcenie, tj. $cov(s_i, q_i) > 0$ (co nie jest mocnym założeniem), to pojawia się problem endogeniczności: $cov(s_i, e_i) > 0$.

Decydujemy się na MZI, a zatem poszukujemy instrumentu z dla zmiennej s, który:

- nie wpływa bezpośrednio na zarobki
- jest skorelowany z wykształceniem
- jest nieskorelowany ze składnikiem losowym (czyli z inteligencją)

Przykład 13.2. Problem pominiętych zmiennych

Jakie są dobre instrumenty dla wykształcenia w modelu płac?

Wielu ekonomistów sugeruje zmienne związane z charakterystyką najbliższej rodziny:

- Wykształcenie matki jest pozytywnie skorelowana z wykształceniem dziecka
  [spełnia warunek *relevance*]
- Wykształcenie matki może jednak być skorelowana ze inteligencją dziecka
  [złamanie warunku *exogeneity*]

Kolejna często wykorzystywana zmienna instrumentalna to liczba rodzeństwa.

- Zazwyczaj posiadanie dużej liczby rodzeństwa wiąże się ze słabszym wykształceniem
  [spełnia warunek *relevance*]
- Liczba rodzeństwa nie wpływa na wrodzoną inteligencję dziecka
  [spełnia warunek *exogeneity*]

Zatem liczba rodzeństwa może być dobrym instrumentem endogenicznej edukacji.

Pytanie: czy znasz inne przykłady dobrych zmiennych instrumentalnych dla wykształcenia?

Równanie płac

- **Card (1995)**

Card (1995) zakłada, że edukacja w równaniu płac jest endogeniczna z powodu pominiętej zdolności lub błędu pomiaru edukacji. Równanie płac oszacowano metodą 2MNK wykorzystując binarny instrument przyjmujący wartość 1, jeżeli w pobliżu miejsca zamieszkania znajduje się college, 0 w przeciwnym przypadku. Występowanie college'u w pobliżu zmniejsza koszty dalszego kształcenia (dojazdy, wynajęcie mieszkania itd.) i w największym stopniu wpływa na decyzję odnośnie kontynuowania edukacji dzieci z biedniejszych rodzin.

Dane dostępne online http://davidcard.berkeley.edu/datan_sets.html.

- **Angrist and Krueger (1991)**

Równanie płac jest szacowane metodą 2MNK z wykorzystaniem kwartału urodzenia jako instrumentu dla endogenicznego wykształcenia. Zaobserwowano, że osoby urodzone w pierwszym półroczu spędzają przeciętnie mniej lat w systemie edukacji niż osoby urodzone w drugim półroczu. Powodem jest prawo o obowiązku szkolnym, gdzie osoby urodzone wcześniej w roku uzyskują wiek pozwalający na zakończenie edukacji w niższej klasie, więc mogą legalnie opuścić szkołę z mniejszą liczbą ukończonych lat.

Dane dostępne online http://economics.mit.edu/faculty/angrist/data1/data/angkru1991.

Przykład 13.3: Empiryczny model płac

Na podstawie danych w pliku `mroz.gdt` oszacowano równanie płac wykorzystując MNK:

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 largecity_i + \beta_4 unemployment_i + \varepsilon_i$$

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 1-428
Zmienna zależna (Y): l_wage

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const         −0,392798        0,202447        −1,940     0,0530    *
  educ           0,107878        0,0144368        7,472     4,56e-013 ***
  exper          0,0156353       0,00402712       3,882     0,0001    ***
  largecity      0,0636633       0,0693048        0,9186    0,3588
  unemployment  −0,00318791      0,0109246       −0,2918    0,7706

Średn.aryt.zm.zależnej  1,190173   Odch.stand.zm.zależnej  0,723198
Suma kwadratów reszt    189,8074   Błąd standardowy reszt  0,669864
Wsp. determ. R-kwadrat  0,150094   Skorygowany R-kwadrat   0,142057
```

Pytanie: Jakie są korzyści z edukacji?

Ze względu na szereg argumentów, że zmienna $educ$ może być endogeniczna, postanowiono zmienić metodę estymacji na 2MNK.

Przykład 13.3: Empiryczny model płac

Wykorzystany instrument dla zmiennej endogenicznej: liczba rodzeństwa ($siblings$)
Uwaga: egzogeniczne regresory są traktowane jako instrumenty!

Krok 1 w 2MNK:

$$educ_i = \alpha_0 + \alpha_1 siblings_i + \alpha_2 exper_i + \alpha_3 largecity_i + \alpha_4 unemployment_i + u_i$$

```
Model 2: Estymacja KMNK, wykorzystane obserwacje 1-753, Zmienna zależna (Y): educ

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const          11,2286         0,299248        37,52     4,12e-174 ***
  siblings        0,0362526      0,0355316        1,020     0,3079
  exper           0,0182131      0,0101621        1,792     0,0735    *
  largecity       0,728212       0,174355         4,177     3,31e-05  ***
  unemployment    0,0323116      0,0267804        1,207     0,2280
```

Krok 2 w 2MNK:

```
Model 3: Estymacja 2MNK, wykorzystane obserwacje 1-428, Zmienna zależna (Y): l_wage
Zmodyfikowane przez instrumenty: educ; Instrumenty: const siblings largecity unemployment exper

