Ekonometria
– notatki z wykładu –

dr Marcin Pitera
Uniwersytet Jagielloński

29 listopada 2020

Spis treści

1 Wstęp

Kurs ten zakłada znajomość podstawowych pojęć z zakresu analizy, rachunku prawdopodobieństwa i algebry liniowej. W szczególności odnosi się do notacji, pojęć oraz podstawowych twierdzeń i własności związanych z funkcjami wielu zmiennych, pochodnych cząstkowych, całek, zmiennych losowych, rozkładów wielowymiarowych, niezależności, bezwarunkowych i warunkowych wartości oczekiwanych, wariancji i kowariancji wektorów losowych, rozkładu normalnego i χ^2, mnożenia i odwracania macierzy, rzędu macierzy, itd. Kurs ten opiera się przede wszystkim na książkach [Hay00] oraz [Gre18].

Uwaga: przykłady będą pokazane i omówione na ćwiczeniach; wykład skupia się w głównej mierze na części teoretycznej kursu. Bardzo dobre opracowanie praktycznych aspektów modelu regresji liniowej można znaleźć w książce [Far15]. Warto jednak wspomnieć, że podsumowanie podstawowych (praktycznych) metod weryfikacji założeń, identyfikacji nietypowych obserwacji, czy doboru właściwego modelu dostępne jest w Rozdziale 5, który ma bardziej praktyczny charakter. W Rozdziale 1.1 omawiamy ogólny schemat budowy modelu statystycznego, co również ma bardziej praktyczny charakter, a przynajmniej nie jest ściśle związane z matematyczną (właściwą) częścią tego wykładu.

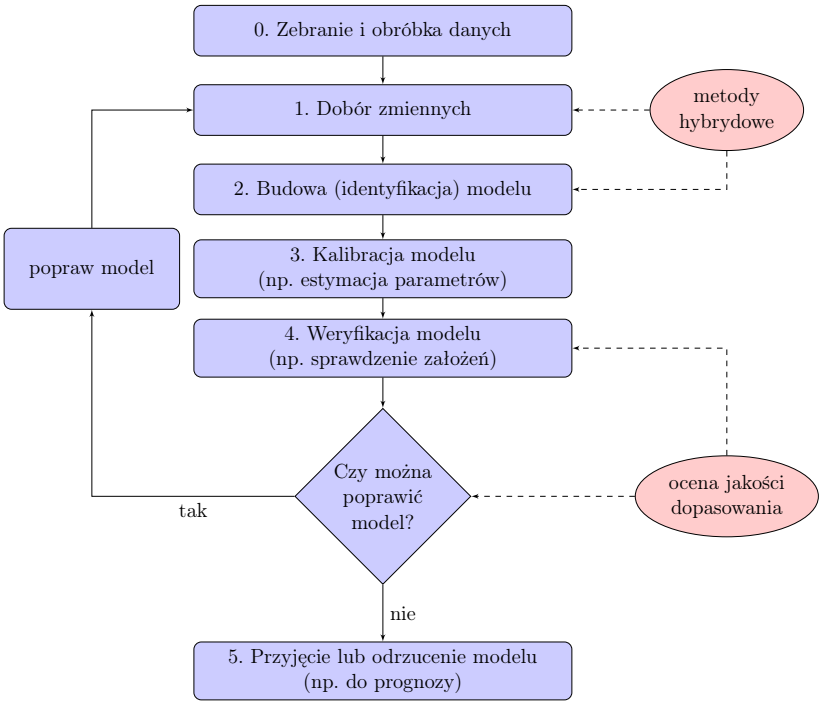
1.1 Budowa modelu statystycznego

Model regresji liniowej jest jednym z wielu modeli statystycznych, które mogą służyć do wyjaśnienia zjawisk, czy prognozy przyszłych wartości. Ogólny przykładowy schemat budowy modeli statystycznych przedstawiony jest na Rysunku 1 i składa się z kroków takich, jak: (0.) zebranie i obróbka danych; (1.) dobór zmiennych; (2.) budowa modelu; (3.) kalibracja modelu; (4.) weryfikacja modelu; (5.) przyjęcie lub odrzucenie modelu. W kursie tym będziemy omawiać głównie kroki (2.), (3.) oraz (4.) – w odniesieniu do modelu regresji liniowej – skupiając się na aspekcie matematycznym problemu; przedstawimy też wybrane techniki związane z (1). Omówmy teraz pokrótce każdy z kroków w nawiązaniu do modelu regresji liniowej:

0. **Zebranie i obróbka danych.** Aby móc zbudować model statystyczny oparty o dane, potrzebujemy je zebrać i obrobić. Ta część budowy modelu często nie jest związana z jego matematyczną częścią, choć warto już na tym etapie popatrzeć holistycznie na całość (statystycznie poprawne zaplanowanie eksperymentu wpłynie na jakość danych, poprawnie dodane znaczniki czasowe mogą pomóc w dopasowaniu modeli dynamicznych, itd.). W praktyce krok ten zajmuje najwięcej – około 80% – czasu poświęconego na budowę modelu. Zawiera on w sobie etapy takie jak czyszczenie danych, właściwa reprezentacja danych czy uzupełnienie brakujących informacji. Często dane wpływają do modelu w sposób dynamiczny, co wymaga ich zautomatyzowanej obróbki.

1. **Dobór zmiennych.** Mając interesujące nas dane musimy określić, jakie zjawisko chcemy objaśnić oraz jakich danych chcemy użyć do jego objaśnienia. W rozważanym przez nas modelu dane przekształcimy w tablicę, w której wiersze będą odpowiadać różnym obserwacjom, a kolumny różnym zmiennym (charakterystykom). W tablicy tej wyróżniamy pierwszą kolumnę która będzie odpowiadała za obserwowalną **zmienną objaśnianą**, której wartości będziemy chcieli wytłumaczyć poprzez pozostałe kolumny, tj. **zmienne objaśniające**.

2. **Budowa modelu.** W kroku tym określa się postać strukturalną modelu, czyli związek między zmienną objaśnianą, a zmiennymi objaśniającymi. Model regresji liniowej będzie zakładał, że zmienną objaśnianą będzie można przedstawić jako sumę liniowej kombinacji zmiennych objaśniających oraz nieobserwowalnego czynnika losowego modelu; omówimy to dokładniej w Rozdziale 1.3. Warto zaznaczyć, że kroki (1.) oraz (2.) są ze sobą często ściśle powiązane – właściwy



Rysunek 1: *Przykładowy uproszczony schemat budowy modelu statystycznego. W modelu regresji liniowej zadana jest postać strukturalną modelu (w 2.), ale poprzez dobór zmiennych i ich modyfikację (w 1.) uzyskujemy dużą swobodę; możemy również dokonać przekształceń zmiennej objaśnianej (w 2.). W praktyce najwięcej – około 80% czasu – zajmuje początkowy krok (0.).*

dobór zmiennych skutkuje lepiej zbudowanym modelem. Istnieje wiele metod hybrydowych, które jednocześnie budują i dobierają zmienne do modelu, czy dokonują właściwej selekcji zmiennych (ang. *feature extraction* lub *feature selection*). Dotyczy to również technik służących do transformacji zmiennych, np. poprzez obłożenie ich wartości funkcją nieliniową. Oczywiście model musi być tak zbudowany, aby spełnione były jego założenia. W tym kroku można dokonać wstępnej tego weryfikacji; krok (4.) również sprawdza zasadność założeń. Analiza budowy modelu regresji liniowej, zbioru założeń, czy jego (matematycznych) własności jest podstawowym tematem tego kursu. Informacje na ten temat dostępne są w Rozdziale 2 oraz Rozdziale 4. Opis wybranych alternatywnych modeli powiązanych z regresją liniową można znaleźć w Rozdziale 3.

3. **Kalibracja modelu.** Mając daną ogólną postać strukturalną modelu chcemy dobrać parametry modelu tak, aby zapewnić jego jak najlepsze dopasowanie. Wprowadza się tutaj zazwyczaj tzw. *funkcję celu* (ang. *objective function*), którą należy zminimalizować. W kursie tym będziemy bazować na *metodzie najmniejszy kwadratów* (ang. *least squares*) i badać strukturę powiązanych estymatorów parametrów modelu. Definicję podstawowych estymatorów najmniejszych kwadratów oraz ich własności są omawiane w Rozdziale 2 oraz Rozdziale 4.

4. **Weryfikacja modelu modelu.** Mając dany skalibrowany model chcielibyśmy sprawdzić, czy jest on poprawny. Techniki walidacyjne obejmują różne aspekty modelu takie, jak: analiza poprawności założeń, identyfikacje nietypowych obserwacji zaburzających działanie modelu, analiza jakości dopasowania modelu, czy analiza istotności zmiennych w modelu. Podstawowe techniki weryfikujące model oraz jakość dopasowania omówione są w Rozdziale 5. Podstawowe testy staty-

styczne istotności współczynników modelu – oparte o wielowymiarową normalność – są omawiane w Rozdziale 2.8.

5. **Przyjęcie lub odrzucenie modelu.** Na podstawie wyników testów – przeprowadzanych głównie w kroku (4.) – należy podjąć decyzję, czy zbudowany model jest poprawny i czy chcemy go użyć. Warto wspomnieć, że, w praktyce, modele są monitorowane na bieżąco. Przy modelach służących do prognozy, jakość dopasowania można monitorować przy użyciu tzw. testowania wstecznego (ang. *backtesting*).

1.2 Podstawowa notacja statystyczna

Podczas tego kursu będziemy używać podstawowej notacji statystycznej. We wszystkich definicjach zakładamy istnienie referencyjnej przestrzeni probabilistycznej $(\Omega, \Sigma, \mathbb{P})$, gdzie miara $\mathbb{P}$ może być nieznana; zazwyczaj rozważa się rodzinę miar $(\mathbb{P}_\theta)_{\theta\in\Theta}$, gdzie Θ jest przestrzenią parametrów oraz $\mathbb{P} = \mathbb{P}^{\theta_0}$ dla pewnego (nieznanego) parametru $\theta_0 \in \Theta$. Wszystkie definicje będą podane (dla uproszczenia) tylko względem referencyjnej przestrzeni, choć formalnie własność ta powinna być spełniona dla dowolnego $\theta \in \Theta$ oraz powiązanej przestrzeni probabilistycznej $(\Omega, \Sigma, \mathbb{P}^\theta)$. Zaczniemy od zdefiniowania próby prostej.

Definicja 1.1 (Próba prosta)**.** Dla ustalonego $n \in \mathbb{N}$ **próbą prostą** (lub próbką prostą) ze zmiennej losowej X będziemy nazywać zbiór zmiennych losowych $X_1, \ldots, X_n$ takich, że zbiór $(X, X_1, \ldots, X_n)$ jest zbiorem zmiennych niezależnych o takim samym rozkładzie (i.i.d.). Konkretne **realizacje** zmiennych losowych $X_1, \ldots, X_n$ będziemy oznaczać przez $x_1, \ldots, x_n$.

Kolejnym ważnym pojęciem jest definicja estymatora i jego podstawowych własności takich jak nieobciążoność, czy zgodność.

Definicja 1.2 (Estymatory i powiązane pojęcia)**.** Dla ustalonego $n \in \mathbb{N}$ **estymatorem** parametru $\theta \in \mathbb{R}$ nazywamy dowolną funkcję mierzalną $\hat{\rho}_n : \mathbb{R}^n \to \mathbb{R}$ której celem jest szacowanie wartości θ. Mając daną próbę prostą (odp. jej realizację) będziemy pisać skrótowo

$$\hat{\rho} = \hat{\rho}_n(X_1, \ldots, X_n) \quad (\text{odp. } \hat{\rho} = \hat{\rho}_n(x_1, \ldots, x_n)),$$

oraz traktować $\hat{\rho}$ jako zmienną losową (odp. jej realizację). Estymator $\hat{\rho}$ będziemy nazywać **nieobciążonym**, jeżeli $\mathbb{E}[\hat{\rho}] = \theta$ oraz **asymptotycznie nieobciążonym**, jeżeli $\mathbb{E}[\hat{\rho}] \to \theta$ (gdy $n \to \infty$). Estymator nazywamy (słabo) **zgodnym** jeżeli $\hat{\rho} \to \theta$ (według prawdopodobieństwa, gdy $n \to \infty$).

Podstawowymi estymatorami są estymatory średniej oraz odchylenia standardowego.

Definicja 1.3 (Estymatory średniej i odchylenia standardowego)**.** Przez standardowe **estymatory średniej** oraz **odchylenia standardowego** będziemy rozumieć zmienne losowe

$$\bar{X} := \tfrac{1}{n}\textstyle\sum_i^n X_i \quad \text{oraz} \quad \bar{\sigma} := \sqrt{\tfrac{1}{n-1}\textstyle\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Czasami (z drobną kolizją oznaczeń) będziemy również używać notacji $\bar{x}$ oraz $\bar{\sigma}$ w odniesieniu do konkretnych realizacji powyższych zmiennych.

Definicje 1.1, 1.2 oraz 1.3 można oczywiście rozszerzyć na d-wymiarowe wektory losowe.

1.3 Wstęp do regresji liniowej

W rozdziale tym przedstawimy podstawową notację związaną z modelem regresji liniowej opierającą się o uproszczony model probabilistyczny. Załóżmy, że mamy daną zmienną losową Y oraz k-wymiarowy wektor losowy $(X_1, \ldots, X_k)$. Naszym celem jest prognoza (średniej) wartości Y w oparciu o wartości $X_1, \ldots, X_k$. W modelu liniowym zakłada się, że prognozy tej można dokonać przy użyciu funkcji liniowej; wprowadźmy pojęcie tzw. postaci strukturalnej modelu.

Definicja 1.4 (Postać strukturalna uproszczonego modelu regresji liniowej)**. Postacią strukturalną uproszczonego modelu regresji liniowej** nazywamy równanie

$$Y = \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon, \tag{1.1}$$

gdzie $(\beta_i)_{i=1}^k$ są (nieznanymi) parametrami rzeczywistymi, a ϵ jest nieobserwowalną częścią (błędem) modelu.

Jeżeli $k = 1$, to często mówi się o **regresji prostej** (ang. *simple regression*), natomiast jeżeli $k > 1$, to o **regresji wielorakiej** (ang. *multple regression*). Podstawowym założeniem modelu regresji liniowej jest liniowa zależność między **zmienną objaśnianą** (endogeniczną) Y, a **zmiennymi objaśniającymi** (egzogenicznymi) $X_1, \ldots, X_k$. Wektor parametrów $(\beta_i)_{i=1}^k$ nazywa się często **wektorem współczynników regresji liniowej**, człon modelu $\beta_1 X_1 + \ldots + \beta_k X_k$ **regresją** bądź **funkcją regresji**, natomiast ϵ **błędem losowym** lub **czynnikiem losowym (stochastycznym)** modelu.[1]

Oczywiście, nawet w uproszczonym modelu probabilistycznym, postać strukturalna jest bardzo ogólna i będziemy potrzebowali dodatkowych założeń, aby móc estymować β. Przy standardowych (uproszczonych) założeniach, będziemy rozważali n-wymiarową próbkę prostą z wektora $(Y, X_1, \ldots, X_k)$. Oznaczając próbkę przez $(Y_i, X_{i1}, \ldots, X_{ik})_{i=1}^n$, a powiązany wektor realizacji przez $(y_i, x_i')_{i=1}^n$, gdzie $x_i' := (x_{i1}, \ldots, x_{ik})$, $i = 1, 2, \ldots, n$, liniowa zależność między realizacjami wyraża się poprzez zbiór równań

$$y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, 2, \ldots, n. \tag{1.2}$$

Zbiór równań (1.2) można przedstawić w notacji macierzowej

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}, \tag{1.3}$$

gdzie

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} := \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Macierz $\boldsymbol{X}$ często nazywa się **macierzą danych** (ang. *data matrix*).

[1] często zmienne objaśniające są deterministyczne (w przeciwieństwie do zmiennej objaśnianej) stąd zwyczajowa nazwa *czynnik losowy* w odniesieniu do powiązanego błędu modelu.

Uwaga 1.5 (Wyraz wolny)**.** W większości przypadków w modelu uwzględnia się tzw. **wyraz wolny** (ang. *intercept coefficient*) poprzez dodanie zmiennej objaśniającej $X_0 \equiv 1$ oraz powiązanego współczynnika regresji β_0. Równanie macierzowe (1.3) przyjmuje wtedy postać $\boldsymbol{y} = \tilde{\boldsymbol{X}}\tilde{\beta} + \boldsymbol{\epsilon}$, gdzie

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \text{oraz} \quad \tilde{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Dla uproszczenia często wyróżnia się współczynnik β_0, oznaczając go przez α, i rozważa powiązane równanie

$$\boldsymbol{y} = \alpha + \boldsymbol{X}\beta + \boldsymbol{\epsilon}.$$

W dalszej części wykładu będziemy używać tych notacji zamiennie.

Uwaga 1.6 (Przekształcenia modelu)**.** Na pierwszy rzut oka postać modelu (1.1) może wydawać się ograniczona na przykład w związku z tym, iż dopuszcza jedynie addytywny rodzaj błędu i liniowy typ zależności. Poprzez transformację zmiennych można tak przekształcić równanie (1.1), aby uwzględniało nieliniowe zależności oraz multiplikatywny typ błędu. Przykładowo, zakładając $Y > 0$ oraz $(X_1, \ldots, X_k) > 0$, możemy rozważyć równanie postaci

$$\log(Y) = \beta_1 \log(X_1) + \ldots + \beta_k \log(X_k) + \epsilon,$$

które odpowiada zależności $Y = X_1^{\beta_1} \cdot \ldots \cdot X_k^{\beta_k} e^{\epsilon}$ z multiplikatywnym błędem. Możemy również dokonać transformacji zmiennych objaśniających i rozważyć przykładowe równanie

$$Y = \beta_1 |X_1| + \beta_2 X_1^2 + \beta_3 e^{X_2} + \ldots + \epsilon,$$

które dopuszcza nieliniową zależność między Y, a wejściowymi zmiennymi $X_1, \ldots, X_k$. Oba równania wciąż przyjmują strukturalną postać modelu regresji liniowej.

W dalszej części wykładu zajmiemy się analizą tego, jakie założenia należy nałożyć na y, $\boldsymbol{X}$ oraz ϵ, aby można było efektywnie estymować współczynniki regresji liniowej. Model regresji może służyć m.in. do predykcji wartości y na podstawie wartości $\boldsymbol{X}$, więc kontrola nad błędem ϵ oraz jego minimalizacja będzie centralną częścią tego wykładu.

Uwaga 1.7 (Zmienne jakościowe)**.** W modelu regresji liniowej dopuszczamy występowanie zmiennych o rozkładach nieciągłych. Oprócz standardowych zmiennych o rozkładzie dyskretnym obejmuje to tzw. zmienne jakościowe (ang. *categorical variables* lub *factors*). Są to zmienne, które zazwyczaj służą do opisu cechy (np. koloru oczy, czy płci) do której ciężko przyporządkować wartość liczbową (np. gdy nie ma naturalnego porządku między wartościami). Zmienne takie zazwyczaj uwzględnia się w modelu strukturalnym poprzez wprowadzenie do wektora $(X_1, \ldots, X_k)$ funkcji charakterystycznych określających występowanie danej cechy. Innymi słowy, mając do czynienia ze zmienną jakościową Z przyjmującą m różnych wartości $(a_1, \ldots, a_m)$, w której każda wartość odpowiada konkretnej wartości cechy, możemy stworzyć wektor $(Z_1, \ldots, Z_m)$, taki, że

$$Z_l := \mathbb{1}_{\{Z=a_l\}}, \quad l = 1, 2 \ldots, m.$$

Zmienne Z_l często nazwa się zmiennymi identyfikującymi (ang. *dummy variables* lub *indicator variable*). Aby uniknąć zdegenerowania modelu, zazwyczaj włącza się tylko pierwsze $m-1$ zmiennych; na tej podstawie można zidentyfikować jednoznacznie wartość Z_m.

2 Klasyczny model regresji liniowej

W rozdziale tym omówimy podstawowe założenia modelu regresji liniowej i metody estymacji współczynników regresji liniowej β. Dla uproszczenia będziemy używać notacji macierzowej przedstawionej w (1.3), tzn. rozważać równanie macierzowe

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}. \tag{2.1}$$

Od teraz, dla uproszczenia, będziemy używać notacji (2.1) również w odniesieniu do wektorów losowych, a nie tylko ich realizacji. Warto zwrócić uwagę, że w (2.1) zakładamy, iż mamy dane n-równań, które niekoniecznie muszą odnosić się do próbki prostej (jak miało to miejsce w uproszczonym modelu).

2.1 Założenia modelu regresji liniowej

Niech $n \in \mathbb{N}$ oznacza ilość obserwacji, a $k \in \mathbb{N}$ ilość zmiennych w modelu (zakładamy $n > k$). Mamy daną $n \times k$ wymiarową (losową) macierz danych $\boldsymbol{X}$, $n \times 1$ wymiarowy wektor losowy $\boldsymbol{y}$, $n \times 1$ wymiarowy wektor losowy $\boldsymbol{\epsilon}$, oraz $k \times 1$ wymiarowy wektor współczynników β. Poniżej przedstawiamy podstawowe założenia modelu regresji liniowej:

(A.1) **Liniowa zależność** (ang. *linearity*). Model określa liniową zależność między zmienną objaśnianą, a zmiennymi objaśniającymi poprzez równanie (2.1), tzn. zależność $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$.

(A.2) **Brak współliniowości** (ang. *full rank*, *no multicollinearity*). Między zmiennymi objaśniającymi nie występuje współliniowość, tj. rząd (losowej) macierzy danych $\boldsymbol{X}$ to k (prawie na pewno).

(A.3) **Egzogeniczność czynnika losowego** (ang. *exogenity of independent variables*, *strict exogenity*). Czynnik losowy $\boldsymbol{\epsilon}$ ma zerową wartość oczekiwaną i jest niezależny (stochastycznie, liniowo) od zmiennych objaśniających, tzn. dla $i = 1, \ldots, n$ mamy $\mathbb{E}[\epsilon_i \mid \boldsymbol{X}] = 0$.

(A.4) **Homoskedastyczność i niezależność liniowa błędów** (ang. *homoscedasticity and nonautocorrelation*, *spherical error variance*). Czynnik losowy $\boldsymbol{\epsilon}$ ma stałą wariancję i nie występuję w nim zjawisko autokorelacji, tzn. istnieje $\sigma > 0$, taka że $\mathbb{E}[\epsilon_i^2 \mid \boldsymbol{X}] = \sigma^2$ oraz $\mathbb{E}[\epsilon_i \epsilon_j \mid \boldsymbol{X}] = 0$ dla $i, j = 1, \ldots, k$ (gdzie $i \neq j$).

(A.5) **Normalny rozkład błędów** (ang. *Normality of the error term*, *normality*). Czynnik losowy $\boldsymbol{\epsilon}$ ma (warunkowy) wielowymiarowy rozkład normalny pod warunkiem $\boldsymbol{X}$, tzn $\boldsymbol{\epsilon}|\boldsymbol{X} \sim \mathcal{N}_n$.

Definicja 2.1 (Model regresji liniowej)**.** Model spełniający założenia (A.1)–(A.4) nazywamy **modelem regresji liniowej**.[a] Jeżeli model spełnia dodatkowo założenie (A.5) mówimy o **standardowym (klasycznym)** modelu regresji liniowej.

[a] Czasami, z drobną kolizją oznaczeń, będziemy mówić o modelu regresji liniowej, jeżeli będą spełnione tylko założenia (A.1)–(A.2), bądź (A.1)–(A.3).

W założeniach (A.3) i (A.4) zakładamy istnienie (i skończoność) dwóch pierwszych momentów czynnika losowego $\boldsymbol{\epsilon}$. Mając dane (A.3), założenie (A.4) można przedstawić równoważnie w bardziej zwięzłej formie

$$\mathrm{Var}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \sigma^2 \mathrm{I}_n,$$

gdzie I_n to macierz jednostkowa $n \times n$; założenia (A.3)–(A.5) można również ująć łącznie w uproszczonej postaci

$$\boldsymbol{\epsilon}|\boldsymbol{X} \sim \mathcal{N}_n[0, \sigma^2 \mathrm{I}_n], \tag{2.2}$$

gdzie $\mathcal{N}_n(\mu, \Sigma)$ to n-wymiarowy rozkład normalny o wektorze średnich μ i macierzy kowariancji Σ. W tym kursie nie wykorzystaliśmy jednak uproszczonej postaci, gdyż wiele własności nie wymaga (warunkowej) normalności czynnika losowego, tzn. wiele rezultatów będziemy dowodzonych tylko przy wykorzystaniu założeń (A.1)–(A.4). Omówmy teraz pokrótce każde z założeń:

- Założenie (A.1) określa typ zależności między zmienną objaśniającą, a zmiennymi objaśnianymi. Jak wspomnieliśmy w Uwadze 1.6, założenie to obejmuje szeroką klasę modeli i odnosi się bardziej do tego jak parametry β są uwzględnione w modelu oraz typu błędu, niż do samej zależności między czynnikami modelującymi.

- Założenie (A.2) powiązane jest z tzn. identyfikacją modelu i umożliwia efektywną estymację współczynników. Warto zwrócić uwagę, iż założenie to często powiązane jest z właściwą reprezentacją macierzy danych. Uwzględnienie tych samych danych wielokrotnie może doprowadzić do sytuacji w której nie będziemy w stanie jednoznacznie estymować współczynników regresji.

- Założenie (A.3) odnosi się do czynnika losowego. Wartość średnia błędu nie powinnna zależeć (liniowo) od macierzy danych i wynosić zero.

Propozycja 2.2 (Niezależnośc czynnika losowego od zmiennych objaśniających)**.** W modelu regresji liniowej czynnik losowy jest niezależny liniowo od zmiennych objaśniających, tzn. dla $i, j = 1, 2, \ldots, n$ zachodzi

$$\mathbb{E}[x_j \cdot \epsilon_i] = (\mathbb{E}[x_{j1}\epsilon_i], \ldots, \mathbb{E}[x_{jk}\epsilon_i])' = 0.$$

Dowód Propozycji 2.2 pozostawiony jest jako proste ćwiczenie (należy skorzystać z prawa wieży oraz liniowości operatora wartości oczekiwanej). Warto wspomnieć, iż poprzez uwzględnienie wyrazu wolnego w modelu, łatwo jest dostać (bezwzględną) zerową wartość oczekiwaną czynnika losowego. Między innymi dlatego praktycznie we wszystkich zastosowaniach uwzględnia się wyraz wolny. Oczywiście tutaj założenie jest dużo silniejsze (tzn. warunkowa wartość może być niezerowa, gdy bezwarunkowa wartość jest zerowa).

- Założenie (A.4) uściśla jaki typ błędów bierzemy pod uwagę, zarówno w odniesieniu do wielkości (homoskedastyczność), jak i zależności między błędami (autokorelacja). Jest ono szczególnie ważne w odniesieniu do szeregów czasowych, gdzie często obserwujemy trendy w zmienności. W dalszej części wykładu pokażemy, jak je osłabić.

- Założenie (A.5) jest istotne z praktycznego punktu widzenia, gdyż pozwala nam na (statystyczną) kontrolę nad czynnikiem losowym. Dodatkowo, Centralne Twierdzenie Graniczne sugeruje taki rozkład czynnika losowego w wielu przypadkach. Z Reprezentacji (2.2) wynika również, iż rozkład $\boldsymbol{\epsilon}$ jest niezależny od $\boldsymbol{X}$, co daje nam od razu rozkład bezwarunkowy, tzn. zachodzi $\boldsymbol{\epsilon} \sim \mathcal{N}_n[0, \sigma^2 \mathrm{I}_n]$

Na koniec przedstawmy kilka uwag dotyczących uproszczonych wersji modelu, jego związku z warunkową wartością oczekiwaną, i powiązaniem z szeregami czasowymi.

Uwaga 2.3 (Model regresji liniowej dla próbki prostej)**.** W uproszczonym modelu, gdy $(\boldsymbol{y}, \boldsymbol{X})$ jest próbką prostą, założenia (A.3) oraz (A.4) można wyrazić w uproszczonej postaci

$$\mathbb{E}[\epsilon_i \mid x_i] = 0 \quad (i = 1, 2, \ldots, n),$$
$$\mathbb{E}[\epsilon_i^2 \mid x_i] = \sigma^2 \quad (i = 1, 2, \ldots, n),$$

gdzie x_i to wektor losowy odpowiadający i-tej obserwacji. Warto zwrócić uwagę, że gdy $(\boldsymbol{y}, \boldsymbol{X})$ jest próbką prostą, to automatycznie dostajemy równość $\mathbb{E}[\epsilon_i^2] = \sigma^2$ dla pewnego $\sigma > 0$ bez dodatkowych założeń. Nie jest to jednak równoważne (A.4), które odnosi się do warunkowej postaci wariancji. Dla odróżnienia, często pierwszą własność nazywa się **bezwarunkową homoskedastycznością** (ang. *unconditional homoskedasticity*), a drugą, **warunkową homoskedastycznością** (ang. *conditional homoskedasticity*).

Uwaga 2.4 (Model regresji liniowej przy braku losowości zmiennych objaśniających)**.** W wielu zastosowaniach (szczególnie ekonomicznych) zakłada się, iż jedynym czynnikiem stochastycznym jest czynnik losowy. Innymi słowy, macierz danych nie jest macierzą losową, tylko stałą. Wtedy założenia (A.3) oraz (A.4) również się upraszczają, gdyż $\mathbb{E}[\cdot \mid \boldsymbol{X}] = \mathbb{E}[\cdot]$.

Uwaga 2.5 (Regresja, a warunkowa wartość oczekiwana)**.** Założenia (A.1) oraz (A.3) są kluczowe dla modelu, gdyż implikują równość

$$\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] = \boldsymbol{X}\beta,$$

co pozwala nam na interpretację funkcji regresji w języku warunkowych wartości oczekiwanych (rzutowań ortogonalnych).

Uwaga 2.6 (Regresja liniowa, a szeregi czasowe)**.** Dla szeregów czasowych założenie (A.3) oznacza niezależność czynników losowych względem przeszłości, teraźniejszości, jak i przyszłości, co często nie jest spełnione (np. dla klasycznego modelu autoregresji). Zajmiemy się tym w dalszej części wykładu.

2.2 Metoda najmniejszych kwadratów i estymator OLS parametru β

Wartości błędów losowych nie są bezpośrednio obserwowane, ale dla każdego wektora $\hat{\beta} \in \mathbb{R}^k$ możemy policzyć różnicę między wartościami objaśnianymi, a jej prognozami, tzn. wartościami funkcji regresji $\boldsymbol{X}\hat{\beta}$. Przy ustalonym $\hat{\beta} \in \mathbb{R}^k$ dostajemy ciąg różnic

$$y_i - x_i'\,\hat{\beta}, \quad (i = 1, 2, \ldots, n), \tag{2.3}$$

które często nazywane są **residuami** (ang. *residuals*) modelu. W optymalizacji często prowadza się tzw. **funkcję celu**, która na podstawie wartości (2.3) bada nam jakość dopasowania modelu do danych. Najbardziej popularnym wyborem jest funkcja odpowiadająca normie kwadratowej. Mając dany wektor współczynników $\hat{\beta} \in \mathbb{R}^k$ i powiązane residua, zdefiniowane w (2.3), suma kwadratów residuów (ang. *Residual Sum of Squares* – RSS)[2] dana jest przez

$$\mathrm{RSS}(\hat{\beta}) := \sum_{i=1}^{n} (y_i - x_i'\hat{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}). \tag{2.4}$$

[2] W literaturze można spotkać też inne oznaczenia, np. *Sum of Squared Residuals* (SSR), lub *Sum of Squared Errors of prediction* (SSE).

Standardowy estymator najmniejszych kwadratów (nieznanego) parametru β, zwany w skrócie estymatorem OLS (ang. *Ordinary Least Squares*) parametru β , uzyskujemy poprzez minimalizacje funkcji celu RSS, tzn. rozwiązanie problemu

$$\mathrm{RSS}(\hat{\beta}) \to \min . \tag{2.5}$$

Otrzymana wartość $\hat{\beta} \in \mathbb{R}^k$ będzie oczywiście zależeć od konkretnej realizacji $(\boldsymbol{y}, \boldsymbol{X})$ i tylko przybliży nieznaną wartość β. Z postaci funkcji RSS widzimy, iż penalizuje ona duże odchylenia. Estymator OLS nakłada wysoką karę nawet na niewielką liczbę dużych residuów, kosztem niewielkich kar dla małych residuów. Dokładne własności statystyczne tego estymatora zostaną omówione w dalszej części wykładu.

Uwaga 2.7 (Funkcja celu)**.** Funkcja celu RSS jest ściśle związana z metodą najmniejszych kwadratów, ale nie jest jedynym możliwym wyborem. Przykładowo, zamiast rozważać sumę kwadratów błędów, możemy rozważyć sumę modułów błędów (ang. *Least Absolute Deviation* – LAD), tzn. funkcję celu

$$\mathrm{LAD}(\hat{\beta}) = \sum_{i=1}^{n} |y_i - x_i'\hat{\beta}|.$$

Funkcja ta prowadzi często do bardziej odpornych (ang *robust*) wyników, ale nie gwarantuje jednoznaczności i stabilności rozwiązań. W funkcji celu możemy też uwzględniać koszt związany z uwzględnieniem wielu zmiennych w modelu (np. metody LASSO). W tym wykładzie nie będziemy jednak poruszać tego typu zagadnień.

Zanim przejdziemy do rozwiązania (2.5) wprowadźmy definicję estymatora OLS parametru β.

Definicja 2.8 (Estymator OLS parametru β)**.** Dla modelu regresji liniowej **estymatorem OLS parametru** β nazywamy wektor $\boldsymbol{b}$, który definiujemy jako

$$\boldsymbol{b} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.6}$$

Możemy następnie powiązać estymator $\boldsymbol{b}$ z wyjściowym problemem optymalizacyjnym (2.5).

Propozycja 2.9 (Estymator $\boldsymbol{b}$ estymatorem najmniejszych kwadratów)**.** Załóżmy, iż mamy dany model regresji liniowej spełniający (A.1)–(A.3). Wtedy, rozwiązaniem problemu (2.5) jest estymator $\boldsymbol{b}$ dany w (2.6).

Dowód. Aby rozwiązać problem (2.5) można skorzystać ze standardowych metod gradientowych (warto zwrócić uwagę, iż funkcja RSS ma postać gładkiej funkcji kwadratowej). Sprawdźmy najpierw warunek konieczny pierwszego rzędu, tzn. przyrównajmy wartość $\frac{\partial \mathrm{RSS}(\hat{\beta})}{\partial \hat{\beta}}$ do zera. Dla uproszczenia ustalmy dodatkową notację $a := \boldsymbol{X}'\boldsymbol{y}$ oraz $A := \boldsymbol{X}'\boldsymbol{X}$. Dla każdego ustalonego $\hat{\beta}$ dostajemy

$$\begin{aligned}\mathrm{RSS}(\hat{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = (\boldsymbol{y}' - \hat{\beta}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = \boldsymbol{y}'\boldsymbol{y} - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{y}'\boldsymbol{y} - 2a'\hat{\beta} + \hat{\beta}'A\hat{\beta}.\end{aligned}$$

Czynnik $\boldsymbol{y}'\boldsymbol{y}$ nie zależy od $\hat{\beta}$ i może być pominięty w pochodnej funkcji RSS. Łatwo zauważyć również, iż ponieważ A jest symetryczna, to dostajemy

$$\frac{\partial(a'\hat{\beta})}{\partial\hat{\beta}} = a \quad \text{oraz} \quad \frac{\partial(\hat{\beta}'A\hat{\beta})}{\partial\hat{\beta}} = 2A\hat{\beta} \qquad \text{(ćwiczenie do domu)},$$

co daje nam gradient

$$\frac{\partial \operatorname{RSS}(\hat{\beta})}{\partial \hat{\beta}} = -2a + 2A\hat{\beta}. \tag{2.7}$$

Przyrównanie (2.7) do zera pokazuje, że potencjalny optymalny estymator, oznaczany przez $\boldsymbol{b}$, musi spełniać równanie $A\boldsymbol{b} = a$, tzn.

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}. \tag{2.8}$$

Równanie (2.8) nazywane jest często równaniem normalnym regresji liniowej (ang. *normal equations*). Korzystając z założenia (A.2), wiemy, iż macierz A jest niezdegenerowana i dodatnio określona. Mnożąc obie strony (2.8) przez $A^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ dostajemy (potencjalne) unikalne rozwiązanie postaci

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Aby zakończyć dowód, wystarczy sprawdzić warunek konieczny drugiego rzędu, tzn. zbadać Hessian. Zauważając, iż A jest dodatnio określona oraz hessian funkcji ma (stałą) postać

$$\frac{\partial^2 \operatorname{RSS}(\hat{\beta})}{\partial^2 \hat{\beta}} = 2A,$$

wiemy, że funkcje RSS osiąga w punkcie $\boldsymbol{b}$ (globalne) minimum, co kończy dowód. □

Estymator $\boldsymbol{b}$ można równoważnie zapisać jako

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X}/n)^{-1}(\boldsymbol{X}'\boldsymbol{y}/n) = S_{xx}^{-1}s_{xy}, \tag{2.9}$$

gdzie S_{xx} to średnia z próbki $(x_i x_i')_{i=1}^n$, a s_{xy} to średnia z próbki $(x_i y_i)_{i=1}^n$, tzn.

$$S_{xx} := \overline{\boldsymbol{X}'\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i',$$

$$s_{xy} := \overline{\boldsymbol{X}\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i.$$

Reprezentacja 2.6 jest używana zazwyczaj w odniesieniu do skończonych próbek, podczas gdy Reprezentacja 2.9 jest używana przy analizie asymptotycznej.

2.3 Podstawowe statystyki związane z regresja liniową OLS

Z estymatorem OLS parametru β wyrażonym przez (2.6) związanych jest wiele statystyk i powiązanych wartości, które mogą służyć np. do oceny jakości modelu. Poniżej przedstawiamy listę wybranych statystyk wraz z krótkim komentarzem.

Definicja 2.10 (Statystyki związane z OLS)**.** Załóżmy, że mamy dany model regresji liniowej oraz powiązany estymator OLS parametru β oznaczany przez $\boldsymbol{b}$. Wtedy,

1) **Residua OLS** (ang. *OLS residuals*) oznaczamy przez $\boldsymbol{e} = (e_1, \ldots, e_n)$ i definiujemy jako
$$\boldsymbol{e} := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}. \tag{2.10}$$

2) **Wektor predykcji** (ang *fitted value*) oznaczamy przez $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ i definiujemy jako
$$\hat{\boldsymbol{y}} := \boldsymbol{X}\boldsymbol{b}$$

3) **Suma kwadratów reszt modelu** (ang. *OLS residual sum of squares*) oznaczana przez RSS zdefiniowana jest jako
$$\mathrm{RSS} := \boldsymbol{e}'\boldsymbol{e}$$

4) **Współczynnik determinacji** (ang. *coefficient of determination*) modelu oznaczamy przez R^2 i definiujemy jako
$$R^2 := 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Wyjaśnimy teraz pokrótce idee stojącą za definicjami:

- Wektor residuów $\boldsymbol{e}$ i wektor predykcji $\hat{\boldsymbol{y}}$ można powiązać równaniem $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$. Wektor predykcji określa teoretyczną (średnią) wartość, którą prognozuje model, a residua określają błąd między prognozą, a prawdziwą wartością. Ponieważ prawdziwy błąd $\boldsymbol{\epsilon}$ jest nieobserwowalny, często używamy $\boldsymbol{e}$ aby go przybliżyć. Z równania (2.8), dostajemy $\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = 0$, co daje nam
$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = 0. \tag{2.11}$$
Wiemy też, że $\boldsymbol{X}'\boldsymbol{e} = n(\frac{1}{n}\sum_{i=1}^n x_i \cdot e_i)$, więc równanie to można traktować jako odpowiednik warunku ortogonalności $\mathbb{E}[x_i\epsilon_i] = 0$ dla próbki, gdzie nieznaną wartość $\boldsymbol{\epsilon}$ zastępujemy przez $\boldsymbol{e}$. Zamiast rozważać residua, często też dokonuje się bezpośredniego porównania $\boldsymbol{y}$ z $\hat{\boldsymbol{y}}$, aby sprawdzić jakość dopasowania modelu. Warto też zaznaczyć, iż z (2.11) wynika, że $\hat{\boldsymbol{y}}$ oraz $\boldsymbol{e}$ są do siebie ortogonalne (ozn. $\hat{\boldsymbol{y}} \perp \boldsymbol{e}$) , tzn. zachodzi
$$\hat{\boldsymbol{y}}'\boldsymbol{e} = \boldsymbol{b}'(\boldsymbol{X}'\boldsymbol{e}) = 0 \tag{2.12}$$

- Suma kwadratów reszt modelu OLS jest niczym innym jak zrealizowaną wartością funkcji celu zadanej w (2.4), tzn. mamy równość $\mathrm{RSS} = \mathrm{RSS}(\boldsymbol{b})$.