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const           1,59902        2,95031         0,5420     0,5881
  educ           −0,0626515      0,252348       −0,2483     0,8040
  exper           0,0146317      0,00487474      3,002      0,0028   ***
  largecity       0,178422       0,187351        0,9523     0,3415
  unemployment    0,00927582     0,0223026       0,4159     0,6777
```

Pytanie:
A jakie wyniki, jeżeli wykorzystano by więcej instrumentów, np. $mothereduc$ i $fathereduc$?

Przykład 13.3: Empiryczny model płac

Wykorzystajmy teraz więcej instrumentów:
liczba rodzeństwa, wykształcenie matki oraz wykształcenie ojca:

```
Model 4: Estymacja 2MNK, wykorzystane obserwacje 1-428
Zmienna zależna (Y): l_wage
Zmodyfikowane przez instrumenty: educ
Instrumenty: const siblings largecity unemployment exper fathereduc
  mothereduc

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const          0,193209        0,396977        0,4867      0,6267
  exper          0,0153400       0,00408779      3,753       0,0002   ***
  largecity      0,0974260       0,0729622       1,335       0,1825
  unemployment   0,000479008     0,0112816       0,04246     0,9662
  educ           0,0577070       0,0325654       1,772       0,0771   *

Średn.aryt.zm.zależnej  1,190173   Odch.stand.zm.zależnej  0,723198
Suma kwadratów reszt    195,2266   Błąd standardowy reszt  0,679359
Wsp. determ. R-kwadrat  0,136956   Skorygowany R-kwadrat   0,128795
F(4, 423)               5,370242   Wartość p dla testu F   0,000316
```

Pytania:

- Jak duże są teraz korzyści z edukacji?
- Czy są statystycznie istotne?
- Jak zwiększenie liczby instrumentów wpływa na błędy standardowe?

Zadania

Zadanie 13.1

Dany jest model liniowy $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, gdzie:

- $\mathbf{y}$ wektor $N \times 1$ dla zmiennej objaśnianej
- $\mathbf{X}$ to macierz $N \times (K+1)$(być może endogenicznych) regresorów

Niech $\mathbf{Z}$ będzie $N \times (L+1)$ macierzą zmiennych instrumentalnych (z $L \geq K$). Oznaczmy

$$\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}; \quad \hat{\mathbf{y}} = \mathbf{P}_{\mathbf{Z}}\mathbf{y} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

Udowodnij, że poniższe trzy definicje estymatora $\widehat{\boldsymbol{\beta}}^{2\text{MNK}}$ są identyczne
(wskazówka: wykorzystaj własności macierzy P_Z):

a. $\widehat{\boldsymbol{\beta}}^{2\text{MNK}}$ to estymator MNK w regresji $\mathbf{y}$ na $\widehat{\mathbf{X}}$

b. $\widehat{\boldsymbol{\beta}}^{2\text{MNK}}$ to estymator MNK w regresji $\hat{\mathbf{y}}$ na $\widehat{\mathbf{X}}$.

c. $\widehat{\boldsymbol{\beta}}^{2\text{MNK}}$ to estymator MZI w regresji $\mathbf{y}$ na $\mathbf{X}$ z $\widehat{\mathbf{X}}$ jako macierz instrumentów.

Zadanie 13.2

W każdym w poniższych przypadków wymyśl zmienne z, które byłyby dobrymi instrumentami dla endogenicznego regresora x (pamiętaj o dwóch warunkach, jakie muszą spełniać instrumenty). Poszukaj odpowiedzi w poniższych artykułach.

a. Jak konkurencja między szkołami (x, mierzone jako liczba szkół w mieście) wpływa na oceny uczniów na końcowym egzaminie (y)?
Hoxby, C.M. (2000) „Does competition among public schools benefit students and taxpayers?" Am. Econ. Review 90: 1209-38

b. Jak liczba policjantów patrolujących ulice (x) wpływa na liczbę przestępstw (y)?
Klick, J. and Tabarrok, A. (2005) „Using terror alert levels to estimate the effect of police on crime"
Journal of Law and Economics 48: 267-279

c. Jak cena ryb (x) wpływa na popyt na ryby (y)?
J. Angrist, K. Graddy, G. Imbens (2000) „The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish" Review of Economic Studies 67: 499-527

d. Jak jakość instytucji (x) wpływa PKB per capita (y)
Acemoglu, D., Johnson, S. and Robinson, J.A. (2001) „The Colonial Origins of Comparative Development: An Empirical Investigation" American Economic Review 91: 1369-1401

Zadanie 13.3

Jednym z ważniejszych zagadnień ekonomii pracy jest wpływ dzietności kobiet na ich podaż pracy. W szczególności interesujące jest o ile spada liczba godzin przepracowanych przez kobietę wraz z urodzeniem kolejnych dzieci. Na podstawie $N = 30\,000$ obserwacji (US Census z 1980) dla zmiennych:

$supply$	liczba przepracowanych tygodni w roku 1979
$morekids = 1$	jeżeli kobieta ma więcej niż dwoje dzieci
$samesex = 1$	jeżeli pierwsza dwójka dzieci jest tej samej płci
age	wiek kobiety
$hispan = 1$	jeżeli kobieta jest latynoską.

metodą MNK oszacowano model regresji i otrzymano następujące wyniki:

$$\widehat{supply}_i = 40 - 6.6 morekids_i - 0.05 age_i + 0.03 hispan_i.$$

a. Czy oszacowania zgodne są z teoretycznymi rozważaniami na temat znaku zależności między zmiennymi? Czy wielkości oszacowanych parametrów są sensowne?

b. W modelu pomijamy ważną (nieobserwowalną) zmienną opisującą preferencje kobiety względem pozostania w domu / pracowania. Rezultatem jest potencjalna endogeniczność zmiennej *morekids*. Czy zmienna *samesex* jest dobrym instrumentem dla zmiennej *morekids*? Uzasadnij.

$$\widehat{morekids}_i = \underset{(0.05)}{0.30} + \underset{(0.02)}{1.30} samesex_i - \underset{(0.08)}{0.16} age_i + \underset{(0.06)}{0.90} hispan_i$$

c. Zaproponuj inną zmienną instrumentalną odpowiednią do użycia w tym modelu.

d. Korzystając z instrumentu $samesex_i$ otrzymaliśmy następujące wyniki estymacji MZI:

$$\widehat{supply}_i = 43.1 - 3.7 morekids_i - 0.03 age_i + 0.06 hispan_i.$$

Czy wartości oszacowań MZI różnią się od wartości oszacowań MNK? O czym może to świadczyć?

Zadanie 13.4

Dany jest model:

$$INFLAT = \beta_0 + \beta_1 MONEY + \beta_3 OUTPUT + \varepsilon$$

gdzie:

$INFLAT$:	tempo wzrostu cen
$MONEY$:	tempo wzrostu podaży pieniądza,
$OUTPUT$:	tempo wzrostu produkcji.

Teoria ekonomii sugeruje, że: $\beta_0 = 0, \beta_1 = 1$ oraz $\beta_2 = -1$.

Korzystając z danych dla 76 krajów i 1995 roku, zawartych w pliku `brumm.gdt`:

a. Oszacuj MNK parametry modelu i zweryfikuj:
   i. Mocną hipotezę: $\beta_0 = 0, \beta_1 = 1$ i $\beta_2 = -1$.
   ii. Słabszą hipotezę: $\beta_1 = 1$ *i* $\beta_2 = -1$.

b. Zmienna $OUTPUT$ może być endogeniczna. Zaproponowano cztery zmienne instrumentalne:

$INITIAL$:	początkowy poziom PKB
$SCHOOL$:	miara poziomu wykształcenia społeczeństwa
INV:	inwestycje jako proporcja PKB
$POPRATE$:	średni przyrost populacji

   Uzasadnij i sprawdź, które z powyższych zmiennych są dobrymi instrumentami dla $OUTPUT$.

c. Wykorzystaj pojedyncze oraz wszystkie cztery instrumenty w 2MNK estymacji modelu. Czy otrzymane oszacowania różnią się od tych w (a)?

d. Zweryfikuj ponownie mocną i słabszą hipotezę z (a) wykorzystując oszacowania IV. Czy wnioski płynące z tej weryfikacji różnią się od tych otrzymanych w (a)?

Zadanie 13.5

Analizujemy model podaży kurczaków, które przez amerykańskie ministerstwo rolnictwa nazywane są „broilers". Dane roczne z okresu 1950-2001 znajdują się w pliku `newbroiler.gdt`. Na podstawie danych z okresu 1960-1999 oszacuj parametry modelu:

$$\ln(QPROD_t) = \beta_0 + \beta_1 \ln(P_t) + \beta_2 \ln(PF_t) + \beta_3 TIME_t + \ln(QPROD_{t-1}) + \varepsilon_t,$$

gdzie:

$QPROD_t$:	produkcja kurczaków
P_t:	indeks cen kurczaków
PF_t:	indeks cen paszy dla kurczaków

Potencjalne zmienne instrumentalne dla zmiennej P_t to:

$\ln(Y_t)$:	PKB per capita (logarytm)
$\ln(PB_t)$:	indeks cen wołowiny (logarytm)
POPGRO:	roczny procentowy przyrost populacji
$\ln(P_{t-1})$:	ceny kurczaków (logarytm, opóźnienie o 1 okres)
$\ln(EXPTS_t)$:	eksport kurczaków (logarytm)

a. Czy zmienna $\ln(QPROD_{t-1})$ jest endo czy egzogeniczna?

b. Oszacuj model podaży MNK. Przedyskutuj wyniki estymacji.

c. Oszacuj model podaży MZI z wszystkimi instrumentami. Porównaj wyniki z (b).

d. Wypróbuj inne kombinacje instrumentów (jeden, dwa, inne dwa, itd.), żeby otrzymać estymatory z jak najmniejszą wariancją.

Zadanie 13.6

W pliku `endSim.gdt` znajdują się sztucznie wygenerowane zmienne z następującego DGP:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \quad \varepsilon_i \sim \mathcal{N}(0,1);$$

gdzie $cor(x_i, \varepsilon_i) = 0.6$, $\beta_0 = 0$ oraz $\beta_1 = 1$.

a. Oszacuj MNK parametry modelu i dokonaj weryfikacji hipotezy $\beta_1 = 1$.

Ze względu na endogeniczność regresora, zaproponowano stworzono trzy instrumenty:

z_{1i}: $cor(z_{1i}, x_i) = 0.3$, $cor(z_{1i}, \varepsilon_i) = 0$

z_{2i}: $cor(z_{2i}, x_i) = 0.5$, $cor(z_{1i}, \varepsilon_i) = 0$

z_{3i}: $cor(z_{3i}, x_i) = 0.5$, $cor(z_{1i}, \varepsilon_i) = 0.3$

b. Oszacuj MZI parametry modelu, korzystając tylko z jednego instrumentu.
   - który instrument jest najlepszy?
   - który z instrumentów spełnia warunki egzogeniczności, a który skorelowania?
   - zweryfikuj hipotezę $\beta_1 = 1$