- Współczynnik determinacji R^2 jest jedną z podstawowych miar dopasowania modelu. Dla uproszczenia załóżmy, iż w modelu występuje wyraz wolny.[3] Wtedy, zachodzi $0 \leq R^2 \leq 1$ oraz
$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2};$$
dowody tych faktów zostaną pokazane w późniejszej części wykładu w Propozycji 2.22. W skrócie, współczynnik ten informuje nas o tym, jaka część zmienności (wariancji) zmiennej objaśnianej ($\sum_{i=1}^n (y_i - \bar{y})^2$) została wyjaśniona przez model ($\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$). Oczywiście model idealnie dopasowany powinien dawać $R^2 \equiv 1$.

[3] Współczynnik ten zdefiniowaliśmy w ogólnym przypadku, ale bez wyrazu wolnego traci on swoją podstawową interpretację (np. może być ujemny). Przy braku zmiennej objaśniające często rozważa się jego drobną modyfikację (ang. *uncentered* R^2).

2.4 Geometryczna interpretacja estymatora OLS parametru β

Łatwo zauważyć, że $\boldsymbol{b}$ to nic innego jak współczynniki rzutowania wektora $\boldsymbol{y}$ na przestrzeń rozpiętą przez kolumny $\boldsymbol{X}$ (porównaj Uwaga 2.5). Mówiąc dokładniej, załóżmy, iż mamy dany punkt $\boldsymbol{y}$ w przestrzeni $\mathbb{R}^n$ oraz k-wymiarową podprzestrzeń U rozpiętą przez wektory tworzące kolumny macierzy $\boldsymbol{X}$, tzn. $U := \operatorname{im} \boldsymbol{X}$. Naszym zadaniem jest znalezienie takiego punktu $\hat{\boldsymbol{y}} \in U$, aby odległość w normie (euklidesowej) między $\hat{\boldsymbol{y}}$, a $\boldsymbol{y}$ była jak najmniejsza. Oczywiście najbliższy punkt otrzymujemy przez rzut ortogonalny $\boldsymbol{y}$ na U, czyli punkt $\hat{\boldsymbol{y}}$ dla którego zachodzi

$$\boldsymbol{y} - \hat{\boldsymbol{y}} \in (\operatorname{im} \boldsymbol{X})^{\perp} = \ker \boldsymbol{X}',$$

co równoważnie można zapisać jako $\boldsymbol{X}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$. Zauważając, że $\hat{y} = \boldsymbol{X}\hat{\beta}$ dla pewnego $\hat{\beta} \in \mathbb{R}^k$ (gdyż $\hat{y} \in U$), otrzymujemy warunek

$$\boldsymbol{X}'(\boldsymbol{X}\hat{\beta} - \boldsymbol{y}) = 0,$$

który jest równoważny warunkowi (2.8). Rozwiązanie tego problemu daje nam estymator $\boldsymbol{b}$.

Aby wytłumaczyć związek między postacią algebraiczną a postacią geometryczną posłużymy się uproszczonym modelem regresji liniowej, w którym $(\boldsymbol{y}, \boldsymbol{X})$ jest próbką prostą z wektora losowego (Y, X). Łatwo pokazać, iż z jednej strony warunkowa wartość oczekiwana $\mathbb{E}[Y|X]$ jest najlepszym $\sigma(X)$-mierzalnym predyktorem modelu przy błędzie pomiaru mierzonym za pomocą funkcji kwadratowej, a z drugiej może być traktowana jako rzut ortogonalny. Zacznijmy od dowodu pierwszego faktu.

Propozycja 2.11 (Warunkowa wartość oczekiwana jako najlepszy predyktor)**.** Niech (Y, X) będzie wektorem losowym, w którym Y jest całkowalną z kwadratem zmienną losową. Wtedy, dla każdej całkowalnej z kwadratem $\sigma(X)$-mierzalnej zmiennej losowej Z dostajemy

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2] \leq \mathbb{E}[(Y - Z)^2].$$

Dowód. Niech Z będzie całkowalną z kwadratem zmienną $\sigma(X)$-mierzalną. Korzystając z prawa wieży dostajemy

$$\begin{aligned}
\mathbb{E}[(Y - Z)^2] &= \mathbb{E}[((Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X]) - Z))^2] \\
&\geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + 2\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|Z]) - Z)] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] - 2\mathbb{E}[(Y - \mathbb{E}[Y|X])Z].
\end{aligned}$$

Następnie, dostajemy

$$\begin{aligned}
\mathbb{E}[(Y - \mathbb{E}[Y|X])Z] &= \mathbb{E}\big[\mathbb{E}[(Y - \mathbb{E}[Y|X])Z \,|\, X]\big] \\
&= \mathbb{E}[Z\,\mathbb{E}[(Y - \mathbb{E}[Y|X]) \,|\, X]] \\
&= \mathbb{E}[Z\,(\mathbb{E}[Y|X] - \mathbb{E}[Y|X])] \\
&= 0,
\end{aligned}$$

co kończy dowód. □

W uproszczonym modelu regresji liniowej zakładamy, iż warunkową wartość oczekiwaną $\mathbb{E}[Y|X]$ możemy wyrazić w postaci $X\beta$ (por. Uwaga 2.5), gdzie problem dopasowania sprowadza się do znalezienia jak najlepszego estymatora β. Jak wspomnieliśmy, warunkową wartość oczekiwaną można również traktować jako rzut ortogonalny (dla całkowalnych z kwadratem zmiennych losowych).

Propozycja 2.12 (Warunkowa wartość oczekiwana jako rzut ortogonalny)**.** Niech (Y, X) będzie wektorem losowym, w którym Y jest całkowalną z kwadratem zmienną losową. Wtedy $\mathbb{E}[Y|X]$ jest rzutem ortogonalnym zmiennej Y na przestrzeń $L^2(\Omega, \sigma(X), \mathbb{P})$.

Dowód. Dokładne wytłumaczenie i dowód Propozycji 2.12 wykracza poza materiał tego wykładu (powinien być przeprowadzony na podstawowym kursie z analizy funkcjonalnej). Aby lepiej zrozumieć ten fakt, spróbujmy przeprowadzić ogólny szkic dowodu. Przestrzeń $L^2(\Omega, \Sigma, \mathbb{P})$ jest przestrzenią Hilberta z iloczynem skalarnym zadanym przez $\langle X, Y\rangle = \mathbb{E}[XY]$, a przestrzeń $L^2(\Omega, \sigma(X), \mathbb{P})$ jest jej domkniętą podprzestrzenią liniową. Aby pokazać, że wektor $Y - \mathbb{E}[Y|X])$ jest ortogonalny do podprzestrzeni $L^2(\Omega, \sigma(X), \mathbb{P})$ należy pokazać, iż dla każdego $Z \in L^2(\Omega, \sigma(X), \mathbb{P})$ zachodzi $\langle Z, Y - \mathbb{E}[Y|X])\rangle = 0$, co wynika już wprost z drugiej części dowodu Propozycji 2.11 . □

Uwaga 2.13 (Geometryczna interpretacja estymatora $\boldsymbol{b}$)**.** Z Propozycji 2.12 można również szybko wywnioskować postać estymatora najmniejszych kwadratów $\boldsymbol{b}$. Załóżmy, iż mamy daną postać strukturalną w uproszczonym modelu regresji liniowej i liniowość rzutu, tzn. warunek $\mathbb{E}[Y|X] = \beta X$. Skoro $X \perp (Y - \mathbb{E}[Y|X])$, to musi zachodzić $\mathbb{E}[X(Y - \beta X)] = 0$, czyli

$$\beta = \mathbb{E}[X'X]^{-1} \cdot \mathbb{E}[XY].$$

Aby wyestymować β możemy więc wziąć estymatory średnich dla $\mathbb{E}[X'X]$ oraz $\mathbb{E}[XY]$, co daje nam dokładnie estymator $\boldsymbol{b}$ w postaci (2.9). Warto też zaznaczyć, iż gdy w modelu występuje wyraz wolny, to estymatory współczynników regresji liniowej można wyrazić w języku wariancji i kowariancji. Istotnie, zakładając postać $\mathbb{E}[Y|X] = \beta X + \alpha$ i obkładając ją bezwarunkową wartością oczekiwaną, dostajemy $\alpha = \mathbb{E}[Y] - \beta\mathbb{E}[X]$, co daje nam reprezentacje $\mathbb{E}[(Y - \mathbb{E}[Y])|X] = \beta(X - \mathbb{E}[X])$ pozwalającą wyrazić β jako

$$\beta = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}.$$

Przedstawmy teraz pojęcia, które definiują obiekty znane z algebry liniowej w odniesieniu do metody OLS. Są one często pomocne w dowodach i pozwalają na geometryczną interpretację modelu.

Definicja 2.14 (Macierz projekcji i anihilator)**.** Załóżmy, iż mamy dany model regresji liniowej. Wtedy,

1) **Macierz projekcji** (ang. *projection matrix*) oznaczamy przez $\boldsymbol{P}$ i definiujemy jako

$$\boldsymbol{P} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

2) **Anihilator** (ang. *annihilator, residual maker*) oznaczamy przez $\boldsymbol{M}$ i definiujemy jako

$$\boldsymbol{M} := \mathrm{I}_n - \boldsymbol{P}.$$

Zarówno $\boldsymbol{P}$ i $\boldsymbol{M}$ są symetryczne, idempotentne oraz spełnione są równości

$$\boldsymbol{P}\boldsymbol{y} = \hat{\boldsymbol{y}}, \qquad \boldsymbol{M}\boldsymbol{y} = \boldsymbol{e}.$$

Innymi słowy, macierz $\boldsymbol{P}$ jest projekcją wektora $\boldsymbol{y}$ na $\boldsymbol{X}$ natomiast macierz $\boldsymbol{M}$ można traktować jako macierz tworzącą reszty modelu. Zachodzi również

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}, \qquad \boldsymbol{M}\boldsymbol{X} = 0,$$

tzn. projekcja $\boldsymbol{X}$ na $\boldsymbol{X}$ daje nam wyjściowy wektor dla którego reszty są zerowe. Łatwo również pokazać, iż $\boldsymbol{PM} = \boldsymbol{MP} = 0$, co daje nam dekompozycję

$$\boldsymbol{y} = \boldsymbol{Py} + \boldsymbol{My} \tag{2.13}$$

na części, które są do siebie ortogonalne, tzn. zachodzi $\hat{\boldsymbol{y}} \perp \boldsymbol{e}$; por. (2.18). Przy pomocy anihilatora można też powiązać wartość RSS z prawdziwymi błędami korzystając z równości

$$\text{RSS} = \boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon}. \tag{2.14}$$

2.5 Własności estymatora OLS parametru β i Twierdzenie Gaussa-Markowa

Zbadamy teraz podstawowe własności estymatora OLS parametru β. Na początek pokażmy, iż $\boldsymbol{b}$ jest nieobciążonym estymatorem parametru β.

Propozycja 2.15 (Nieobciążoność estymatora OLS parametru β)**.** Załóżmy, że mamy dany model regresji liniowej spełniający założenia (A.1)-(A.3). Wtedy, estymator $\boldsymbol{b}$ jest (warunkowo) nieobciążonym estymatorem β, tzn.

$$\mathbb{E}[\boldsymbol{b} \mid \boldsymbol{X}] = \beta.$$

Przy dodatkowym założeniu (A.4) dostajemy również $\text{Var}[\boldsymbol{b} \mid \boldsymbol{X}] = \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Dowód. Załóżmy (A.1)–(A.3). Niech $\boldsymbol{A} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Korzystając z (A.1) dostajemy

$$\begin{aligned}
\boldsymbol{b} - \beta &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \beta \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) - \beta \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} - \beta \\
&= \boldsymbol{A}\boldsymbol{\epsilon}.
\end{aligned} \tag{2.15}$$

Następnie, korzystając z (A.3), mamy

$$\mathbb{E}[\boldsymbol{b} \mid \boldsymbol{X}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{\epsilon} \mid \boldsymbol{X}] + \beta = A\mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] + \beta = \beta,$$

co kończy dowód nieobciążoności. Załóżmy teraz dodatkowo (A.4). Korzystając z (2.15) oraz (A.4) dostajemy

$$\begin{aligned}
\text{Var}[\boldsymbol{b} \mid \boldsymbol{X}] = \text{Var}[\boldsymbol{b} - \beta \mid \boldsymbol{X}] = \text{Var}[\boldsymbol{A}\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \boldsymbol{A}\,\text{Var}[\boldsymbol{\epsilon} \mid \boldsymbol{X}]\boldsymbol{A}' = \boldsymbol{A}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \boldsymbol{X}]\boldsymbol{A}' \\
= \boldsymbol{A}\mathbb{E}[\sigma^2 \mathrm{I}_n \mid \boldsymbol{X}]\boldsymbol{A}' = \sigma^2 \boldsymbol{A}\boldsymbol{A}'.
\end{aligned} \tag{2.16}$$

Aby zakończyć dowód, wystarczy zauważyć, że z (A.2) dostajemy

$$\boldsymbol{A}\boldsymbol{A}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}((\boldsymbol{X}'\boldsymbol{X})^{-1})' = ((\boldsymbol{X}'\boldsymbol{X})^{-1})' = ((\boldsymbol{X}'\boldsymbol{X})')^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.17}$$

□

Warto zauważyć, że Propozycja 2.5 implikuje nieobciążoność w standardowym (bezwarunkowym) sensie, tzn. własność

$$\mathbb{E}[\boldsymbol{b}] = \beta.$$

Zanim przedstawimy jedno z ważniejszych twierdzeń związanych z regresją liniową, tzw. Twierdzenie Gaussa-Markowa, zdefiniujmy specjalną klasę estymatorów nazywanych estymatorami liniowymi.

Definicja 2.16 (Estymator liniowy)**.** Załóżmy, że mamy dany model regresji liniowej. Estymator $\hat{\boldsymbol{\beta}}$ parametru β będziemy nazywać (warunkowo) **liniowym** ze względu na $\boldsymbol{y}$ jeżeli można go przedstawić w postaci $\hat{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$, gdzie $\boldsymbol{C}$ jest pewną (losową) $\sigma(\boldsymbol{X})$-mierzalną macierzą.

Estymatory liniowe są liniowe ze względu na wartości zmiennych objaśnianych $\boldsymbol{y}$, a co za tym idzie ze względu na wielkość błędu $\boldsymbol{\epsilon}$. Pokażemy teraz, że w klasie estymatorów liniowych, które dodatkowo są nieobciążone, estymator $\boldsymbol{b}$ jest w pewnym sensie najlepszy.

Twierdzenie 2.17 (Twierdzenie Gaussa-Markowa)**.** Załóżmy, że mamy dany model regresji liniowej spełniający założenia (A.1)–(A.4). Wtedy, w klasie estymatorów parametru β, które są liniowe oraz nieobciążone, estymator $\boldsymbol{b}$ jest estymatorem o najmniejszej wariancji, czyli tzw. estymatorem BLUE (ang. *Best Linear Unbiased Estimator*). Innymi słowy, dla każdego nieobciążonego estymatora $\hat{\boldsymbol{\beta}}$ parametru β, liniowego ze względu na $\boldsymbol{y}$, dostajemy

$$\operatorname{Var}[\boldsymbol{b} \mid \boldsymbol{X}] \leq \operatorname{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}]. \tag{2.18}$$

Dowód. Niech $\hat{\boldsymbol{\beta}}$ będzie nieobciążonym i liniowym estymatorem parametru β. Z liniowości wiemy, że $\hat{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$, gdzie $\boldsymbol{C}$ jest pewną losową macierzą zależną od $\boldsymbol{X}$. Niech $\boldsymbol{A} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Definiujemy $\boldsymbol{D} := \boldsymbol{C} - \boldsymbol{A}$. Wtedy

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{D} + \boldsymbol{A})\boldsymbol{y} = \boldsymbol{D}\boldsymbol{y} + \boldsymbol{A}\boldsymbol{y} = \boldsymbol{D}(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) + \boldsymbol{b} = \boldsymbol{D}\boldsymbol{X}\beta + \boldsymbol{D}\boldsymbol{\epsilon} + \boldsymbol{b}. \tag{2.19}$$

Obkładając obie strony (2.19) warunkową wartością oczekiwaną dostajemy

$$\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \mathbb{E}[\boldsymbol{D}\boldsymbol{X}\beta + \boldsymbol{D}\boldsymbol{\epsilon} + \boldsymbol{b} \mid \boldsymbol{X}] = \boldsymbol{D}\boldsymbol{X}\beta + \mathbb{E}[\boldsymbol{D}\boldsymbol{\epsilon} \mid \boldsymbol{X}] + \mathbb{E}[\boldsymbol{b} \mid \boldsymbol{X}]. \tag{2.20}$$

Ponieważ estymatory $\hat{\boldsymbol{\beta}}$ i $\boldsymbol{b}$ są nieobciążone oraz $\mathbb{E}[\boldsymbol{D}\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \boldsymbol{D}\mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = 0$, wiemy, że (2.20) implikuje $\boldsymbol{D}\boldsymbol{X}\beta = 0$. Aby własność ta była prawdziwa dla każdego β, musimy mieć $\boldsymbol{D}\boldsymbol{X} = 0$. Korzystając z (2.19) oraz własności (2.15), czyli równości $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\epsilon} + \beta$, dostajemy

$$\hat{\boldsymbol{\beta}} = \boldsymbol{b} + \boldsymbol{D}\boldsymbol{\epsilon} = (\boldsymbol{A} + \boldsymbol{D})\boldsymbol{\epsilon} + \beta.$$

Wynika stąd, że

$$\begin{aligned}
\operatorname{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] &= \operatorname{Var}[\hat{\boldsymbol{\beta}} - \beta \mid \boldsymbol{X}] \\
&= \operatorname{Var}[(\boldsymbol{A} + \boldsymbol{D})\boldsymbol{\epsilon} \mid \boldsymbol{X}] \\
&= (\boldsymbol{A} + \boldsymbol{D}) \operatorname{Var}[\boldsymbol{\epsilon} \mid \boldsymbol{X}](\boldsymbol{A} + \boldsymbol{D})' \\
&= (\boldsymbol{A} + \boldsymbol{D}) \operatorname{Var}[\boldsymbol{\epsilon} \mid \boldsymbol{X}](\boldsymbol{A}' + \boldsymbol{D}') \\
&= \sigma^2 \cdot (\boldsymbol{A} + \boldsymbol{D})(\boldsymbol{A}' + \boldsymbol{D}') \\
&= \sigma^2 \cdot (\boldsymbol{A}\boldsymbol{A}' + \boldsymbol{D}\boldsymbol{A}' + \boldsymbol{A}\boldsymbol{D}' + \boldsymbol{D}\boldsymbol{D}').
\end{aligned} \tag{2.21}$$

Z równości $\boldsymbol{D}\boldsymbol{X} = 0$ wynika, że $\boldsymbol{D}\boldsymbol{A}' = \boldsymbol{D}\boldsymbol{X}((\boldsymbol{X}'\boldsymbol{X})^{-1})' = 0$ oraz $\boldsymbol{A}\boldsymbol{D}' = (\boldsymbol{D}\boldsymbol{A}')' = 0$. Wracając do (2.21), zauważając, że $\boldsymbol{D}\boldsymbol{D}'$ jest pozytywnie określona, oraz przypominając (2.16), dostajemy

$$\operatorname{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \sigma^2 \cdot (\boldsymbol{A}\boldsymbol{A}' + \boldsymbol{D}\boldsymbol{D}') \geq \sigma^2 \cdot \boldsymbol{A}\boldsymbol{A}' = \operatorname{Var}[\boldsymbol{b} \mid \boldsymbol{X}],$$

co kończy dowód. □

Twierdzenia 2.17 implikuję również (bezwarunkową) własność $\mathrm{Var}[\boldsymbol{b}] \leq \mathrm{Var}[\hat{\boldsymbol{\beta}}]$. Zauważmy również, że z (2.18) wiemy, iż macierz $\boldsymbol{K} := \mathrm{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] - \mathrm{Var}[\boldsymbol{b} \mid \boldsymbol{X}]$ jest dodatnio półokreślona (ang. *positive semi-definite*), tzn dla każdego k-wymiarowego wektora $\boldsymbol{a}$ mamy $\boldsymbol{a}\boldsymbol{K}\boldsymbol{a}' \geq 0$. W szczególności, dla każdego $j = 1, 2, \ldots, k$, przyjmując za $\boldsymbol{a}$ wektor, który ma jedynkę na j-tym miejscu, oraz zera wszędzie indziej, dostajemy

$$\mathrm{Var}[\hat{\beta}_j \mid \boldsymbol{X}] \geq \mathrm{Var}[b_j \mid \boldsymbol{X}], \quad (j = 1, 2, \ldots, k).$$

Innymi słowy, dla każdego współczynnika regresji liniowej, wariancja współczynnika b_j jest najmniejsza wśród współczynników liniowych i nieobciążonych estymatorów. Warto też wspomnieć o kowariancji między residuami, a wartościami estymatora $\boldsymbol{b}$.

Propozycja 2.18 (Korelacja między estymatorem OLS parametru β a residuami OLS)**.** Załóżmy, że mamy dany model regresji liniowej spełniający założenia (A.1)–(A.4). Wtedy

$$\mathrm{Cov}[\boldsymbol{b}, \boldsymbol{e} \mid \boldsymbol{X}] = 0,$$

gdzie wektor residuów $\boldsymbol{e}$ zdefiniowany jest w (2.10).

Dowód Propozycji 2.18 pozostawiamy jako ćwiczenie.

Uwaga 2.19 (Problem pominiętych zmiennych)**.** Własności estymatora $\boldsymbol{b}$ pokazane w Propozycji 2.15 oraz Twierdzeniu 2.17 są ściśle związane z założeniami modelu regresji liniowej. Warto tutaj zwrócić uwagę, iż niepoprawny (niepełny) dobór zmiennych objaśniających, tj. niepoprawna specyfikacja modelu i naruszenie założenia (A.1), może prowadzić do systematycznego obciążenia estymatora $\boldsymbol{b}$. Dzieje się tak, gdy nieuwzględniona zmienna jest liniowo zależna od macierzy danych oraz wzbogaca wartość prognostyczną modelu (tzn. wartość współczynnika regresji liniowej przy tej zmiennej jest różny od zera jeżeli uwzględnimy ją jako dodatkową zmienną w modelu). Problem ten często oznacza się skrótowo przez OMV (ang. *Omitted-Variable Bias*). Więcej na ten temat można znaleźć np. w Rozdziale 4.3.2 w [Gre18] lub Rozdziale 3.9 w [Hay00].

2.6 Estymator OLS parametru σ^2 i jego własności

W modelu regresji liniowej zdefiniowanym w Definicji 2.1 mamy tak naprawdę do czynienia z dwoma nieznanymi parametrami, tzn. wektorem współczynników regresji liniowej β oraz wariancją czynnika losowego σ^2. Wprowadźmy teraz estymator OLS parametru σ^2.

Definicja 2.20 (Estymator OLS parametru σ^2)**.** Dla modelu regresji liniowej **estymatorem OLS parametru** σ^2 nazywamy liczbę (a dokładniej zmienną losową) s^2 zdefiniowaną jako

$$s^2 := \frac{\mathrm{RSS}}{n-k}. \tag{2.22}$$

Powiązany **błąd standardowy residów** (ang. *Residual Standard Error*) oznaczamy przez RSE i definiujemy jako

$$\mathrm{RSE} := \sqrt{s^2}. \tag{2.23}$$

Warto przypomnieć, że założyliśmy nierówność $n > k$, więc estymatory te są dobrze zdefiniowane. Wprost z definicji dostajemy $s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}$, co tłumaczy nazwę RSE. Warto zwrócić uwagę, że w (2.22) mianownik jest równy $n-k$ zamiast n. Jest to związane z tzw. *liczbą stopni swobody* modelu i będzie nam dawało nieobciążoność estymatora, co wyjaśnia dokładniej Propozycja 2.21.

Propozycja 2.21 (Nieobciążoność estymatora OLS parametru σ^2)**.** Załóżmy, że mamy dany model regresji liniowej spełniający założenia (A.1)–(A.4). Wtedy s^2 jest (warunkowo) nieobciążonym estymatorem parametru σ^2, tzn. zachodzi

$$\mathbb{E}[s^2 \mid \boldsymbol{X}] = \sigma^2.$$

Dowód. Korzystając z (2.14) dostajemy $\boldsymbol{e}\boldsymbol{e}' = \boldsymbol{\epsilon}\boldsymbol{M}\boldsymbol{\epsilon}'$, gdzie $\boldsymbol{M} = (m_{ij})$ jest anihilatorem. Pokażmy najpierw własność

$$\operatorname{tr}(\boldsymbol{M}) = n - k, \tag{2.24}$$

gdzie $\operatorname{tr}(\boldsymbol{M})$ oznacza ślad macierzy $\boldsymbol{M}$ (sumę elementów na głównej przekątnej). Korzystając z definicji $\boldsymbol{M}$ i liniowości operatora śladu dostajemy

$$\operatorname{tr}(\boldsymbol{M}) = \operatorname{tr}(\mathrm{I}_n) - \operatorname{tr}(\boldsymbol{P}) = n - \operatorname{tr}(\boldsymbol{P}).$$

Przypominając definicję operatora projekcji $\boldsymbol{P}$ oraz korzystając z własności $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A})$, dostajemy

$$\operatorname{tr}(\boldsymbol{P}) = \operatorname{tr}(\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}') = \operatorname{tr}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \operatorname{tr}(\mathrm{I}_k) = k,$$

co kończy dowód (2.24). Korzystając z (2.24) wystarczy więc pokazać, że

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{M}\boldsymbol{\epsilon}' \mid \boldsymbol{X}] = \sigma^2 \cdot \operatorname{tr}(\boldsymbol{M}).$$

Korzystając z założenia (A.4) i zauważając, że $\boldsymbol{M}$ jest funkcją $\boldsymbol{X}$, dostajemy

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{M}\boldsymbol{\epsilon}' \mid \boldsymbol{X}] = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\mathbb{E}[\epsilon_i\epsilon_j \mid \boldsymbol{X}] = \sum_{i=1}^{n} m_{ii}\sigma^2 = \sigma^2 \cdot \operatorname{tr}(\boldsymbol{M}),$$

co kończy dowód. □

Zauważmy, jak poprzednio, że Propozycja 2.21 implikuje bezwarunkową nieobciążoność, tzn. własność $\mathbb{E}[s^2] = \sigma^2$. Z Propozycji 2.5 wiemy, że $\operatorname{Var}[\boldsymbol{b} \mid \boldsymbol{X}] = \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Pozwala nam to na estymację błędu (wariancji) estymatora $\boldsymbol{b}$. Estymator ten definiujemy jako

$$\widehat{\operatorname{Var}[\boldsymbol{b} \mid \boldsymbol{X}]} := s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

2.7 Wybrane dodatkowe własności modelu OLS

W rozdziale tym przedstawimy kilka dodatkowych własności modelu OLS, które bywają pomocne w praktyce.

2.7.1 Regresja liniowa z wyrazem wolnym

Zacznijmy od dodatkowych własności modelu, które wynikają z uwzględnienia w nim wyrazu wolnego.

Propozycja 2.22 (Model OLS z wyrazem wolnym)**.** Załóżmy, że dany mamy model regresji liniowej w którym występuje wyraz wolny. Wtedy

1) $\sum_{i=1}^{n} e_i = 0$, tzn. residua modelu sumują się do zera.

2) $\overline{y} = \overline{\boldsymbol{X}\boldsymbol{b}}$, Średnia wartość prognozy odpowiada średniej wartości zaobserwowanych wartości.

3) $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ oraz $0 \leq R^2 \leq 1$.

Dowód. Załóżmy, iż mamy dany model regresji liniowej z wyrazem wolnym odpowiadającym pierwszej kolumnie macierzy $\boldsymbol{X}$.

1) Równanie (2.8) daje nam

$$\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) = \boldsymbol{X}'\boldsymbol{e} = \tfrac{1}{n}\sum_{i=1}^{n} x_i \cdot e_i = 0.$$

Wystarczy zauważyć, iż $x_{1i} = 1$ dla $i = 1, \ldots, n$, co daje nam $\sum_{i=1}^{n} e_i = 0$.

2) Wynika to wprost z 1) oraz równości $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$.

3) Z (2.12), tzn. warunku ortogonalności $\boldsymbol{e} \perp \hat{y}$, oraz z 1), dostajemy

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}\left[e_i^2 + 2e_i(\hat{y}_i-\bar{y}) + (\hat{y}_i-\bar{y})^2\right] = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2, \quad (2.25)$$

co daje nam

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \quad (2.26)$$

Nierówność $0 \leq R^2 \leq 1$ wynika wprost z (2.25) oraz (2.26). □

2.7.2 Twierdzenie Frischa–Waugha–Lovella

Podstawowym twierdzeniem stojącym u podstaw tzw. regresji sekwencyjnej jest Twierdzenie Frischa–Waugha–Lovella, które mówi, że współczynniki estymatora OLS parametru β można uzyskać w sposób sekwencyjny, poprzez podział zmiennych objaśniających na dwa zbiory, a następnie regresję zmiennej objaśnianej oraz jednej części zmiennych objaśniających na drugą część zmiennych objaśniających.

Twierdzenie 2.23 (Twierdzenie Frischa–Waugha–Lovella)**.** Załóżmy, że mamy dany model regresji liniowej postaci

$$\boldsymbol{y} = \boldsymbol{X}\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{\epsilon}, \quad (2.27)$$

tzn. niech $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ oraz $\beta = (\beta_1, \beta_2)$ w modelu $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$.[a] Wtedy estymator OLS parametru β_2 w modelu (2.27) jest równy estymatorowi OLS parametru β_2 w modelu

$$(\boldsymbol{M}_1\boldsymbol{y}) = (\boldsymbol{M}_1\boldsymbol{X}_2)\beta_2 + \boldsymbol{\epsilon}_1, \quad (2.28)$$

gdzie $\boldsymbol{M}_1$ jest anihilatorem pierwszego członu modelu[b] oraz $\boldsymbol{\epsilon}_1 := \boldsymbol{M}_1\boldsymbol{\epsilon}$.

[a] tutaj β_1 oraz β_2 oznacza wektorów współczynników przy $\boldsymbol{X}_1$ oraz $\boldsymbol{X}_2$, a nie dwie współrzędne wektora β.
[b] tzn. $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{P}_1$, gdzie $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$; zobacz Definicja 2.14.

Dowód. Niech $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2)$ będzie estymatorem OLS parametru $\beta = (\beta_1, \beta_2)$ w modelu (2.27). Równanie normalne regresji liniowej dla (2.8), tzn. równanie $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}$, można zapisać jako

$$\begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{pmatrix} \quad (2.29)$$

Korzystając z (2.29) dostajemy $\boldsymbol{X}_1'\boldsymbol{y} = (\boldsymbol{X}_1'\boldsymbol{X}_1)\boldsymbol{b}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_2)\boldsymbol{b}_2$, co pozwala nam przedstawić $\boldsymbol{b}_1$ jako

$$\boldsymbol{b}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{b}_2) \quad (2.30)$$

Wstawiając (2.30) do równości $\boldsymbol{X}_2'\boldsymbol{y} = (\boldsymbol{X}_2'\boldsymbol{X}_1)\boldsymbol{b}_1 + (\boldsymbol{X}_2'\boldsymbol{X}_2)\boldsymbol{b}_2$ otrzymanej ponownie z (2.29) dostajemy

$$\begin{aligned}\boldsymbol{X}_2'\boldsymbol{y} =& \boldsymbol{X}_2'[\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'](\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{b}_2) + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\\ =& \boldsymbol{X}_2'\boldsymbol{P}_1\boldsymbol{y} - \boldsymbol{X}_2'\boldsymbol{P}_1\boldsymbol{X}_2\boldsymbol{b}_2 + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\\ =& \boldsymbol{X}_2'\boldsymbol{P}_1\boldsymbol{y} + \boldsymbol{X}_2'[\boldsymbol{I}_n - \boldsymbol{P}_1]\boldsymbol{X}_2\boldsymbol{b}_2\\ =& \boldsymbol{X}_2'\boldsymbol{P}_1\boldsymbol{y} + \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{b}_2,\end{aligned}$$

co można równoważnie zapisać jako $\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{b}_2$. Dostajemy stąd

$$\boldsymbol{b}_2 = [\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}. \tag{2.31}$$

Korzystając z tego, że $\boldsymbol{M}_1$ jest idempotentna i symetryczna dostajemy

$$\boldsymbol{b}_2 = [\boldsymbol{X}_2'\boldsymbol{M}_1'\boldsymbol{M}_1\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1'\boldsymbol{M}_1\boldsymbol{y} = [(\boldsymbol{M}_1\boldsymbol{X}_2)'(\boldsymbol{M}_1\boldsymbol{X}_2)]^{-1}(\boldsymbol{M}_1\boldsymbol{X}_2)'(\boldsymbol{M}_1\boldsymbol{y}),$$

z czego wynika już, że $\boldsymbol{b}_2$ jest estymatorem OLS parametru β_2 w modelu (2.28).

□

Uwaga 2.24 (Regresja sekwencyjna - wyjaśnienie)**.** Wartości $(\boldsymbol{M}_1\boldsymbol{y})$ oraz $(\boldsymbol{M}_1\boldsymbol{X}_2)$ w Twierdzeniu 2.23 odpowiadają residuom, gdy dokonujemy regresji $\boldsymbol{y}$ na macierz $\boldsymbol{X}_1$ oraz pojedynczych kolumn macierzy $\boldsymbol{X}_2$ na macierz $\boldsymbol{X}_1$. Z Twierdzenia 2.23 wynika więc, że współczynniki pełnego modelu można odzyskać w sposób sekwencyjny poprzez:

(a) Regresję $\boldsymbol{X}_1$ na $\boldsymbol{y}$; uzyskujemy stąd wektora residuów $\boldsymbol{M}_1\boldsymbol{y}$.

(b) Regresję $\boldsymbol{X}_1$ na kolejne kolumny macierzy $\boldsymbol{X}_2$; uzyskujemy stąd macierz residuów $\boldsymbol{M}_1\boldsymbol{X}_2$.

(c) Regresję $\boldsymbol{M}_1\boldsymbol{X}_2$ na $\boldsymbol{M}_1\boldsymbol{y}$; uzyskujemy stąd estymator OLS parametru β_2, tzn. $\boldsymbol{b}_2$.

(d) Regresję $\boldsymbol{X}_1$ na $(\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{b}_2)$; uzyskujemy stąd estymator OLS parametru β_1, tzn. $\boldsymbol{b}_1$.

Warto zwrócić uwagę, że w podpunkcie (d) estymator $\boldsymbol{b}_1$ uzyskujemy poprzez *korektę* zmiennej objaśniającej $\boldsymbol{y}$ o $\boldsymbol{X}_2\boldsymbol{b}_2$, można to traktować jako

Uwaga 2.25 (Regresja sekwencyjna dla ortogonalnych zmiennych)**.** W szczególnym przypadku, gdy macierze $\boldsymbol{X}_1$ oraz $\boldsymbol{X}_2$ w (2.27) są do siebie ortogonalne, estymacja $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2)$ znacznie się upraszcza. Z równania (2.29) wynika, że proces estymacji można od siebie całkowicie oddzielić, tzn. zachodzi $\boldsymbol{b}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$ oraz $\boldsymbol{b}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}$. Ortogonalność (niezależność liniowa) zmiennych objaśniających upraszcza więc znacząco estymację modelu.

2.8 Testowanie hipotez statystycznych przy założeniu normalności

Mając dany estymator (nieznanych) współczynników regresji liniowej β chcielibyśmy zbadać, czy otrzymane współczynniki są sensowne. W szczególności, jednym z ważniejszych testów jest zbadanie, czy dana zmienna objaśniająca ma istotny wpływ na jakość modelu. Zakładając, że chcemy sprawdzić istotność pierwszej zmienną objaśniającej, możemy zbadać hipotezę statystyczną zakładającą, iż $\beta_1 = 0$ wobec alternatywy $\beta_1 \neq 0$. Często mówimy, że model jest poprawnie określony (ang. *correctly specified*) jeżeli hipoteza zerowa ($\beta_1 = 0$) jest fałszywa.

Testy statystyczne powiązane z modelem regresji liniowej zazwyczaj opierają się na zbadaniu różnicy między estymowanymi współczynnikami modelu ($\boldsymbol{b}$), a prawdziwymi współczynnikami modelu (β). Wartość $\boldsymbol{b} - \beta$ często nazywa się błędem próbki (ang. *sampling error*).

Aby dostać rozkład powiązanej statystyki testowej musimy wiedzieć, jaki ma ona rozkład przy założeniu prawdziwości hipotezy zerowej. W tym celu należy określić teoretyczny rozkład czynnika losowego. Przy założeniu warunkowej normalności odnosi się to bezpośrednio do założenia (A.5). Założenie to mówi, że czynnik losowy $\boldsymbol{\epsilon}$ ma (warunkowy) wielowymiarowy rozkład normalny pod warunkiem $\boldsymbol{X}$. Warto przy tym zwrócić uwagę, że założenie to odnosi się tylko do czynnika losowego warunkowanego przez $\boldsymbol{X}$, tzn. nie musimy dodawać założeń na rozkład (losowej) macierzy danych $\boldsymbol{X}$. Zanim przejdziemy do testowania hipotez, warto przypomnieć kilka podstawowych faktów związanych z wielowymiarowym rozkładem normalnym i ich związku z modelem regresji liniowej:

- Wielowymiarowy rozkład normalny jest wyznaczony jednoznacznie przez pierwsze dwa momenty, tzn. wektor średnich $\boldsymbol{\mu}$ oraz macierz kowariancji $\boldsymbol{\Sigma}$. W szczególności z reprezentacji (2.2), tzn. własności $\boldsymbol{\epsilon}|\boldsymbol{X} \sim \mathcal{N}_n[0, \sigma^2 \mathrm{I}_n]$, wiemy, że bezwarunkowy rozkład czynnika losowego $\boldsymbol{\epsilon}$ również ma rozkład normalny z takim samą średnią i macierzą kowariancji. Wynika to z faktu, że rozkład warunkowy nie zależy od wartości $\boldsymbol{X}$.
- Jeżeli zmienne losowe X oraz Y mają wielowymiarowy rozkład normalny to są one niezależne wtedy i tylko wtedy jeżeli są nieskorelowane. Przenosi się to również na rozkłady warunkowe $X|\boldsymbol{X}$ oraz $Y|\boldsymbol{X}$.
- Dowolna kombinacja liniowa zmiennych $X_1, \ldots, X_k$ o wielowymiarowym rozkładzie normalnym również ma rozkład normalny. Dodatkowo wektor stworzony przez dodanie kombinacji liniowej tych zmiennych do wektora $(X_1, \ldots, X_k)$ ma również (zdegenerowany) rozkład normalny.