c. Oszacuj 2MNK parametry modelu wykorzystując 2 lub 3 instrumenty. Który z modeli z punktów **a.**, **b.** i **c.** jest twoim zdaniem najlepszy? Odpowiedź uzasadnij.

Temat 14

Testy w metodzie zmiennych instrumentalnych

Katarzyna Bech-Wysocka i Michał Rubaszek

- Warunek egzogeniczności: test J
- Słaby vs silny instrument
- Test Hausmana

Etapy budowy modelu ekonometrycznego

1. Postawienie hipotezy badawczej
2. Wybór postaci funkcyjnej
3. Zebranie danych
4. Estymacja
5. **Weryfikacja**
6. Zastosowanie

Weryfikacja w MZI

Modele oszacowane metodą zmiennych instrumentalnych podlegają identycznej weryfikacji jak te oszacowane MNK tj. powinniśmy zbadać o:

- poprawność specyfikacji
- odpowiednie wartości / istotność parametrów
- własności składnika losowego (homoskedastyczność, brak autokorelacji, itd.)

Różnicą między MZI i MNK jest to, że możemy przeprowadzić dodatkowe testy związane z endogenicznością regresorów:

- sprawdzenie, czy regresory rzeczywiście są endogeniczne (i stosowanie MZI ma sens)
- spełnienie przez instrumenty warunku egzogeniczności (*exogeneity*)
- spełnienie przez instrumenty warunku skorelowania (*relevance*)

Warunek egzogeniczności

Jednym z dwóch warunków, które muszą spełniać dobre zmienne instrumentalne, jest:

Egzogeniczność: zmienne $\mathbf{z}_l$ nie są skorelowane z $\boldsymbol{\varepsilon}$, tj. $E(\mathbf{Z}'\boldsymbol{\varepsilon}) = \mathbf{0}$

Istnieją dwa podejścia do weryfikacji, czy ten warunek jest spełniony

Metoda nieformalna. W przypadku modelu nadmiernie zidentyfikowanego, tj. gdy jest więcej instrumentów niż regresorów ($L > K$), możliwe jest otrzymanie kilku zgodnych estymatorów IV z różnych kombinacji instrumentów. Jeżeli uzyskane w ten sposób oszacowania są podobne, to instrumenty są egzogeniczne. Jeżeli natomiast oszacowania różnią się w zależności od zbioru wybranych instrumentów może to oznaczać, że niektóre lub wszystkie instrumenty nie są egzogeniczne.

Metoda formalna: polega na przeprowadzeniu formalnego testu, który jest nazywany testem restykcji nałożonych na nadmierne warunki identyfikujące (*overidentifying restrictions*), lub krócej jako J-test Sargana. Hipoteza zerowa testu:

$$H_0: E(z_{li}\varepsilon_i) = 0 \text{ dla } l = 1,2,\dots,L$$

implikuje egzogeniczność wszystkich instrumentów, zaś hipoteza alternatywna mówi, że co najmniej jeden z instrumentów jest skorelowany ze składnikiem losowym modelu.

Warunek egzogeniczności: J-test

Intuicja testu J Sargana

Oszacowania parametrów, których liczba w modelu $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ wynosi $K + 1$, może zostać uzyskana jako rozwiązanie następujących układów równań

MNK: $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ → $K + 1$ równań i $K + 1$ niewiadomych

MZI: $\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ → $L + 1$ równań i $K + 1$ niewiadomych

Oznacza to, że w przypadku nadmiernych warunków identyfikujących ($L > K$) nie istnieją takie wartości $\widehat{\boldsymbol{\beta}}$, dla których wyrażenie $\mathbf{Z}'\hat{\boldsymbol{\varepsilon}}$ jest dokładnie równe $\mathbf{0}$ (w przypadku MNK $\mathbf{X}'\hat{\boldsymbol{\varepsilon}}$ jest zawsze równe zeru). Możemy natomiast sprawdzić na ile odległości od zera, tj. $\mathbf{u} = \mathbf{Z}'\hat{\boldsymbol{\varepsilon}} - \mathbf{0}$ są istotne, a na ile nie. Statystyka J jest ważoną sumą kwadratów:

$$J = \mathbf{u}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{u} \overset{H_0}{\sim} \chi^2(L - K)$$

Prostsza procedura testu J:

- Oszacuj model MZI i wyznacz reszty $\hat{\varepsilon}_i^{MZI}$
- Dla regresji pomocniczej, w której $\hat{\varepsilon}_i^{MZI}$ są objaśniane przez wszystkie instrumenty, tj. egzogeniczne regresory ($z_{li} = x_{li}$ dla $l \le L_1$) i instrumenty dla zmiennych endo (z_{li} dla $l > L_1$):

  $$\hat{\varepsilon}_i^{MZI} = \gamma_0 + \sum_{l=1}^{L_1}\gamma_l x_{li} + \sum_{l=L_1+1}^{L}\gamma_l z_{li} + u_i,$$

  oblicz wartość statystyki F dla restrykcji $\gamma_{L_1+1} = \gamma_{L_1+1} = \cdots = \gamma_L = 0$.
- Przy prawdziwości hipotezy zerowej (egzogeniczność zmiennych instrumentalnych):

$$J = (L - K)F \overset{H_0}{\sim} \chi^2(L - K)$$

Przykład 14.1: Warunek egzogeniczności

Na podstawie danych z pliku `mroz.gdt` oszacowano 2MNK dwa modele równania płac (por. Przykład 13.3)