Łącząc te fakty dostajemy, że różnica między estymatorem OLS, a prawdziwą wartością parametrów ma rozkład normalny. W dowodzie Propozycji 2.15 pokazaliśmy, że $\boldsymbol{b} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}$. Różnica ta jest więc liniową funkcją $\boldsymbol{\epsilon}$ i ma (warunkowy) rozkład normalny. Z Propozycji 2.15 wiemy, że $\mathrm{Var}[\boldsymbol{b} \mid \boldsymbol{X}] = \mathrm{Var}[\boldsymbol{b} - \beta \mid \boldsymbol{X}] = \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Podsumowując, przy założeniach (A.1)–(A.5) dostajemy

$$(\boldsymbol{b} - \beta)|\boldsymbol{X} \sim \mathcal{N}_k\left(0, \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1}\right). \tag{2.32}$$

Pokażmy teraz jak wykorzystać (2.32) do testowania podstawowych hipotez statystycznych związanych z modelem regresji liniowej.

2.8.1 Testowanie pojedynczych współczynników regresji liniowej

Na początek zajmiemy się najprostszym typem testów odnoszących się do pojedynczego współczynnika regresji liniowej. Niech $j \in \{1, 2, \ldots, k\}$ będzie ustalonym indeksem. Dla ustalonej liczby $\overline{\beta}_j \in \mathbb{R}$ zakładamy hipotezę zerową

$$\mathrm{H}_0 : \beta_j = \overline{\beta}_j, \tag{2.33}$$

wobec hipotezy alternatywnej $\mathrm{H}_1 : \beta_j \neq \overline{\beta}_j$. Załóżmy, że przedział ufności, dla którego chcemy dokonać testu to α. Oczywiście metodyka tutaj przedstawiona będzie odnosiła się też do innych typów hipotez (np. z nierównościami, czy jednostronnych) – kluczowe jest znalezienie rozkładu statystyki testowej. Przy założeniu prawdziwości H_0, dostajemy

$$(b_j - \overline{\beta}_j)|\boldsymbol{X} \sim \mathcal{N}_k\left(0, \sigma^2 \cdot [(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}\right),$$

gdzie $[\cdot]_{ab}$ odnosi się do elementu macierzy w a-tym wierszu i b-tej kolumnie. Następnym krokiem jest unormowanie zmiennej $(b_j - \overline{\beta}_j)$ tak, aby miała jednostkowe odchylenie. Teoretycznie można wprowadzić nową zmienną

$$z_j := \frac{b_j - \overline{\beta}_j}{\sqrt{\sigma^2 \cdot [(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}}.$$

Wiemy, że $z_j|\boldsymbol{X} \sim N(0,1)$, ale (gdy) błąd σ^2 jest nieznany, to nie jesteśmy w stanie obliczyć (wprost) wartości z_k.

Uwaga 2.26 (Statystyka z_j)**.** Statystyka z_j ma wiele przydatnych własności. Można ją policzyć z próbki (zakładając, że znamy σ^2), jej rozkład warunkowy względem $\boldsymbol{X}$ nie zależy od wartości $\boldsymbol{X}$, znamy jej rozkład warunkowy (i bezwarunkowy), nie zależy ona od (nieznanych) parametrów β.

Naturalnym podejściem jest zastąpienie parametru σ^2 przez jego estymator OLS, tzn. wartość s^2 zdefiniowaną w (2.23). Otrzymaną w ten sposób statystykę często nazywa się t-statystyką (ang. *t-statistic*, *t-ratio*, *t-value*) estymatora OLS, a jej mianownik błędem standardowym (ang. *standard error*) estymatora OLS wyrażonym przez

$$\mathrm{SE}(\beta_j) := \sqrt{s^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}.$$

Oczywiście podstawienie s^2 zamiast σ^2 wpływa na rozkład, co pokazuje Propozycja 2.27.

Propozycja 2.27 (Rozkład statystyki t_j)**.** Załóżmy (A.1)–(A.5). Zakładając prawdziwość hipotezy zerowej $\mathrm{H}_0 : \beta_j = \overline{\beta}_j$, statystyka t_j wyrażona przez

$$t_j := \frac{b_j - \overline{\beta}_j}{\mathrm{SE}(\beta_j)},$$

ma rozkład t-studenta o $n-k$ stopniach swobody.

Dowód. Przypominając (2.23) oraz (2.14) dostajemy

$$t_j = \frac{z_j}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{z_j}{\sqrt{\frac{\mathrm{RSS}}{(n-k)\sigma^2}}} = \frac{z_j}{\sqrt{\frac{\boldsymbol{e}'\boldsymbol{e}}{(n-k)\sigma^2}}} = \sqrt{(n-k)}\frac{z_j}{\sqrt{q}},$$

gdzie $q = \boldsymbol{e}'\boldsymbol{e}/\sigma^2$. Wiemy (np. z wykładu Statystyka), że rozkład t-studenta o m stopniach swobody można wyrazić jako

$$\sqrt{m}\frac{X}{\sqrt{Y}},$$

gdzie $X \sim N(0,1)$, a $Y \sim \chi^2(m)$ są niezależnymi zmiennymi losowymi. Wiemy, że $z_j \mid \boldsymbol{X} \sim N(0,1)$. Pokażmy teraz, że $q|\boldsymbol{X} \sim \chi^2(n-k)$ oraz q jest (warunkowo) niezależne od z_j (pod warunkiem $\boldsymbol{X}$). Korzystając z drugiej równości w (2.14) dostajemy

$$q = \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} = \frac{\boldsymbol{\epsilon}'}{\sigma}\boldsymbol{M}\frac{\boldsymbol{\epsilon}}{\sigma}.$$

Wiemy, że $\boldsymbol{M}$ jest symetryczna, idempotentna oraz $\boldsymbol{\epsilon}/\sigma \sim N(0, \mathrm{I}_n)$. Stąd wynika, że $q \sim \chi^2(\mathrm{rank}(\boldsymbol{M}))$; zob. Propozycja A.3. Ponieważ $\boldsymbol{M}$ jest idempotentna, zachodzi również $\mathrm{rank}(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M})$; zob. Propozycja A.3. Korzystając z (2.24) dostajemy więc

$$\mathrm{rank}(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}) = n - k,$$

co kończy dowód faktu, iż $q|\boldsymbol{X} \sim \chi^2(n-k)$. Pokażmy teraz, iż q jest (warunkowo) niezależne od z_j (pod warunkiem $\boldsymbol{X}$). Wiemy, iż zarówno $\boldsymbol{b}$ jak i $\boldsymbol{e}$ są liniowymi funkcjami $\boldsymbol{\epsilon}$. Liniowość $\boldsymbol{b}$ wynika z (2.15) natomiast liniowość $\boldsymbol{e}$ z równości

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \epsilon) = (\mathrm{I}_n - \boldsymbol{P})\boldsymbol{\epsilon} = \boldsymbol{M}\boldsymbol{\epsilon}.$$

Z (A.5) wynika, iż wektor $(\boldsymbol{b}, \boldsymbol{e})$ ma wielowymiarowy rozkład normalny. Z Propozycji 2.18 wiemy, iż $\boldsymbol{e}$ oraz $\boldsymbol{b}$ są nieskorelowane, co implikuje niezależność i kończy dowód. □

Znając rozkład t_j korzystamy ze standardowych metod statystycznych służących do testowania hipotezy zerowej wobec hipotezy alternatywnej. Dla przypomnienia:

- Przedział o współczynniku ufności $1-\alpha$ dla statystyki t_j (przykładowy, obustronny) dany jest przez
$$\mathrm{C} = \left[t_{n-k}^{-1}\left(\frac{\alpha}{2}\right), t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right)\right] = \left[-t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right), t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right)\right],$$
gdzie t_m^{-1} oznacza funkcję odwrotną dystrybuanty dla rozkładu t-studenta o m stopniach swobody; druga równość wynika z symetryczności rozkładu t-studenta względem zera.

- Jeżeli $t_j \notin \mathrm{C}$ to odrzucamy H_0 na rzecz H_1. W przeciwnym wypadku ($t_j \in \mathrm{C}$) mówimy, że nie mamy podstaw do odrzucenia H_0 na rzecz H_1. Przypomnijmy, iż brak odrzucenia hipotezy nie może być uznany (zazwyczaj) za podstawę do jej przyjęcia.

- Często w programach statystycznych zamiast t_j podawana jest wartość p-value, która odpowiada największemu poziomowi ufności, dla którego odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. W tym wypadku definiujemy $p = 2(1 - t_{n-k}(|t_j|))$ i odrzucamy H_0 na rzecz H_1 jeżeli $p \leq \alpha$.

Warunek $t_j \in \mathrm{C}$ możemy też zapisać przy wykorzystaniu błędu standardowego jako

$$b_j - \mathrm{SE}(b_j)\cdot t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right) < \overline{\beta}_k < b_j + \mathrm{SE}(b_j)\cdot t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right).$$

Hipotezę H_0 należy więc odrzucić, jeżeli hipotetyczna wartość $\overline{\beta}_j$ nie wpada w przedział

$$\mathrm{CI} = \left[b_j - \mathrm{SE}(b_j)\cdot t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right),\ b_j + \mathrm{SE}(b_j)\cdot t_{n-k}^{-1}\left(1-\frac{\alpha}{2}\right)\right],$$

Przedział CI często nazywa się przedziałem ufności (ang. *confidence interval*).

2.8.2 Testowanie liniowych ograniczeń na współczynniki regresji liniowej

Oprócz testowania wartości pojedynczych współczynników chcielibyśmy też testować hipotezę zerową postaci

$$\mathrm{H}_0 : \boldsymbol{R}\beta = r, \tag{2.34}$$

gdzie $\boldsymbol{R}$ i r są z góry zadane i odnoszą się do liniowych restrykcji nałożonych na zależność między współczynnikami regresji liniowej; hipoteza alternatywna jest postaci $\mathrm{H}_1 : \boldsymbol{R}\beta \neq r$. Długość wektora r będziemy oznaczać przez $\#r$; $\dim(\boldsymbol{R}) = \#r \times k$. Aby uniknąć współliniowości (np. dublujących się warunków) zakładamy, iż macierz $\boldsymbol{R}$ jest pełnego rzędu, tzn $\mathrm{rank}(\boldsymbol{R}) = \#r$.

Podstawową statystyką służącą do testowania hipotez postaci (2.34) przy założeniu (A.5) jest F-statystyka. Jej postać oraz rozkład jest podany w Propozycji 2.28.

Propozycja 2.28 (Rozkład statystyki F)**.** Załóżmy (A.1)–(A.5). Zakładając prawdziwość hipotezy zerowej $\mathrm{H}_0 : \boldsymbol{R}\beta = r$ (gdzie $\dim(\boldsymbol{R}) = \#r \times k$ oraz $\mathrm{rank}(\boldsymbol{R}) = \#r$) statystyka F wyrażona przez

$$F := \frac{(\boldsymbol{R}\boldsymbol{b} - r)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\boldsymbol{b} - r)}{\#r \cdot s^2}, \tag{2.35}$$

ma rozkład Fishera-Snedecora o stopniach swobody $\#r$ oraz $n-k$, tzn. $F(\#r, n-k)$.

Dowód. Niech $q := \boldsymbol{e}'\boldsymbol{e}/\sigma^2$ oraz $w := (\boldsymbol{R}\boldsymbol{b} - r)'[\sigma^2 \cdot \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\boldsymbol{b} - r)$. Korzystając z (2.23) oraz (2.14) dostajemy

$$s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k} = \frac{\sigma^2 q}{n-k},$$

co pozwala nam zapisać statystykę F zadaną w (2.35) jako

$$F = \frac{w}{\#r} \cdot \left(\frac{q}{n-k}\right)^{-1}.$$

Wiemy (z wykładu Statystyka), iż rozkład $F(d_1, d_2)$, tzn. rozkład Snedecora o d_1 oraz d_2 stopniach swobody, można wyrazić jako

$$\frac{X}{d_1}\left(\frac{Y}{d_2}\right)^{-1},$$

gdzie $X \sim \chi^2(d_1)$ oraz $Y \sim \chi^2(d_2)$ są niezależnymi zmiennymi losowymi. Z dowodu Propozycji 2.27 wiemy, iż $q \mid \boldsymbol{X} \sim \chi^2(n-k)$. Wystarczy zatem pokazać, iż $w \mid \boldsymbol{X} \sim \chi^2(\#r)$ oraz niezależność (warunkową) zmiennej q i w.

Niech $v = \boldsymbol{R}\boldsymbol{b} - r$. Zakładając H_0 wiemy, że $v = \boldsymbol{R}(\boldsymbol{b} - \beta)$. Korzystając z (2.32) dostajemy, iż v ma rozkład normalny o średniej 0 oraz wariancji

$$\operatorname{Var}[v \mid \boldsymbol{X}] = \operatorname{Var}[\boldsymbol{R}(\boldsymbol{b} - \beta) \mid \boldsymbol{X}] = \boldsymbol{R}\operatorname{Var}[(\boldsymbol{b} - \beta) \mid \boldsymbol{X}]\boldsymbol{R}' = \sigma^2 \cdot \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'.$$

Wynika stąd, iż

$$w = v' \operatorname{Var}[v \mid \boldsymbol{X}]^{-1} v.$$

Ponieważ $\operatorname{rank}(\boldsymbol{R}) = \#r$ oraz $\boldsymbol{X}\boldsymbol{X}'$ nie jest zdegenerowana (singularna), macierz $\sigma^2 \cdot \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$ również nie jest zdegenerowana (ćwiczenie). Korzystając z faktu, iż dla dowolnego m-wymiarowego wektora normalnego $X \sim N(\mu, \Sigma)$, gdzie Σ jest niezdegenerowana zachodzi

$$(X - \mu)'\Sigma^{-1}(X - \mu) \sim \chi^2(m), \qquad \text{(ćwiczenie)}$$

dostajemy $w \mid \boldsymbol{X} \sim \chi^2(\#r)$. Pokażemy teraz (warunkową) niezależność zmiennych w i q. Wiemy, iż w jest funkcją $\boldsymbol{b}$ natomiast q jest funkcją $\boldsymbol{e}$. W Propozycji 2.27 pokazaliśmy (warunkową) niezależność $\boldsymbol{b}$ i $\boldsymbol{e}$, co implikuje (warunkową) niezależność w i q (pod warunkiem $\boldsymbol{X}$) i kończy dowód. □

Podobnie jak wcześniej, znając rozkład F korzystamy ze standardowych metod statystycznych służących do testowania hipotezy zerowej wobec hipotezy alternatywnej. Dla dużych wartości F hipoteza zerowa powinna być odrzucona. Przy istotności statystycznej $1 - \alpha$ odrzucamy H_0 jeżeli wartość odrzuconej statystyki jest większa niż α-kwantyl rozkładu $F(\#r, n-k)$.

Warto tutaj zwrócić uwagę, iż interesuje nas tylko prawy ogon rozkładu F. Jest to związane z faktem, iż F bada różnicę wariancji, która powinna być duża, jeżeli model jest źle dopasowany (w szczególności F jest zawsze dodatnia). Aby lepiej zrozumieć ten fakt pokażmy inny sposób wyliczenia F oparty o zawężenie zbioru estymatorów do tych spełniających warunek hipotezy zerowej i zastosowanie metod związanych z zawężoną regresją liniową (ang. *restricted regression*, *restricted least squares*). Definiujemy minimalną wartość zawężonej sumy kwadratów residuów (ang. *Restricted Residual Sum of Squares*) jako

$$\mathrm{RSS}_R := \min_{\tilde{\beta} \in R} \mathrm{RSS}(\tilde{\beta}), \tag{2.36}$$

gdzie zbiór R to zbiór estymatorów $\hat{\beta}$ spełniający równanie $\boldsymbol{R}\hat{\beta} = r$. Zachodzi wtedy równość

$$F = \frac{n-k}{\#r} \cdot \frac{\mathrm{RSS}_R - \mathrm{RSS}}{\mathrm{RSS}}. \tag{2.37}$$

Co ciekawe, w praktyce statystykę F wylicza się w oparciu o wzór (2.37), a nie (2.28); czasochłonne procedury związane z bezpośrednim odwracaniem i mnożeniem macierzy można zastąpić przez szybkie algorytmy (zawężonej) regresji. Przedstawienie konkretnych metod numerycznych nie jest częścią tego wykładu.

Uwaga 2.29 (Postać zawężonego estymatora OLS)**.** Stosunkowo łatwo można pokazać, iż zawężony estymator OLS dla którego zachodzi równość w (2.36) ma postać

$$\boldsymbol{b}^* := \boldsymbol{b} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\boldsymbol{b} - r).$$

Dowód tego faktu pozostawiamy jako ćwiczenie domowe; zobacz [Gre18, Strona 126].

Uwaga 2.30 (Związek między t-statystyką, a F-statystyką)**.** Warto zaznaczyć, iż hipoteza podana w (2.34) jest uogólnieniem poprzedniego przypadku (2.33): możemy nałożyć ograniczenie tylko na jeden współczynnik i jego wartość. Przy ustalonej hipotezie zerowej $\mathrm{H}_0 : \beta_j = \overline{\beta}_j$ można pokazać, iż F-statystyka jest tak naprawdę kwadratem t-statystyki, tzn. zachodzi $F = t_j^2$.

2.9 Związek między estymatorami OLS, a estymatorami ML

Przy założeniu (A.5) znamy rozkład czynnika losowego $\boldsymbol{\epsilon}$. Pozwala nam to na znalezienie estymatorów największej wiarygodności (ang. *Maximum Likelihood (ML) estimators*) parametrów β i σ^2. Pokażemy, iż estymator OLS parametru β jest równy estymatorowi ML parametru β, oraz zachodzi ścisły związek między powiązanymi estymatorami parametru σ^2.

Główna idea metody największej wiarygodności opiera się znalezieniu parametrów, które maksymalizują prawdopodobieństwo (funkcję wiarygodności) wystąpienia zaobserwowanych wartości. Mając daną (skończenie wymiarową) przestrzeń parametrów Ψ oraz powiązaną rodzinę funkcji gęstości $\{f(\cdot;\psi), \psi \in \Psi\}$, gęstość $(\boldsymbol{y}, \boldsymbol{X})$ dla parametru $\psi \in \Psi$ możemy rozbić na gęstość brzegową $\boldsymbol{X}$ oraz warunkową gęstość $\boldsymbol{y}$, tzn.

$$f(\boldsymbol{y}, \boldsymbol{X}; \psi) = f(y|\boldsymbol{X}; \theta) \cdot f(\boldsymbol{X}; \xi),$$

gdzie θ to podzbiór wektora ψ odpowiadający warunkowej gęstości $\boldsymbol{y}$, a ξ to podzbiór wektora ψ odpowiadający rozkładowi brzegowemu $\boldsymbol{X}$.[4] W rozdziale tym będzie nas interesowało znalezienie (warunkowego) estymatora największej wiarygodności, który maksymalizuję wiarygodność związaną z warunkową gęstością $f(y|\boldsymbol{X};\theta)$; warto zwrócić przy tym uwagę, że klasyczny model regresji liniowej nie daje nam informacji o członie $f(\boldsymbol{X};\xi)$.

Przy założeniu normalności czynnika losowego (A.5), korzystając z (A.1), tzn. postaci liniowej $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$, dostajemy

$$\boldsymbol{y} \mid \boldsymbol{X} \sim N(\boldsymbol{X}\beta, \sigma^2 \mathrm{I}_n). \tag{2.38}$$

Zdefiniujmy powiązaną (warunkową) przestrzeń parametrów $\Theta := \mathbb{R}^k \times \mathbb{R}_+$, gdzie każdy element przestrzeni $\theta \in \Theta$ można przedstawić jako $\theta = (\tilde{\beta}, \tilde{\sigma}^2)$ dla pewnego ustalonego wektora współczynników regresji liniowej $\tilde{\beta}$ i wariancji błędu $\tilde{\sigma}^2$. Korzystając z (2.38) prawdziwą gęstość $\boldsymbol{y} \mid \boldsymbol{X}$ można wyrazić jako

$$f(\boldsymbol{y} \mid \boldsymbol{X}) := f(\boldsymbol{y} \mid \boldsymbol{X}; \theta_0) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\right), \tag{2.39}$$

[4] Dla uproszczenia przyjęliśmy, iż gęstość łączna oraz gęstość brzegowa X istnieje i jest dobrze określona. Rozważanie ogólnego przypadku wymagałoby dodatkowych wytłumaczeń, które pomijamy na potrzeby tego kursu.

gdzie $\theta_0 = (\beta, \sigma^2)$ jest prawdziwym (nieznanym) zestawem parametrów (β, σ^2). Powiązana funkcja log-wiarygodności (ang. *log likelihood function*) dla dowolnego parametru $\theta = (\tilde{\beta}, \tilde{\sigma}^2)$, czyli zlogarytmowany odpowiednik (2.39), w którym zastępujemy (β, σ^2) przez $(\tilde{\beta}, \tilde{\sigma}^2)$, wyrażona jest przez

$$\log L(\tilde{\beta}, \tilde{\sigma}^2) := -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta}). \tag{2.40}$$

W Propozycji 2.31 pokażemy, że przy założeniu normalności estymatory ML są ściśle związane z estymatorami OLS.

Propozycja 2.31 (Estymatory ML)**.** Załóżmy (A.1)–(A.5). Wtedy estymatorami (ML) największej wiarygodności parametrów (β, σ^2), maksymalizującymi funkcje (2.40), są estymatory dane przez

$$\hat{\beta}_{ML} := \boldsymbol{b}, \qquad \hat{\sigma}^2_{ML} := \frac{n-k}{n}s^2, \tag{2.41}$$

gdzie $\boldsymbol{b}$ i s^2 to estymatory OLS parametrów β oraz σ^2.

Dowód. Z postaci funkcji (2.40) widać, iż funkcję tą możemy najpierw zmaksymalizować względem parametru $\tilde{\beta}$, tzn. szukając minimum wartości $(\boldsymbol{y}-\boldsymbol{X}\tilde{\beta})'(\boldsymbol{y}-\boldsymbol{X}\tilde{\beta})$, a następnie, względem parametru $\tilde{\sigma}^2$. Przypominając (2.4) oraz (2.5) dostajemy

$$\min_{\tilde{\beta}}[(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta})] = \min_{\tilde{\beta}} \mathrm{RSS}(\tilde{\beta}) = \mathrm{RSS} = \boldsymbol{e}'\boldsymbol{e}, \quad \text{oraz} \quad \hat{\beta}_{ML} = \boldsymbol{b},$$

co kończy dowód faktu, iż estymator OLS jest również estymatorem ML. Skoncentrowana funkcji największej log-wiarygodności (ang. *concentrated log likelihood*) wyrażona jest więc przez

$$\log L(\boldsymbol{b}, \tilde{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}\boldsymbol{e}'\boldsymbol{e}$$

i zależy tylko od parametru $\tilde{\sigma}^2$. Definiując $\tilde{\gamma} = \tilde{\sigma}^2$ pozostaje nam obliczyć

$$\arg\max_{\tilde{\gamma}\in\mathbb{R}_+}\left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\gamma}) - \frac{1}{2\tilde{\gamma}}\boldsymbol{e}'\boldsymbol{e}\right] = \arg\min_{\tilde{\gamma}\in\mathbb{R}_+}\left[\frac{n}{2}\log(\tilde{\gamma}) + \frac{1}{2\tilde{\gamma}}\boldsymbol{e}'\boldsymbol{e}\right].$$

Zauważając, iż $\boldsymbol{e}'\boldsymbol{e}$ nie zależy od $\tilde{\gamma}$ dostajemy

$$\frac{\partial \log L(\boldsymbol{b}, \tilde{\gamma})}{\partial \tilde{\gamma}} = \frac{-n\tilde{\gamma} + \boldsymbol{e}'\boldsymbol{e}}{2\tilde{\gamma}^2}. \tag{2.42}$$

Przyrównując (2.42) do zera dostajemy równość

$$\hat{\gamma} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n} = \frac{n-k}{n}\frac{\mathrm{RSS}}{n-k} = \frac{n-k}{n}s^2,$$

co implikuje $\hat{\sigma}^2_{ML} = \frac{n-k}{n}s^2$ i kończy dowód. □

Pokażemy następnie, że przy założeniach (A.1)–(A.5) estymator OLS parametru β jest najlepszym nieobciążonym estymatorem parametru β. Warto zauważyć, iż stwierdzenie to jest mocniejsze od Twierdzenia Gaussa-Markowa (Twierdzenia 2.17), które odnosiło się do klasy nieobciążonych oraz liniowych estymatorów parametrów β. Zanim to jednak zrobimy, przypomnijmy podstawowe informacje o macierzy informacji Fishera oraz nierówność Rao-Craméra.

Uwaga 2.32 (Macierz informacji Fishera i nierówność Rao-Craméra)**.** Dla wektora losowego $\boldsymbol{z}$, na przykład wektora obserwacji $(\boldsymbol{y}, \boldsymbol{X})$, o gęstości łącznej $f(\boldsymbol{z};\theta)$, gdzie $\theta \in \Theta$ jest nieznanym parametrem, definiujemy powiązaną funkcję wiarygodności $L(\hat{\theta};\boldsymbol{z}) = f(\boldsymbol{z};\hat{\theta})$, $\hat{\theta} \in \Theta$. Macierz informacji Fishera mierzy ilość informacji o parametrach jaką niosą obserwacje. Dana jest ona wzorem

$$\boldsymbol{I}(\theta) = \mathbb{E}[s(\theta)s(\theta)'],$$

gdzie $s(\tilde{\theta}) = \frac{\partial}{\partial\tilde{\theta}} \log L(\tilde{\theta};\boldsymbol{z})$, $\hat{\theta} \in \Theta$, jest miarą zmienności funkcji wiarygodności (*ang. score function*), a $\boldsymbol{z}$ jest wektorem losowym obserwacji.[5] Przy pewnych ogólnych założeniach nałożonych na gęstość f, mając dany nieobciążony estymator parametrów $\hat{\theta}(\boldsymbol{X})$ o skończonej macierzy wariancji-kowariancji, dostajemy tzw. dolną nierówność Rao-Craméra

$$\mathrm{Var}[\hat{\theta}(\boldsymbol{z})] \geq \boldsymbol{I}(\theta)^{-1}, \tag{2.43}$$

która daje dolne ograniczenie na wariancję estymatora $\hat{\theta}$.[6] Dla odpowiednio regularnej funkcji f, macierz informacji Fishera można wyrazić w postaci

$$\boldsymbol{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log L(\boldsymbol{z};\theta)\right]. \tag{2.44}$$

Dokładniejsze informacje na ten temat można znaleźć np. w książce [Ame85].

Propozycja 2.33 (Estymator OLS jest najlepszym nieobciążonym estymatorem β)**.** Załóżmy (A.1)–(A.5). Wtedy estymator $\boldsymbol{b}$ jest najlepszym nieobciążonym estymatorem parametru β, czyli tzw. estymatorem BUE (ang. *Best Unbiased Estimator*). Innymi słowy, dla każdego nieobciążonego estymatora $\hat{\boldsymbol{\beta}}$ parametru β (niekoniecznie liniowego) dostajemy

$$\mathrm{Var}[\boldsymbol{b} \mid \boldsymbol{X}] \leq \mathrm{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}]. \tag{2.45}$$

Dowód. Aby dowieść (2.45) wystarczy pokazać, iż estymator OLS osiąga dolne ograniczenie nierówności Rao-Craméra (2.43). Dla modelu regresji liniowej spełniającego założenia (A.1)–(A.5) funkcja wiarygodności podana w (2.39) spełnia założenia regularności z Uwagi 2.32. Dowód tego faktu można znaleźć w Rozdziale 1.3 w [Ame85]. Warunkowa macierz informacji Fishera (2.44) w modelu regresji liniowej zadana jest więc przez

$$\boldsymbol{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log L(\theta)\middle|\boldsymbol{X}\right], \tag{2.46}$$

gdzie $\theta = (\beta, \sigma^2)$ to prawdziwy parametr modelu, a funkcja wiarygodności $\log L$ zadana jest przez

$$\log L(\tilde{\beta}, \tilde{\sigma}^2) := -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta}).$$

Aby policzyć (2.46) wyznaczmy najpierw Hesjan (macierz drugich pochodnych) w punkcie (β, σ^2), tzn. policzmy wartości macierzy

$$\begin{bmatrix} \frac{\partial^2}{\partial\tilde{\beta}\partial\tilde{\beta}'} \log L(\beta,\sigma^2) & \frac{\partial^2}{\partial\tilde{\beta}\partial\tilde{\sigma}^2} \log L(\beta,\sigma^2) \\ \frac{\partial^2}{\partial\tilde{\sigma}^2\partial\tilde{\beta}'} \log L(\beta,\sigma^2) & \frac{\partial^2}{\partial\tilde{\sigma}^2\partial\tilde{\sigma}^2} \log L(\beta,\sigma^2) \end{bmatrix}.$$

[5] Warto zwrócić uwagę, iż operator wartości oczekiwanej liczony jest względem miary $\mathbb{P}_\theta$.
[6] Warto zauważyć, iż macierz informacji liczona jest dla prawdziwej wartości $\theta \in \Theta$.

Dokonując prostych obliczeń otrzymujemy pierwsze pochodne cząstkowe funkcji $\log L$ w punkcie (β, σ^2) dane przez

$$\frac{\partial}{\partial \tilde{\beta}} \log L(\beta, \sigma^2) = \frac{1}{\sigma^2} \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\beta),$$
$$\frac{\partial}{\partial \tilde{\sigma}^2} \log L(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta),$$

oraz drugie pochodne cząstkowe w punkcie (β, σ^2) dane przez

$$\frac{\partial^2}{\partial \tilde{\beta} \partial \tilde{\beta}'} \log L(\beta, \sigma^2) = -\frac{1}{\sigma^2} \boldsymbol{X}'\boldsymbol{X}, \tag{2.47}$$
$$\frac{\partial^2}{\partial \tilde{\sigma}^2 \partial \tilde{\sigma}^2} \log L(\beta, \sigma^2) = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\tilde{\sigma}^2)^3}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta), \tag{2.48}$$
$$\frac{\partial^2}{\partial \tilde{\sigma}^2 \partial \tilde{\beta}'} \log L(\beta, \sigma^2) = -\frac{1}{(\sigma^2)^2} \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\beta). \tag{2.49}$$

Obkładając (2.47)–(2.49) warunkową wartością oczekiwaną, zauważając, iż $\boldsymbol{y} - \boldsymbol{X}\beta = \epsilon$ oraz $\mathbb{E}[\boldsymbol{\epsilon}'\boldsymbol{\epsilon}|\boldsymbol{X}] = n\sigma^2$, a następnie wstawiając otrzymane wartości do (2.46) dostajemy ostatecznie

$$\boldsymbol{I}(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{n}{2\sigma^4} \end{bmatrix}. \tag{2.50}$$

Odwracając macierz (2.51) dostajemy

$$\boldsymbol{I}(\theta)^{-1} = \begin{bmatrix} \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{2\sigma^4}{n} \end{bmatrix}. \tag{2.51}$$

Korzystając z Propozycji 2.15 oraz nierówności Rao-Craméra (2.43) dla dowolnego nieobciążonego estymatora $\hat{\boldsymbol{\beta}}$ parametru β dostajemy

$$\mathrm{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] \geq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \mathrm{Var}[\boldsymbol{b} \mid \boldsymbol{X}],$$

co kończy dowód. □

Uwaga 2.34 (Własność BUE dla estymatora OLS parametru σ^2)**.** Estymator ML parametru σ^2 jest obciążony, więc nie można zastosować do niego nierówności Rao-Craméra. Wiemy natomiast, iż estymator OLS parametru σ^2 jest nieobciążony. Pozostaje pytanie, czy jest on najlepszy w podobnym sensie co estymator $\boldsymbol{b}$. Można pokazać (ćwiczenie), iż

$$\mathrm{Var}[s^2|\boldsymbol{X}] = \frac{2\sigma^2}{n-k},$$

więc dolne ograniczenie w (2.51) nie jest osiągnięte. Można jednak pokazać, iż nie istnieje nieobciążony estymator σ^2, który ma wariancję mniejszą niż s^2. Dowód tego faktu wykracza poza materiał tego wykładu; zobacz [Rao73].

3 Przykłady innych modeli liniowych

W rozdziale tym przedstawimy (wybrane) przykłady modeli liniowych które są ściśle związane z modelem regresji liniowej.

3.1 Uogólniony model regresji liniowej i estymator GLS

W klasycznym modelu regresji liniowej zakładaliśmy, iż błędy mają stałą wariancję i są od siebie niezależne – założenie (A.4) mówiło nam, że warunkowa macierz wariancji-kowariancji dana jest przez $\mathrm{Var}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \sigma^2 \mathrm{I}_n$, dla pewnego nieznanego parametru $\sigma^2 > 0$. W praktyce czynnik losowy (błąd) często nie jest (warunkowo) homoskedastyczny, a błędy mogą być skorelowane.

Definicja 3.1 (Uogólniony model regresji liniowej)**.** Model spełniający założenia (A.1)–(A.3), w którym (warunkowa) unormowana macierz wariancji-kowariancji błędów dana przez

$$\boldsymbol{V}(\boldsymbol{X}) := \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \boldsymbol{X}]/\sigma^2, \tag{3.1}$$

jest znana oraz niezdegenerowana, nazywamy **uogólnionym modelem regresji liniowej** (ang. *generalized least squares model*).

Zazwyczaj, tak jak i w poprzednim wypadku, parametr σ^2 jest nieznany. Zastąpienie założenia (A.4) przez (3.1) powoduje, że standardowy estymator OLS nie spełnia większości własności przedstawionych w poprzednim rozdziale. Przykładowo, nie jest on estymatorem BLUE; przy dodatkowym założeniu (A.5) rozkłady t-statystyki oraz F-statystyki nie odpowiadają rozkładom t-studenta oraz Fishera-Snedecora. Warto jednak zwrócić uwagę, że estymator OLS wciąż jest nieobciążony.

3.1.1 Estymator GLS

Pokażemy, że dla znanej wartości $\boldsymbol{V}(X)$ istnieje estymator BLUE parametru β. Aby to pokazać, sprowadzimy uogólniony model regresji liniowej do modelu spełniającego założenie (A.1)–(A.4). Dla uproszczenia będziemy używać notacji $\boldsymbol{V}$ zamiast $\boldsymbol{V}(\boldsymbol{X})$. Ponieważ $\boldsymbol{V}$ jest (niezdegenerowaną) symetryczną macierzą dodatnio określoną, istnieje niezdegenerowana macierz $\boldsymbol{C}$ taka, że

$$\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}.^{7} \tag{3.2}$$

Transformacja wyjściowego modelu $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\epsilon})$ opiera się o liniowe przekształcenie względem $\boldsymbol{C}$.

Propozycja 3.2 (Transformacja modelu GLS do OLS)**.** Niech $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\epsilon})$ tworzy uogólniony model regresji liniowej z powiązaną macierzą $\boldsymbol{C}$. Wtedy, model $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\epsilon}})$, gdzie

$$\tilde{\boldsymbol{y}} \equiv \boldsymbol{C}\boldsymbol{y}, \quad \tilde{\boldsymbol{X}} \equiv \boldsymbol{C}\boldsymbol{X}, \quad \tilde{\boldsymbol{\epsilon}} \equiv \boldsymbol{C}\boldsymbol{\epsilon}. \tag{3.3}$$

spełnia założenia (zwykłego) modelu regresji liniowej (A.1)–(A.4). Dodatkowo, jeżeli $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\epsilon})$ spełnia (A.5), to $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\epsilon}})$ również spełnia (A.5).

Dowód. Ponieważ transformacja (3.3) jest liniowa, spełniony jest warunek (A.1) oraz (A.2) dla modelu $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\epsilon}})$, tzn. zachodzi $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\beta + \tilde{\boldsymbol{\epsilon}}$ oraz macierz $\tilde{\boldsymbol{X}}$ jest pełnego rzędu (ćwiczenie do

[7] Macierz $\boldsymbol{C}$ nie musi być jednoznacznie określona – wybieramy dowolnego reprezentanta.

domu). Aby dowieść (A.3) wystarczy zauważyć, iż ponieważ $\boldsymbol{C}$ jest niezdegenerowana, dostajemy $\sigma(\boldsymbol{X}) = \sigma(\tilde{\boldsymbol{X}})$, oraz skorzystać z liniowości warunkowej wartości oczekiwanej. Istotnie, korzystając z założenia (A.2) dla modelu $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\epsilon})$ dostajemy

$$\mathbb{E}[\tilde{\boldsymbol{\epsilon}} \mid \tilde{\boldsymbol{X}}] = \mathbb{E}[\tilde{\boldsymbol{\epsilon}} \mid \boldsymbol{X}] = \boldsymbol{C}\mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = 0.$$

Dowód własności (A.4) opiera się na obserwacji $\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' = \mathrm{I}_n$, która wynika z $(\boldsymbol{C}')^{-1}\boldsymbol{V}^{-1}\boldsymbol{C}^{-1} = \mathrm{I}_n$. Dostajemy

$$\mathbb{E}[\tilde{\boldsymbol{\epsilon}}\tilde{\boldsymbol{\epsilon}}' \mid \tilde{\boldsymbol{X}}] = \mathbb{E}[(\boldsymbol{C}\boldsymbol{\epsilon})(\boldsymbol{\epsilon}'\boldsymbol{C}') \mid \boldsymbol{X}] = \boldsymbol{C}\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \boldsymbol{X}]\,\boldsymbol{C}' = \boldsymbol{C}\sigma^2\boldsymbol{V}\boldsymbol{C}' = \sigma^2\mathrm{I}_n.$$

Własność (A.5) wynika z faktu, iż transformacja (3.3) jest liniowa, więc w szczególności zachowuje ona wielowymiarowy rozkład normalny czynnika losowego $\tilde{\boldsymbol{\epsilon}}$, przy założeniu normalności $\boldsymbol{\epsilon}$. □

Estymator OLS dla przekształconego modelu (3.3) nazywamy estymatorem GLS (ang. *generalised least squares*) dla wyjściowego modelu.

Propozycja 3.3 (Estymator GLS parametru β)**.** Dla modelu uogólnionej regresji liniowej z powiązaną macierzą $\boldsymbol{V}$ estymatorem GLS parametru β nazywamy estymator OLS dla powiązanego modelu (3.3). Wyraża się on wzorem

$$\widehat{\beta}_{\mathrm{GLS}} := (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{3.4}$$

Warunkowa wariancja estymatora GLS wynosi $\mathrm{Var}[\widehat{\beta}_{\mathrm{GLS}} \mid \boldsymbol{X}] = \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$.

Dowód. Niech C będzie dowolną macierzą spełniającą (3.2). Z Propozycji 2.9 dostajemy jawny wzór na estymator OLS dla przekształconego modelu (3.3). Z własności $\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$ otrzymujemy

$$\widehat{\beta}_{\mathrm{GLS}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} = ((\boldsymbol{C}\boldsymbol{X})'\boldsymbol{C}\boldsymbol{X})^{-1}(\boldsymbol{C}\boldsymbol{X})'\boldsymbol{C}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{3.5}$$

Wariancję estymatora $\widehat{\beta}_{\mathrm{GLS}}$ liczymy wykorzystując własność $\mathrm{Var}[\boldsymbol{y}|\boldsymbol{X}] = \mathrm{Var}[\boldsymbol{\epsilon}|\boldsymbol{X}]$ oraz symetryczność $\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$. Dostajemy

$$\begin{aligned}
\mathrm{Var}[\widehat{\beta}_{\mathrm{GLS}} \mid \boldsymbol{X}] &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\ \mathrm{Var}[\boldsymbol{y} \mid \boldsymbol{X}]\,((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})' \\
&= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\ \mathrm{Var}[\boldsymbol{y} \mid \boldsymbol{X}]\,\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\quad(\sigma^2\boldsymbol{V})\quad\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\
&= \sigma^2\,(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.
\end{aligned}$$

□

Dla uogólnionego modelu regresji liniowej zarówno estymator OLS, jak i GLS jest nieobciążony. Wiemy jednak, iż (warunkowa) wariancja estymatora GLS jest mniejsza, co wynika z Twierdzenia Gaussa-Markova zastosowanego do przekształconego modelu (3.3). Podsumujmy teraz własności estymatora GLS.