```
Model A: Estymacja 2MNK, wykorzystane obserwacje 1-428, Zmienna zależna (Y): l_wage
Zmodyfikowane przez instrumenty: educ; Instrumenty: const siblings largecity unemployment exper

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------
  const          1,59902         2,95031          0,5420     0,5881
  educ          −0,0626515       0,252348        −0,2483     0,8040
  exper          0,0146317       0,00487474       3,002      0,0028   ***
  largecity      0,178422        0,187351         0,9523     0,3415
  unemployment   0,00927582      0,0223026        0,4159     0,6777
```

```
Model B: Estymacja 2MNK, wykorzystane obserwacje 1-428, Zmienna zależna (Y): l_wage
Zmodyfikowane przez instrumenty: educ; Instrumenty: const siblings largecity unemployment exper fathereduc mothereduc

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------
  const          0,193209        0,396977         0,4867     0,6267
  exper          0,0153400       0,00408779       3,753      0,0002   ***
  largecity      0,0974260       0,0729622        1,335      0,1825
  unemployment   0,000479008     0,0112816        0,04246    0,9662
  educ           0,0577070       0,0325654        1,772      0,0771   *
```

Pytania:

- czy oszacowania są podobne czy różne w zależności od zbioru instrumentów?
- Czy mamy podstawy wątpić w egzogeniczność instrumentów?

Przykład 14.1: Warunek egzogeniczności

```
Model B: Estymacja 2MNK, wykorzystane obserwacje 1-428, Zmienna zależna (Y): l_wage
Zmodyfikowane przez instrumenty: educ; Instrumenty: const siblings largecity unemployment exper fathereduc mothereduc

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------
  const          0,193209        0,396977         0,4867     0,6267
  exper          0,0153400       0,00408779       3,753      0,0002   ***
  largecity      0,0974260       0,0729622        1,335      0,1825
  unemployment   0,000479008     0,0112816        0,04246    0,9662
  educ           0,0577070       0,0325654        1,772      0,0771   *
```

Dla modelu B, liczba instrumentów ($L = 6$) jest większa niż liczba regresorów ($K = 4$). Mamy zatem dwa dodatkowe warunki identyfikujące. Przeprowadzenie testu J prowadzi do następujących wyników:

```
Test Sargana - nadmiernej identyfikacji -
  Hipoteza zerowa: wszystkie instrumenty są ważne - uzasadnione
  Statystyka testu: LM = 0,450248
  z wartością p = P(Chi-kwadrat(2) > 0,450248) = 0,798417
```

Pytania:

- Jak zinterpretować wyniki tego testu?
- Czy wyniki są zgodne z konkluzjami płynącymi z zastosowania metody nieformalnej?
- Czy potrafisz uzyskać powyższe wyniki przeprowadzając regresję pomocniczą?

Warunek skorelowania (relevance)

Problem słabych instrumentów

Drugim warunkiem, które muszą spełniać dobre zmienne instrumentalne, jest:

Skorelowanie (relevance): zmienne instrumentalne $\mathbf{z}_l$ są skorelowane z $\mathbf{x}_k$

- W celu ilustracji, rozważmy najprostszy model liniowy z jedną zmienną objaśniającą:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

oraz jedną zmienną instrumentalną z_i.

- Jeżeli $cor(z_i, x_i)$ jest bliska 0, to z_i nazywamy **słabym instrumentem**. Dlaczego?

Estymator MZI parametru β_1: $\widehat{\beta_1}^{MZI} = \frac{\sum(z_i-\bar{z})(y_i-\bar{y})}{\sum(z_i-\bar{z})(x_i-\bar{x})} = \frac{\frac{1}{n}\sum(z_i-\bar{z})(y_i-\bar{y})}{\frac{1}{n}\sum(z_i-\bar{z})(x_i-\bar{x})}$

Zauważ, że WLLN implikuje: $\frac{1}{n}\sum(z_i-\bar{z})(y_i-\bar{y}) \xrightarrow{p} cov(z,y)$

$\frac{1}{n}\sum(z_i-\bar{z})(x_i-\bar{x}) \xrightarrow{p} cov(z,x)$

zaś twierdzenie Słuckiego: $\operatorname{plim} \widehat{\beta_1}^{MZI} = \frac{\operatorname{plim}\frac{1}{n}\sum(z_i-\bar{z})(y_i-\bar{y})}{\operatorname{plim}\frac{1}{n}\sum(z_i-\bar{z})(x_i-\bar{x})}$

A zatem, jeżeli $cov(z,x) \approx 0$, wtedy $\widehat{\beta_1}^{MZI}$, bo w mianowniku jest wartość zbliżona do 0.

Przykład 14.2. Słabe instrumenty

Rozważmy model wyjaśniający wpływ palenia papierosów w ciąży na wagę urodzeniową noworodka:

$$\ln(bwght) = \beta_0 + \beta_1 packs + \varepsilon,$$

gdzie $packs$ to liczba paczek papierosów wypalanych dziennie przez matkę

- Podejrzewamy, że zmienna $packs$ może być endogeniczna (dlaczego?) i używamy zmiennej $cigprice$, średniej ceny paczki papierosów w rejonie zamieszkania, jako instrumentu.
- Zakładamy, że zmienna $cigprice$ jest nieskorelowana ze składnikiem losowym modelu (po co nam takie założenie? czy uważasz, że jest w tym modelu spełnione?).

W pierwszym kroku 2MNK otrzymujemy:

$$\widehat{packs} = 0.067 + 0.0003 cigprice,$$

czyli bardzo mały (nieistotny) wpływ.

Wykorzystując $cigprice$ jako instrument, w drugim kroku 2MNK otrzymujemy:

$$\ln(\widehat{bwght}) = 4.45 + 2.99 packs.$$

Okazuje się że oszacowanie jest złego znaku i dużej wartości. Dlaczego?

Warunek skorelowania: weryfikacja

- Gdy w modelu występuje tylko jeden endogeniczny regresor, warunek skorelowania można sprawdzić poprzez oszacowanie parametrów modelu z pierwszego kroku 2MNK i obliczyć wartość statystyki F dla testu łącznej istotności całego modelu.
- Jeżeli wartość tej statystyki testowej **przekracza 10** ($F^{obl} > 10$), to nie dowodów, że instrumenty są słabe. Dla wartości mniejszych od 10, może występować problem słabych instrumentów.
- W przypadku nadmiernej liczby warunków identyfikujących ($L > K$), możemy mieć kilka mocnych i kilka słabych instrumentów. W takim przypadku warto wykorzystać tylko mocne instrumenty i nie uwzględniąć tych słabych.
- Co zrobić gdy dysponujemy tylko słabymi instrumentami?
  - Postaraj się znaleźć mocny instrument
  - Zrezygnuj z 2MNK - zmień metodę estymacji
  - Może się okazać, że nie można oszacować modelu

Przykład 14.3. Moc instrumentów w równaniu płac

Na podstawie danych w pliku `mroz.gdt` w równaniu płac wykorzystujemy liczbę rodzeństwa, wykształcenie matki i wykształcenie ojca jako instrumenty.

- Pierwszy krok w procedurze 2MNK:

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 1-753
Zmienna zależna (Y): educ

                współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------
  const          8,12227         0,330599        24,57      4,02e-098 ***
  largecity      0,439333        0,153597         2,860     0,0044    ***
  unemployment   0,0246142       0,0233442        1,054     0,2920
  mothereduc     0,189419        0,0259241        7,307     7,05e-013 ***
  fathereduc     0,175124        0,0246784        7,096     2,99e-012 ***
  exper          0,0310054       0,00889130       3,487     0,0005    ***
  siblings       0,0139480       0,0309872        0,4501    0,6528

średn.aryt.zm.zależnej  12,28685   Odch.stand.zm.zależnej  2,280246
Suma kwadratów reszt    2862,283   Błąd standardowy reszt  1,958786
Wsp. determ. R-kwadrat  0,267966   Skorygowany R-kwadrat   0,262078
F(6, 746)               45,51298   Wartość p dla testu F   1,51e-47
Logarytm wiarygodności −1571,205   Kryt. inform. Akaike'a  3156,410
Kryt. bayes. Schwarza   3188,778   Kryt. Hannana-Quinna    3168,880
```

Wartość statystyki testowej F

Pytanie: czy powinniśmy się martwić problemem słabych instrumentów w tym modelu?

Test Hausmana na endogeniczność regresora

- Przy spełnionych założeniach **A1-A4**, estymator MNK jest BLUE, a zatem estymator MNK jest bardziej efektywny niż estymator MZI.
- Nie chcemy stosować MZI, jeżeli nie ma takiej potrzeby (gdy regresory są egzogeniczne).

Jak sprawdzić endogeniczność regresorów?

- Możemy przeprowadzić test Hausmana, dla którego zespół hipotez jest następujący:

$H_0: cov(x, \varepsilon) = 0$ [stosujemy MNK]

$H_1: cov(x, \varepsilon) \neq 0$ [stosujemy MZI]

Intuicja testu Hausmana:

Jeżeli regresory są egzogeniczny, czyli H_0 jest prawdziwa, to różnica między oszacowaniami MNK i MZI jest „mała". A zatem możemy zbudować test, który sprawdza jak bardzo ważona suma kwadratów różnic oszacowań parametrów jest odległa od zera. W szczególności, **statystyka testu Hausmana to:**

$$H = \left(\hat{\boldsymbol{\beta}}^{\mathrm{MZI}} - \hat{\boldsymbol{\beta}}^{\mathrm{MNK}}\right)\left[Var(\hat{\boldsymbol{\beta}}^{\mathrm{MZI}}) - Var(\hat{\boldsymbol{\beta}}^{\mathrm{MNK}})\right]^{-1}\left(\hat{\boldsymbol{\beta}}^{\mathrm{MZI}} - \hat{\boldsymbol{\beta}}^{\mathrm{MNK}}\right) \overset{H_0}{\sim} \chi^2_M,$$

gdzie M to liczba regresorów, które zostały uznane za endogeniczne w MZI.