Propozycja 3.4 (Własności estymatora GLS parametru β)**.** Dla uogólnionego modelu regresji liniowej estymator $\widehat{\beta}_{\mathrm{GLS}}$ zadany przez (3.4) jest estymatorem

- Nieobciążonym, tzn. $\mathbb{E}[\widehat{\beta}_{\mathrm{GLS}}|\boldsymbol{X}] = \beta$;
- BLUE, tzn. spośród nieobciążonych i liniowych estymatorów β estymator $\widehat{\beta}_{\mathrm{GLS}}$ ma najmniejszą wariancję zadaną przez $\mathrm{Var}[\widehat{\beta}_{\mathrm{GLS}} \mid \boldsymbol{X}] = \sigma^2 \cdot (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$.

Dowód Propozycji (3.4) pozostawiamy jako ćwiczenia.

Uwaga 3.5 (GLS dla nieznanej macierzy $\boldsymbol{V}$)**.** Możemy również rozważyć model uogólnionej regresji liniowej zakładając, iż macierz $\boldsymbol{V}$ jest nieznana. Modele takie nazywa się modelami FGLS (ang. *Feasible Generalized Least Squares*). Sposoby estymacji macierzy $\boldsymbol{V}$ oraz dokładny opis modeli FLGS wykracza poza materiał tego wykładu.

Uwaga 3.6 (Ogólna uwaga o metodzie GLS)**.** Metoda GLS wydaje się na pierwszy rzut oka atrakcyjniejsza od OLS. Kluczowym założeniem obu metod jest jednak warunkowa egzogeniczność czynnika losowego, tzn założenie (A.3). W praktycznych zastosowaniach (np. dla szeregów czasowych) założenie to często jest niespełnione i wymaga zastosowania innego typu modeli. W rozdziale poświęconemu asymptotycznym własnościom pokażemy, iż estymator OLS spełnia pewne własności dla dużych próbek (zgodność, czy asymptotyczna normalność) przy słabszych założeniach nałożonych na czynnik losowy - własności te nie są spełnione przez estymator GLS (np. dla pewnej klasy modeli korekta autokorelacji reszt może doprowadzić do braku zgodności estymatora). Dokładny opis różnic między estymacją OLS, a GLS można znaleźć w Rozdziale 1.6 w [Hay00].

3.2 Ważony model regresji liniowej i estymator WLS

Ważony model regresji liniowej (ang. *weighted linear regression model*) jest specjalnym przypadkiem uogólnionego modelu regresji liniowej, w którym nie występuje zależność między różnymi czynnikami losowymi, tzn. macierz $\boldsymbol{V}$ jest diagonalna. Dla danej diagonalnej macierzy $\boldsymbol{C}$ oraz $i = 1, 2, \ldots, n$, oznaczmy przez $v_i(\boldsymbol{X})$ i-ty element na przekątnej macierzy $\boldsymbol{V}$. Wtedy warunek (3.1) można przedstawić w postaci niemacierzowej jako

$$\mathbb{E}[\epsilon_i \epsilon_j \mid \boldsymbol{X}] = \begin{cases} \sigma^2 v_i(\boldsymbol{X}) & \text{gdy } i = j \\ 0 & \text{gdy } i \neq j \end{cases}, \quad i, j = 1, 2, \ldots, n.$$

Łatwo zauważyć, iż powiązana macierz $\boldsymbol{C}$ określona w (3.2) również jest diagonalna, gdzie i-ty element przekątnej to $1/v_i(\boldsymbol{X})$. Z (3.3) wynika, iż przekształcona zmienna objaśniana oraz zmienne objaśniające mają postać

$$\tilde{\boldsymbol{y}}_i = \frac{y_i}{\sqrt{v_i(\boldsymbol{X})}}, \quad \tilde{x}_i = \frac{x_i}{\sqrt{v_i(\boldsymbol{X})}}, \quad i = 1, 2, \ldots, n.$$

Estymacja parametrów modelu polega więc na **zważeniu** obserwacji poprzez dane dla niej (unormowane) odchylenie standardowe, a następnie policzenie standardowego estymatora OLS.

W szczególności, zakładając, iż $(y_i, x_i)_{i=1}^n$ jest próbką prostą z wektora $(\boldsymbol{y}, \boldsymbol{X})$ dostajemy od razu bezwarunkową homoskedastyczność, tzn. warunek $\mathbb{E}[\epsilon_i^2] = \sigma^2$, ale nie implikuje on warunkowej homoskedastyczności. Estymator GLS może być tutaj pomocny, aby zwiększyć jakość estymacji. Warto zauważyć, że mając daną próbkę prostą istnieje funkcja v, dla której zachodzi $v(x_i) = v_i(\boldsymbol{X})$. Zadanie $\boldsymbol{V}$ sprowadza się więc do określenia funkcji k-zmiennych $v(\cdot)$.

3.3 Przykład uogólnionego modelu liniowego GLM: regresja logistyczna

Oprócz modeli regresji liniowej opartych o metodę najmniejszych kwadratów w statystyce można spotkać wiele innych modeli mających postać liniową. Kluczowe jest równanie zadające zależność między zmienną objaśnianą, a zmiennych objaśniającymi. Dla uogólnionych modeli liniowych (ang. *Generalized Linear Models*) równanie to ma postać

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = f(\boldsymbol{X}\beta), \tag{3.6}$$

gdzie f^{-1} odpowiada tzw. funkcji łączącej (ang. *link function*).

W tym podrozdziale przedstawimy (pokrótce i opisowo) intuicję stojącą za przykładowym modelem GLM. Załóżmy, iż zmienna objaśniana $\boldsymbol{y}$ jest zmienną postaci binarnej (0/1), gdzie 1 oznacza sukces, a 0 porażkę. Mając macierz danych $\boldsymbol{X}$, klasyczne metody regresji liniowej sprowadzały by się do rozpatrywania równań postaci

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta.$$

Regresja tego typu wydaje się jednak nie mieć sensu, gdyż w przypadku binarnej zmiennej $\boldsymbol{y}$, realizacje warunkowej wartości oczekiwanej $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}]$ powinny być w przedziale $[0, 1]$. Aby temu zaradzić możemy wprowadzić funkcję $f\colon \mathbb{R} \to [0, 1]$ oraz rozważyć nieliniowy model (3.6). Zauważając, iż $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \mathbb{P}[\boldsymbol{y} = 1|\boldsymbol{X}]$ określa nam warunkowe prawdopodobieństw sukcesu, naturalnym kandydatem wydaje się tzw. funkcja logistyczna postaci

$$f(x) = \frac{1}{1 + e^{-x}}.$$[8]

W szczególności $f^{-1}(p) = \ln p/(1 - p)$ odpowiada tzw. funkcji logitowej (ang. *logit link function*), pozwalającej na wyrażenie prawdopodobieństwa przez logarytm tzw. funkcji szansy (ang. *log-odds*). Dodatkowym utrudnieniem jest fakt, iż wartość $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}]$ nie jest wartością obserwowaną – tzn. do dyspozycji mamy tylko ciąg obserwacji 0/1. Możemy sobie jednak z tym poradzić poprzez maksymalizację funkcji największej wiarygodności. Algorytm estymacji parametrów może wyglądać następująco:

- Dla ustalonego β oraz $i = 1, 2, \ldots, n$ liczymy wartości $\hat{p}_i(\beta) := f(x_i \cdot \beta)$ otrzymując predykcję prawdopodobieństw sukcesu dla zmiennych $\boldsymbol{y}$.
- Mając dane konkretne realizacje (y_i) oraz prawdopodobieństwa $\hat{p}_i(\beta)$ funkcja wiarygodności dana jest przez
$$L(\beta) = \prod_{i=1}^{n} \left[\hat{p}_i(\beta)^{y_i} \cdot (1 - \hat{p}_i(\beta))^{1-y_i}\right].$$
- Maksymalizując funkcję $\log L$ dostajemy optymalny zestaw parametrów. Estymacja optymalnego parametru β jest dosyć trudnym zadaniem (np. nie istnieją jawne wzory na estymatory o dobrych własnościach); można korzystać np. z algorytmu Newtona-Raphsona.

Ze względu na swoją użyteczność metoda regresji logistycznej doczekała się wielu dedykowanych technik oceny jakości modelu. Jest ona często wykorzystywana np. problemach związanych z uczeniem maszynowym. Przekracza to jednak zakres materiału na tym kursie.

[8] Oczywiście jest wiele tego typu funkcji. Wybór funkcji logitowej podyktowany jest jej naturalną interpretacją oraz dobrymi własnościami. Innym przykładem może być tzw. funkcja probitowa.

4 Asymptotyczny model regresji liniowej

W rozdziale tym zajmiemy się asymptotycznymi własnościami estymatorów OLS takimi jak zgodność i asymptotyczna normalność. Będziemy korzystać z wielu definicji, własności i twierdzeń podanych na wykładzie *Rachunek Prawdopodobieństwa* oraz *Statystyka* (1 oraz 2). W szczególności dotyczy to różnych typów zbieżności zmiennych losowych i zależności między nimi, praw wielkich liczb oraz centralnych twierdzeń granicznych. Zmienimy (i częściowo osłabimy) również zbiór założeń modelu regresji liniowej, dopasowując je do modelu asymptotycznego. Zanim do tego przejdziemy, zdefiniujmy kilka podstawowych pojęć oraz pokażmy proste własności, których będziemy używać w tym rozdziale; większość z nich związana jest z *procesami stochastycznymi* w czasie dyskretnym.[9]

4.1 Wstępne definicje i przypomnienie

Zacznijmy od zdefiniowania *procesu stochastycznego* w czasie dyskretnym. Dla uproszczenie wszystkie definicje podane będą dla zmiennych losowych indeksowanych liczbami naturalnymi.

Definicja 4.1 (Proces stochastyczny)**. Procesem stochastycznym** nazywamy dowolny ciąg zmiennych losowych $(Z_i)_{i\in\mathbb{N}}$ określonych na tej samej przestrzeni probabilistycznej. Jeżeli liczby $i \in \mathbb{N}$ odpowiadają kolejnym chwilom w czasie, wtedy proces stochastyczny nazywamy **szeregiem czasowym**.

Często będziemy używać pojęcia *szeregu czasowego* zarówno w odniesieniu do zmiennych losowych, jak ich konkretnych realizacji – patrz definicja poniżej.

Definicja 4.2 (Realizacja procesu)**. Realizacją (trajektorią) procesu dla zdarzenia elementarnego** $\omega \in \Omega$ nazywamy ciąg liczb rzeczywistych będących wartością procesu dla tego elementu przestrzeni probabilistycznej, tzn. ciąg liczb $(Z_i(\omega))_{i\in\mathbb{N}}$.

Zdefiniujmy teraz podstawowe klasy procesów stochastycznych, których będziemy używać w kontekście modelu regresji liniowej. Zacznijmy od pojęcia stacjonarności.

Definicja 4.3 (Proces stacjonarny)**.** Niech $(Z_i)_{i\in\mathbb{N}}$ będzie procesem stochastycznym. Mówimy, że proces $(Z_i)_{i\in\mathbb{N}}$ jest

- **(Silnie) stacjonarny** (ang. *stationary*), gdy dla dowolnego $n \in \mathbb{N}$, $(i_1, i_2, \ldots, i_n) \in \mathbb{N}^n$ oraz $h \in \mathbb{N}$ wektory losowe $(Z_{i_1}, \ldots, Z_{i_n})$ oraz $(Z_{i_1+h}, \ldots, Z_{i_n+h})$ mają ten sam rozkład.
- **Słabo stacjonarny** (ang *weakly stationary* albo *covariance stationary*), gdy dla dowolnych liczb $i_1, i_2, h \in \mathbb{N}$ zachodzi $E[Z_{i_1}] = E[Z_{i_1+h}]$ oraz $\mathrm{Cov}(Z_{i_1}, Z_{i_2}) = \mathrm{Cov}(Z_{i_1+h}, Z_{i_2+h})$.

Proces jest słabo stacjonarny wtedy i tylko wtedy, gdy dla dowolnych liczb $i, j \in \mathbb{N}$, $i < j$, zachodzi $E[Z_i] = E[Z_j]$ oraz **funkcja auto-kowariancji** (ang. *auto-covariance function*) procesu dana przez $K(i, j) := \mathrm{Cov}(Z_i, Z_j)$ zależy tylko od różnicy $i - j$. Dla procesów (słabo) stacjonarnych często wprowadza się funkcję $\tilde{K}(h) := \mathrm{Cov}(Z_1, Z_{1+h})$, którą będziemy nazywać **uproszczoną funkcją auto-kowariancji**.

[9] Dokładniejsza analiza tego typu zagadnień oraz definicje dla czasu ciągłego będą podane później, na przedmiocie *Procesy Stochastyczne*.

Podstawowym przykładem procesu (silnie) stacjonarnego jest próba prosta. Proces słabo stacjonarny o zerowej wartości średniej oraz zerowej wartości uproszczonej funkcji auto-kowariancji (dla $h > 0$) nazywamy *szumem* (ang. *white noise*). Oczywiście tak określony szum nie musi być procesem silnie stacjonarnym.[10]

Stacjonarność mówi nam, że rozkład procesu nie zależy od przesunięcia w czasie. Dla procesów stacjonarnych często będziemy używać uproszczonej notacji do opisu momentów i ich przekształceń używając wyrażeń typu $\mathbb{E}[Z_i]$, tzn. bez kwantyfikowania indeksu i. Dla procesów stacjonarnych, wystarczy zadać średnią dla jednego indeksu czasu (np. $i = 1$); dla zwiększenia przejrzystości będziemy jednak pozostawiać wskaźnik i w wielu sformułowaniach.

Kolejną interesującą własnością jest tzw. ergodyczność.

Definicja 4.4 (Proces ergodyczny)**.** Niech $(Z_i)_{i\in\mathbb{N}}$ będzie stacjonarnym procesem stochastycznym. Mówimy, że proces $(Z_i)_{i\in\mathbb{N}}$ jest **ergodyczny** (ang. *ergodic*), gdy dla $i, k, l \in \mathbb{N}$ oraz dowolnych dwóch ograniczonych (i mierzalnych) funkcji $f\colon \mathbb{R}^k \to \mathbb{R}$, $g\colon \mathbb{R}^l \to \mathbb{R}$ zachodzi

$$\lim_{n\to\infty} |\mathbb{E}[f(Z_i, \ldots, Z_{i+k}) g(Z_{i+n}, \ldots, Z_{i+l+n})]| = |\mathbb{E}[f(Z_1, \ldots, Z_k)]| \cdot |\mathbb{E}[g(Z_1, \ldots, Z_l)]| .$$

Ergodyczność mówi nam, iż proces jest asymptotycznie niezależny, tzn. dowolne dwa wektory stochastyczne oddalone od siebie dostatecznie daleko są prawie niezależne. Własność ta będzie istotna szczególnie w odniesieniu do tzw. twierdzenia ergodycznego dla procesów stacjonarnych, które można traktować jako uogólnienie prawa wielkich liczb.

Twierdzenie 4.5 (Twierdzenie Ergodyczne)**.** Niech $(Z_i)_{i\in\mathbb{N}}$ będzie stacjonarnym procesem ergodycznym o średniej $\mathbb{E}[Z_i] = \mu$. Wtedy

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \to \mu, \quad (p.n.).$$

Dowód Twierdzenia Ergodycznego 4.5, które jest specjalnym przypadkiem tzw. *Ergodycznego Twierdzenia Birkhoffa* wykracza poza materiał tego wykładu; można go znaleźć np. w [Bil08]. Łatwo zauważyć, że Twierdzenie 4.5 odnosi się również do wyższych momentów (jeżeli tylko istnieją).

Wprowadźmy następnie pojęcie martyngału oraz procesu różnic martyngałowych.

Definicja 4.6 (Własności martyngałowe)**.** Niech $(Z_i)_{i\in\mathbb{N}}$ będzie całkowalnym procesem stochastycznym. Mówimy, że proces $(Z_i)_{i\in\mathbb{N}}$ jest:

- **Martyngałem** (ang. *martingale*) jeżeli dla dowolnego $i \geq 2$ zachodzi

$$\mathbb{E}[Z_i \mid Z_{i-1}, \ldots, Z_1] = Z_{i-1}.^{a}$$

- **Ciągiem różnic martyngałowych** (ang. *martingale difference sequence*), oznaczanym w skrócie jako **MDS**, jeżeli dla dowolnego $i \geq 2$ zachodzi

$$\mathbb{E}[Z_i \mid Z_{i-1}, \ldots, Z_1] = 0.$$

[a] można wprowadzić bardziej ogólną definicję martyngału względem zadanej (dyskretnej) filtracji, do której proces jest adaptowany; nie będzie to jednak potrzebne w tym kursie.

[10] Przykład: $X \sim U[0, 2\pi]$ oraz $Z_i := \cos(iX)$, dla $i \in \mathbb{N}$.

Oczywiście wszystkie podane w tym rozdziale definicje można uogólnić na wielowymiarowe procesy stochastyczne, tzn. ciągi składające się z wektorów losowych. Następne twierdzenie łączy nam własność MDS ze stacjonarnością oraz ergodycznością. Można je traktować jako pewną formę uogólnienie centralnego twierdzenia granicznego (Lindeberga-Levy'ego).

Twierdzenie 4.7 (Centralne Twierdzenie Graniczne dla MDS)**.** Niech $(Z_i)_{i\in\mathbb{N}}$ będzie wielowymiarowym, stacjonarnym i ergodycznym procesem MDS o skończonej macierzy wariancji-kowariancji $\Sigma = \mathbb{E}[Z_1 Z_1']$. Wtedy zachodzi

$$\sqrt{n}\cdot\overline{Z} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \xrightarrow{d} N(0,\Sigma), \quad \text{(gdy } n\to\infty\text{, według rozkładów)}.$$

4.2 Założenia asymptotycznego modelu regresji liniowej

Podamy teraz założenia definiujące asymptotyczny model regresji liniowej. Będą one odnosiły się do $(k+1)$-wymiarowego procesu stochastycznego $(\boldsymbol{y},\boldsymbol{x}) = ((y_i,x_i))_{i\in\mathbb{N}}$, gdzie $k\in\mathbb{N}$ oznacza liczbę elementów procesu stochastycznego $\boldsymbol{x}$ odpowiadającemu informacji.[11] Warto zwrócić uwagę, iż proces $(\boldsymbol{y},\boldsymbol{x})$ można traktować jako ciąg modeli regresji liniowych.[12] Podobnie jak wcześniej, będziemy używać oznaczeń $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_i)_{i\in\mathbb{N}}$ dla (nieznanego) czynnika losowego (procesu stochastycznego) oraz β dla k-wymiarowego wektor współczynników regresji liniowej. Przedstawmy teraz podstawowe założenia asymptotycznego modelu regresji liniowej:

(B.1) **Liniowa zależność** (ang. *linearity*). Model określa liniową zależność między zmienną objaśnianą, a zmiennymi objaśniającymi. Dla $i\in\mathbb{N}$ mamy zależność $y_i = x_i'\beta + \epsilon_i$.

(B.2) **Ergodyczna stacjonarność** (ang. *ergodic stationarity*). Proces stochastyczny $(\boldsymbol{y},\boldsymbol{x})$ jest stacjonarny i ergodyczny.

(B.3) **Ortogonalność błędów względem obserwacji** (ang. *predetermined regressors*). Czynnik losowy $\boldsymbol{\epsilon}$ jest ortogonalny względem wektora $\boldsymbol{x}$, tzn. dla każdego $i\in\mathbb{N}$ zachodzi

$$\mathbb{E}[x_i\epsilon_i] = \mathbb{E}[x_i(y_i - x_i'\beta)] = 0.$$

(B.4) **Brak współliniowości** (ang. *rank condition*). Rząd $k\times k$ wymiarowej macierzy $\Sigma_{xx} := \mathbb{E}[x_i x_i']$ jest pełny, tzn. macierz ta jest niezdegenerowana (i skończona).[13]

(B.5) **Własność MDS**. Proces stochastyczny $\boldsymbol{g} := (g_i)_{i\in\mathbb{N}}$, gdzie $g_i = x_i\cdot\epsilon_i$, dla $i\in\mathbb{N}$, jest procesem MDS o niezdegenerowanej macierzy $S := \mathbb{E}[g_i g_i']$.

[11] Oznaczenie $\boldsymbol{x}$ na k-wymiarowy wektor stochastyczny jest wprowadzone, aby rozróżnić model asymptotyczny od skończonego.