Test Hausmana

Test Hausmana dla modelu $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, w którym przypuszczamy, że M regresorów jest endogenicznych, możemy przeprowadzić też w następujący krokach:

1. Dla każdego endogenicznego regresora x_{k_m}, gdzie $m = 1,2,\dots,M$ oszacuj parametry modelu:
$$\mathbf{x}_{k_m} = \mathbf{Z\gamma}_m + \mathbf{v}_m$$
2. Oblicz reszty $\hat{\mathbf{v}}_m$ oraz stwórz macierz $\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1\ \hat{\mathbf{v}}_2 \dots \hat{\mathbf{v}}_{\mathrm{M}}]$ o wymiarach $\mathrm{N} \times M$
3. Oszacuj MNK parametry modelu rozszerzonego:
$$\mathbf{y} = \mathbf{X\beta} + \hat{\mathbf{V}}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$
4. Dokonaj weryfikacji hipotezy o egzogeniczności regresorów:
$$H_0: \boldsymbol{\delta} = \mathbf{0}$$

Odrzucenie powyższej hipotezy oznacza, że co najmniej jeden regresor jest endogeniczny i powinniśmy stosować MZI

UWAGA: W większości przypadków rozważamy endogeniczność jednego regresora ($M = 1$). Hipoteza zerowa może być zatem weryfikowana za pomocą testu t-Studenta.

Przykład 14.4: Test Hausmana w równaniu płac

Dla modelu z pliku mroz.gdt w równaniu płac, gdzie w MZI wykorzystujemy liczbę rodzeństwa, wykształcenie matki i wykształcenie ojca jako instrumenty.

```
Model A: Estymacja 2MNK, wykorzystane obserwacje 1-428, Zmienna zależna (Y): l_wage
Zmodyfikowane przez instrumenty: educ; Instrumenty: const siblings largecity unemployment exper fathereduc
mothereduc

               współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------
  const         0,193209         0,396977        0,4867      0,6267
  exper         0,0153400        0,00408779      3,753       0,0002   ***
  largecity     0,0974260        0,0729622       1,335       0,1825
  unemployment  0,000479008      0,0112816       0,04246     0,9662
  educ          0,0577070        0,0325654       1,772       0,0771   *
```

Wyniki testu Hausmana są następujące:

```
Hipoteza zerowa: Estymator MNK jest zgodny
Asymptotyczna statystyka testu: Chi-kwadrat(1) = 3,11851
z wartością p = 0,0774076
```

Pytania:

- Czy zastosowanie metody IV było w tym modelu konieczne?
- Czy potrafisz uzyskać dany wynik za pomocą metody opisanej na poprzednim slajdzie?

Zadania

Zadanie 14.1

Dany jest model:

$$INFLAT = \beta_0 + \beta_1 MONEY + \beta_3 OUTPUT + \varepsilon$$

gdzie:

$INFLAT$:	tempo wzrostu cen
$MONEY$:	tempo wzrostu podaży pieniądza,
$OUTPUT$:	tempo wzrostu produkcji.

Teoria ekonomii sugeruje, że: $\beta_0 = 0, \beta_1 = 1$ oraz $\beta_2 = -1$.

Wykorzystaj dane `brumm.gdt` z 1995 dla 76 krajów. Wybierz jedną zmienną instrumentalną z pozostałych zmiennych ($INITIAL, SCHOOL, INV, POPRATE$) oraz:

a. Przeprowadź test Hausmana do zbadania endogeniczności zmiennej $OUTPUT$.
   Jakie wyciągasz wnioski?
b. Sprawdź warunek egzogeniczności dla instrumentów z punktu **a.**
c. Zbadaj warunek skorelowania dla instrumentów wykorzystując test J
d. Powtórz **a.-c.** dla modelu z dwoma, trzema i czterema instrumentami.
   Jaki zbiór instrumentów jest według Ciebie najlepszy?

Zadanie 14.2

Analizujemy model podaży kurczaków, które przez amerykańskie ministerstwo rolnictwa nazywane są „broilers". Dane roczne z okresu 1950-2001 znajdują się w pliku `newbroiler.gdt`.
Na podstawie danych z okresu 1960-1999 oszacuj parametry modelu:

$$\ln(QPROD_t) = \beta_0 + \beta_1 \ln(P_t) + \beta_2 \ln(PF_t) + \beta_3 TIME_t + \ln(QPROD_{t-1}) + \varepsilon_t,$$

gdzie:

$QPROD_t$:	produkcja kurczaków
P_t:	indeks cen kurczaków
PF_t:	indeks cen paszy dla kurczaków

Dla zmiennej P_t wykorzystaj następujące instrumenty:

$\ln(Y_t)$:	PKB per capita (logarytm)
$\ln(PB_t)$:	indeks cen wołowiny (logarytm)
POPGRO:	roczny procentowy przyrost populacji
$\ln(P_{t-1})$:	ceny kurczaków (logarytm, opóźnienie o 1 okres)
$\ln(EXPTS_t)$:	eksport kurczaków (logarytm)

a. Zbadaj endogeniczność zmiennej $\ln(P_t)$ wykorzystując test Hausmana.
   Powtórz test dla różnych zbiorów instrumentów. Czy wnioski zawsze są takie same?
b. Sprawdź, które instrumenty są słabe. Zbuduj model tylko mocnymi instrumentami
c. Sprawdź czy wykorzystywane instrumenty są egzogeniczne.
d. Czy masz teoretyczne wątpliwości co do jakości instrumentów?

Zadanie 14.3

W zbiorze `wine_aus.gdt` znajdują się następujące zmienne opisującymi rynek konsumpcji i produkcji wina w Australii w latach 1955-1974:

- Q: spożycie wina na osobę (w litrach),
- Pw: cena wina (deflowana CPI),
- Pb: cena piwa (deflowana CPI),
- A: średnie wydatki na reklamę wina ($/osobę),
- Y: średni dochód ($/osobę),
- S: koszty magazynowania (indeks).

Rozważmy model popytu na wino:

$$\ln(Q_t) = \beta_0 + \beta_1 \ln(Pw_t) + \beta_2 \ln(Pb_t) + \beta_3 \ln(Y_t) + \beta_4 \ln(A_t) + \varepsilon_t.$$

a. Oszacuj parametry modelu popytu na wino wykorzystując koszty magazynowania jako instrument dla endogenicznej ceny wina.

b. Sprawdź, czy koszty magazynowania są dobrym instrumentem.

c. Zweryfikuj czy cena wina rzeczywiście jest endogeniczna w równaniu popytu.

Zadanie 14.4

W zbiorze `wine_aus.gdt` znajdują się następujące zmienne opisującymi rynek konsumpcji i produkcji wina w Australii w latach 1955-1974:

- Q: spożycie wina na osobę (w litrach),
- Pw: cena wina (deflowana CPI),
- Pb: cena piwa (deflowana CPI),
- A: średnie wydatki na reklamę wina ($/osobę),
- Y: średni dochód ($/osobę),
- S: koszty magazynowania (indeks).

Rozważmy model podaży wino:

$$\ln(Q_t) = \alpha_0 + \alpha_1 \ln(Pw_t) + \alpha_2 \ln(S_t) + u_t.$$

a. Oszacuj parametry modelu podaży wina wykorzystując cenę piwa, wydatki na reklamę oraz dochód jako instrumenty dla endogenicznej ceny wina.

b. Sprawdź czy zaproponowane instrumenty spełniają warunki skorelowania (relevance) i egzogeniczności (exogeneity). Wybierz najlepszą kombinację instrumentów.

c. Zweryfikuj czy cena wina rzeczywiście jest endogeniczna w równaniu podaży.

Zadanie 14.5

[Kontynuacja zadania 13.3]. Na podstawie $N = 30\,000$ obserwacji (US Census z 1980) dla:

$supply$	liczba przepracowanych tygodni w roku 1979
$morekids = 1$	jeżeli kobieta ma więcej niż dwoje dzieci
$samesex = 1$	jeżeli pierwsza dwójka dzieci jest tej samej płci
age	wiek kobiety
$hispan = 1$	jeżeli kobieta jest latynoską
siblings2=1	jeżeli kobieta ma więcej niż dwoje rodzeństwa.

parametry modelu:

$$supply_i = \beta_0 + \beta_1 morekids_i + \beta_2 age_i + \beta_3 hispan_i + \varepsilon_i.$$

Podejrzewamy, że zmienna $morekids$ jest endogeniczna i wykorzystujemy zmienne:
$samesex$ oraz $siblings2$ jako instrumenty

Zadanie 14.5 cd

Regresja pierwszego kroku 2MNK:

$$\widehat{morekids}_i = \underset{(0.07)}{0.20} + \underset{(0.04)}{1.15}\, samesex_i + \underset{(0.03)}{0.90}\, siblings2_i - \underset{(0.06)}{0.13}\, age_i + \underset{(0.08)}{0.30}\, hispan_i,\ R^2 = 0.30$$

Regresje drugiego kroku 2MNK:

Model 1. instrument, *samesex:*

$$\widehat{supply}_i = 43.1 - 3.7 morekids_i - 0.03 age_i + 0.06 hispan_i.$$

Model 2. instrument, *siblings2:*

$$\widehat{supply}_i = 47.3 - 3.5 morekids_i - 0.028 age_i + 0.07 hispan_i.$$

Model 3. instrumenty, *samesex* i *siblings2:*

$$\widehat{supply}_i = 42.7 - 3.8 morekids_i - 0.032 age_i + 0.064 hispan_i.$$

Wyniki testu Hausmana dla modelu 3: $\chi^2(1) = 12.376$.

Odpowiedz na pytania:

a. Czy zmienne *samesex* i *siblings2* są słabymi czy mocnymi instrumentami*?*
b. Czy instrumenty *samesex* i *siblings2* są egzogeniczne?
c. Czy zmienna $morekids_i$ jest endogeniczna?

Zadanie 14.6

W pliku `endSim.gdt` znajdują się sztucznie wygenerowane zmienne z następującego DGP:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \quad \varepsilon_i \sim \mathcal{N}(0,1);$$

gdzie $cor(x_i, \varepsilon_i) = 0.6$, β_0 =0 oraz $\beta_1 = 1$. Ze względu na endogeniczność regresora stworzono trzy instrumenty:

z_{1i}: $cor(z_{1i}, x_i) = 0.3$, $cor(z_{1i}, \varepsilon_i) = 0$

z_{2i}: $cor(z_{2i}, x_i) = 0.5$, $cor(z_{1i}, \varepsilon_i) = 0$

z_{3i}: $cor(z_{3i}, x_i) = 0.5$, $cor(z_{1i}, \varepsilon_i) = 0.3$

b. Oszacuj MZI parametry modelu, korzystając tylko z jednego instrumentu.
   - sprawdź endogeniczność zmiennej x za pomocą testu Hausmana
   - czy instrument jest słaby czy mocny?
   - Dlaczego nie można przeprowadzić testu egzogeniczność?

c. Oszacuj MZI parametry modelu wykorzystując 2 instrumenty:
   - sprawdź endogeniczność zmiennej x za pomocą testu Hausmana
   - czy instrument jest słaby czy mocny?
   - czy instrumenty są egzogeniczne?