[12] tzn. dla każdego $n\in\mathbb{N}$, pierwsze n elementów procesu $(\boldsymbol{y},\boldsymbol{x})$ może odpowiadać zmiennym w skończonym modelu regresji liniowej.

[13] Proces jest stacjonarny, więc Σ_{xx} jest dobrze zdefiniowana i nie zależy od i.

Definicja 4.8 (Asymptotyczny model regresji liniowej)**.** Model $(\boldsymbol{y}, \boldsymbol{x})$ spełniający zestaw założeń (B.1)–(B.4) nazywamy **asymptotycznym modelem regresji liniowej**.[a] Jeżeli model spełnia dodatkowo założenie (B.5), to mówimy o **klasycznym** asymptotycznym modelu regresji liniowej.

[a] Czasami, z drobną kolizją oznaczeń, będziemy mówić o asymptotycznym modelu regresji liniowej, jeżeli będą spełnione tylko założenia (B.1)–(B.2).

Warto zwrócić uwagę, że założenia te mogą być uznane za słabsze od tych nałożonych na klasyczny model regresji liniowej. Po pierwsze, aby określić asymptotyczny rozkład estymatorów OLS nie jest wymagane założenie normalności błędów. Po drugie, warunkowe własności modelu klasycznego (np. ścisłą egzogeniczność czynnika losowego) zastąpiliśmy przez warunek (bezwarunkowej) ortogonalności, co dopuszcza np. modele autoregresyjne. Podobnie jak wcześniej, omówmy pokrótce każde z założeń:

- Założenie (B.1) jest bezpośrednim odpowiednikiem założenia (A.1) i określa typ zależności między zmienną objaśniającą, a zmiennymi objaśnianymi.

- Założenie (B.2) łączy stacjonarność i ergodyczność. W tym wykładzie główny nacisk będzie położony na najprostszy przypadek, kiedy założenie to jest spełnione, tzn. gdy wektor $((x_i, y_i))_{i\in\mathbb{N}}$ jest i.i.d. Warto zwrócić uwagę, że założenie (B.2) implikuje stacjonarność czynnika losowego $\boldsymbol{\epsilon}$. W szczególności bezwarunkowy drugi moment $\mathbb{E}[\epsilon_i^2]$ (jeżeli istnieje) jest stały i nie zależy od i. Nie implikuje to jednak warunkowej homoskedastyczności z (A.4), tzn. $\mathbb{E}[\epsilon_i^2|x_i]$ może zależeć od x_i.

- Założenie (B.3) jest słabsze od (A.3), gdyż odnosi się do bezwarunkowej ortogonalności. Często (w praktyce) założenie to jest przedstawiane we wzmocnionej formie. Warto zwrócić uwagę, że nie wymagamy, aby zachodziło $\mathbb{E}[x_j\epsilon_i] = 0$, dla $i \neq j$, co dopuszcza pewne modele szeregów czasowych. Jeżeli model uwzględnia wyraz wolny (np. gdy $x_{i1} = 1$, dla $i \in \mathbb{N}$) to założenie (B.3) implikuje zerową wartość czynnika losowego, tzn. własność $\mathbb{E}[\epsilon_i] = 0$, $i \in \mathbb{N}$.

- Założenie (B.4) można traktować jako odpowiednik założenia (A.2). Ponieważ macierz $\mathbb{E}[x_i x_i']$ jest skończona, z twierdzenia ergodycznego dostajemy od razu $\lim_{n\to\infty} S_{xx} = \Sigma_{xx}$ (p.n.), gdzie $S_{xx} := \frac{1}{n}\sum_{i=1}^n x_i x_i'$. Wynika stąd, że dla odpowiednio dużych $n \in \mathbb{N}$ macierzy informacji dla klasycznego modelu regresji liniowej (o n obserwacjach) musi być pełnego rzędu, co implikuje (A.2).

- Własność MDS w założeniu (B.5) można traktować jako wzmocnienie założenia (B.3). Procesy spełniające własność MDS są procesami o zerowej średniej, więc (B.5) faktycznie implikuje (B.3). Założenie to będzie potrzebne do pokazania asymptotycznej normalności estymatora OLS. Założenie to jest spełnione na przykład, gdy

$$\mathbb{E}[\epsilon_i|\epsilon_{i-1}, \ldots, \epsilon_1, x_i, \ldots, x_1] = 0, \quad i \in \mathbb{N}. \tag{4.1}$$

  Dowód tego, że warunek (4.1) jest wystarczający dla (B.5) pozostawiamy jako ćwiczenie do domu. Warto zwrócić uwagę, że oprócz poprzednich obserwacji (od momentu 1 do $i-1$) do warunkowej wartości oczekiwanej wchodzi bieżąca informacja, tzn. wektor x_i.

4.3 Asymptotyczne własności estymatora OLS

Pokażemy, że przy założeniach (B.1)–(B.4) estymator OLS parametru β jest zgodny i asymptotycznie normalny. Traktując model spełniający założenia (B.1)–(B.4) jako ciąg modeli klasycznej

regresji liniowej, oraz przypominając (2.9), dla każdej ustalonej ilości obserwacji $n \in \mathbb{N}$ definiujemy odpowiednik estymatora OLS parametru β dany jako

$$\boldsymbol{b}_n := S_{xx}^{-1}(n)\, s_{xy}(n), \tag{4.2}$$

gdzie $S_{xx}(n) := \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ oraz $s_{xy}(n) := \frac{1}{n}\sum_{i=1}^{n} x_i y_i$. Warto zwrócić uwagę, że macierz $S_{xx}(n)$ jest (asymptotycznie) prawie na pewno odwracalna, ale dla ustalonego $n \in \mathbb{N}$ mogą istnieć zdarzenia dla których macierz ta jest singularna, co skutkuję tym, iż estymator (4.2) nie jest wszędzie poprawnie zdefiniowany. Oznaczając taki zbiór przez A_n, dla każdego $\omega \in A_n$ dodefiniowujemy $\boldsymbol{b}_n(\omega) = 0$. Od teraz, będziemy korzystać z uproszczonej notacji statystycznej (z wykładu Statystyką) i pomijać n, tzn. pisać $\boldsymbol{b}$, S_{xx} oraz s_{xy} zamiast $\boldsymbol{b}_n$, $S_{xx}(n)$ oraz $s_{xy}(n)$.

Propozycja 4.9 (Zgodność estymatora $\boldsymbol{b}$)**.** Załóżmy, że mamy dany asymptotyczny model regresji liniowej spełniający założenia (B.1)–(B.4). Wtedy $\boldsymbol{b}$, jest asymptotycznie (słabo) zgodny, tzn. $\boldsymbol{b} \to \beta$ (według prawdopodobieństwa), gdy $n \to \infty$.

Dowód. Załóżmy, że model spełnia (B.1)–(B.4). Dla $\bar{\boldsymbol{g}} := \frac{1}{n}\sum_{i=1}^{n} g_i$, dostajemy[14]

$$\begin{aligned}
\boldsymbol{b} - \beta &= S_{xx}^{-1}\, s_{xy} - \beta \qquad (4.3)\\
&= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \beta \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i(\beta x_i' + \epsilon_i) - \beta \cdot \frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i \epsilon_i\right) \\
&= S_{xx}^{-1}\bar{\boldsymbol{g}}. \qquad (4.4)
\end{aligned}$$

Korzystając z (B.2) wiemy, że proces stochastyczny $(x_i x_i')_{i\in\mathbb{N}}$ jest ergodycznie stacjonarny skąd, na mocy Twierdzenia 4.5, dostajemy

$$S_{xx} \xrightarrow{p.n.} \Sigma_{xx} \quad (n \to \infty). \tag{4.5}$$

Zbieżność (4.5) implikuje z kolei zbieżność według prawdopodobieństwa, tzn. $S_{xx} \xrightarrow{p} \Sigma_{xx}$ dla $n \to \infty$. Z założenia (B.4) wiemy, że Σ_{xx} jest odwracalna, co daje nam

$$S_{xx}^{-1} \xrightarrow{p} \Sigma_{xx}^{-1} \quad (n \to \infty). \tag{4.6}$$

Dowód (4.6) pozostawiamy jako ćwiczenie do domu; patrz Lemat 2.3 w [Hay00]. Analogicznie, korzystając dodatkowo z założenia (B.3), dostajemy $\bar{\boldsymbol{g}} \xrightarrow{p} \mathbb{E}[g_i] = 0$, co wraz z (4.6) implikuje

$$S_{xx}^{-1}\bar{\boldsymbol{g}} \xrightarrow{p} 0 \quad (n \to \infty). \tag{4.7}$$

Przypominając (4.4) dostajemy więc $(\boldsymbol{b} - \beta) \xrightarrow{p} 0$, co implikuje $\boldsymbol{b} \xrightarrow{p} \beta$ i kończy dowód.

□

[14] dla każdego $\omega \in \Omega$ i (odpowiednio dużych) $n \in \mathbb{N}$ dla których $\omega \notin A_n$.

Propozycja 4.10 (Asymptotyczny rozkład estymatora $\boldsymbol{b}$)**.** Załóżmy, że mamy dany klasyczny asymptotyczny model regresji liniowej spełniający założenia (B.1)–(B.5). Wtedy $\boldsymbol{b}$ ma asymptotyczny rozkład normalny, tzn.

$$\sqrt{n}(\boldsymbol{b}-\beta) \xrightarrow{d} N(0, \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}) \quad (n \to \infty),$$

gdzie $\Sigma_{xx} = \mathbb{E}[x_i x_i']$, $S = \mathbb{E}[g_i g_i']$ oraz $g_i = x_i \epsilon_i$, $i \in \mathbb{N}$.

Dowód. Załóżmy, że mamy dany klasyczny asymptotyczny model regresji liniowej spełniający założenia (B.1)–(B.5). Korzystając z (4.4) dostajemy

$$\sqrt{n}(\boldsymbol{b}-\beta) = S_{xx}^{-1}(\sqrt{n}\,\bar{\boldsymbol{g}}).$$

Korzystając z (A.5) oraz CTG dla MDS (tzn. Twierdzenia 4.7) dostajemy

$$\sqrt{n}\,\bar{\boldsymbol{g}} \xrightarrow{d} N(0, S) \quad (n \to \infty),$$

gdzie $S = \mathbb{E}[g_i g_i']$. Przypominając, że $S_{xx}^{-1} \xrightarrow{p} \Sigma_{xx}^{-1}$ oraz korzystając z Twierdzenia Slutsky'ego (zob. Lemat 2.4 w [Hay00]) dostajemy

$$S_{xx}^{-1}(\sqrt{n}\,\bar{\boldsymbol{g}}) \xrightarrow{d} N(0, \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}) \quad (n \to \infty),$$

co kończy dowód. □

Mając dane residua OLS możemy również pokazać, że estymator OLS bezwarunkowej wariancji błędu jest zgodny. Podobnie jak w klasycznym modelu, dla każdego ustalonego $n \in \mathbb{N}$ wprowadzamy oznaczenie $e_i(n) := y_i - x_i \boldsymbol{b}_n$ i dla każdego $n \in \mathbb{N}$ definiujemy

$$s_n^2 := \frac{1}{n-k} \sum_{i=1}^{n} e_i^2(n).$$

Podobnie jak wcześniej będziemy używać skrótowego zapisu s^2 oraz e_i.

Propozycja 4.11 (Zgodność estymatora s^2)**.** Załóżmy, że mamy dany klasyczny asymptotyczny model regresji liniowej spełniający założenia (B.1)–(B.4). Wtedy s^2 jest zgodnym estymatorem drugiego momentu czynnika losowego, tzn. zachodzi

$$s^2 \xrightarrow{p} \mathbb{E}[\epsilon_i^2] \quad (n \to \infty).$$

Dowód. Niech model spełnia (B.1)–(B.4). Zauważając

$$s^2 = \frac{n}{n-k}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right),$$

wystarczy pokazać, że $\frac{1}{n}\sum_{i=1}^{n} e_i^2 \xrightarrow{p} \mathbb{E}[\epsilon_i^2]$, dla $n \to \infty$. Łatwo zauważyć, że (dla każdego ustalonego $n \in \mathbb{N}$ oraz $i = 1, 2, \ldots, n$) dostajemy

$$e_i = y_i - x_i'\boldsymbol{b} = y_i - x_i'\beta - x_i'(\boldsymbol{b}-\beta) = \epsilon_i - x_i'(\boldsymbol{b}-\beta),$$

co daje nam

$$e_i^2 = \epsilon_i^2 - 2(\boldsymbol{b}-\beta)' x_i \epsilon_i + (\boldsymbol{b}-\beta)' x_i x_i' (\boldsymbol{b}-\beta). \tag{4.8}$$

Sumując po $i = 1, 2, \ldots, n$ (dla kadego $n \in \mathbb{N}$) dostajemy

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - 2(\boldsymbol{b}-\beta)'\bar{\boldsymbol{g}} + (\boldsymbol{b}-\beta)' S_{xx} (\boldsymbol{b}-\beta). \tag{4.9}$$

Łatwo pokazać (ćwiczenie do domu), że dwa ostatnie czynniki w (4.9) dążą (wg prawdopodobieństwa) do zera, co daje nam równość (wg prawdopodobieństwa) granic $\frac{1}{n}\sum_{i=1}^{n} e_i^2$ oraz $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$, i kończy dowód; warto zwrócić uwagę że (nieobserwowany) proces $(\epsilon_i^2)_{i\in\mathbb{N}}$ jest stacjonarny i ergodyczny, więc rozkład graniczny $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$ jest dobrze określony (o ile drugi moment istnieje, co milcząco założyliśmy w tezie). □

W Propozycji 4.10 pokazaliśmy, że estymator $\boldsymbol{b}$ ma (asymptotycznie) rozkład normalny o wariancji $\Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}$, gdzie $\Sigma_{xx} = \mathbb{E}[x_i x_i']$ oraz $S = \mathbb{E}[g_i g_i']$. Ponieważ proces $(g_i) = (x_i \epsilon_i)$ nie jest bezpośrednio obserwowany, należy wyestymować S. Naturalnym kandydatem wydaje się estymator

$$\widehat{S} = \frac{1}{n}\sum_{i=1}^{n} e_i x_i x_i'. \tag{4.10}$$

Warto zauważyć, że zachodzi $g_i g_i' = (x_i\epsilon_i)(x_i\epsilon_i)' = \epsilon_i^2 x_i x_i'$, a residua (e_i) przybliżają (nieznane) wartość czynnika losowego (ϵ_i). Aby estymator ten był zgodny, potrzebne jest dodatkowe założenia nałożone na wyższe momenty dla zmiennych objaśniających.

Propozycja 4.12 (Zgodność estymatora $\widehat{S}$)**.** Załóżmy, że mamy dany klasyczny asymptotyczny model regresji liniowej spełniający założenia (B.1)–(B.4) oraz macierz czwartych momentów dla zmiennych objaśniających istnieje i jest skończona, tzn. dla $i \in \mathbb{N}$ zachodzi

$$\mathbb{E}[(x_{ij}x_{im})^2] < \infty, \quad j, m = 1, 2, \ldots, k.$$

Wtedy estymator $\widehat{S}$ wartości S dany przez (4.10) jest (słabo) zgodny.

W wykładzie tym pominiemy (techniczny) dowód Propozycji 4.12. Aby zrozumieć, dlaczego założenie o wyższych momentach jest potrzebne zrobimy szkic dowodu dla jednej zmiennej objaśniającej. Przyjmując $k = 1$ i zauważając, że mamy do czynienia ze zmiennymi losowymi (a nie wektorami losowymi), z (4.8) dostajemy

$$e_i^2 = \epsilon_i^2 - 2(\boldsymbol{b}-\beta) x_i \epsilon_i + (\boldsymbol{b}-\beta)^2 x_i^2. \tag{4.11}$$

Mnożąc obie strony (4.11) przez x_i^2 i biorąc średnią z pierwszych n obserwacji dostajemy

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i^2 - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 x_i^2 = -2(\boldsymbol{b}-\beta)\frac{1}{n}\sum_{i=1}^{n} \epsilon_i x_i^3 + (\boldsymbol{b}-\beta)^2 \frac{1}{n}\sum_{i=1}^{n} x_i^4. \tag{4.12}$$

Jeżeli więc czwarty moment $(\mathbb{E}[x_i^4])$ istnieje i jest skończony, to wyrażenie $\frac{1}{n}\sum_{i=1}^{n} x_i^4$ dąży według prawdopodobieństwa do pewnej ustalonej liczby. Łącząc to z faktem, że $(\boldsymbol{b}-\beta)$ dąży do zera, wiemy, iż drugi człon prawej strony (4.12) dąży do zera; warto tutaj przypomnieć, że proces (x_i) jest ergodycznie stacjonarny, a $\boldsymbol{b}$ jest zgodny. Podobne rozumowanie można zastosować do $\sum_{i=1}^{n} \epsilon_i x_i^3$.

Mając dany estymator S możemy też łatwo zdefiniować estymator asymptotycznej wariancji dla $\boldsymbol{b}$. Przy założeniach wymienionych w Propozycji 4.12 zgodnym estymatorem $\Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}$ jest

$$S_{xx}^{-1} \widehat{S} S_{xx}^{-1},$$

gdzie $S_{xx} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i'$; dowód tego faktu pozostawiamy jako ćwiczenie do domu.

4.4 Kilka dodatkowych uwag dotyczących asymptotycznego modelu regresji liniowej

W rozdziale tym omówimy (bardzo ogólnie) kilka problemów związanych z modelem asymptotycznego modelu regresji liniowej. Więcej szczegółów można znaleźć w Rozdziale 2 w książce [Hay00].

Uwaga 4.13 (Testowanie hipotez statystycznych)**.** Bazując na Propozycji 4.10 możemy testować hipotezy statystyczne podobne do tych przedstawionych w Rozdziale 2.8. Asymptotyczny odpowiednik t-statystyki dla j-tej zmiennej objaśniającej dany jest przez

$$\tilde{t}_j := \frac{\boldsymbol{b}_j - \overline{\beta}_j}{\mathrm{SE}^*(\boldsymbol{b}_j)},$$

gdzie $\mathrm{SE}^*(\boldsymbol{b}_j) := \sqrt{\frac{1}{n} \cdot S_{xx}^{-1} \widehat{S} S_{xx}^{-1}}$ jest zgodnym estymatorem błędu $\boldsymbol{b}$. Asymptotycznym rozkładem statystyki $\tilde{t}_j$ jest standardowy rozkład normalny $N(0,1)$, na którym opiera się konstrukcja przedziałów ufności. Warto zaznaczyć, że asymptotyczny test nie wymaga założenia bezwarunkowej homoskedastyczności, co powoduje, iż w niektórych przypadkach jest on bardziej odporny.

Uwaga 4.14 (Bezwarunkowa homoskedastyczność)**.** Załóżmy, iż asymptotycznego modelu regresji liniowej spełnia dodatkowe założenie o warunkowej homoskedastyczności, tj. warunek

$$\mathbb{E}[\epsilon_i^2 | x_i] = \sigma^2,$$

dla pewnego (nieznanego) parametru $\sigma^2 > 0$. Wtedy możemy zdefiniować alternatywny estymator parametru S dany przez $\widehat{S}' := s^2 S_{xx}$, gdzie s^2 jest estymatorem OLS parametru σ^2. Estymator ten, w przypadku założenia o warunkowej homoskedastyczności, będzie zgodny. Zastępując estymator $\widehat{S}$ przez $\widehat{S}'$ we wzorze na odporny błąd standardowy SE^* dostajemy wartość $\tilde{t}_j = t_j$, gdzie t_j to t-statystyka zdefiniowana dla klasycznego modelu regresji liniowej. Podobna własność zachodzi dla F-statystyki i asymptotycznej statystyki.

Uwaga 4.15 (Testowanie założeń modelu)**.** W następnym rozdziale zajmiemy się testowaniem dopasowania klasycznego modelu regresji liniowej. Wiele z testów, które opiszemy można również zastosować do asymptotycznego modelu. Więcej szczegółów na ten temat można znaleźć w Rozdziale 2 w [Hay00]. W szczególności jeden z najbardziej popularnych testów warunku bezwarunkowej homoskedastyczności, Test White'a, opiera się na badaniu różnicy między estymatorem $\widehat{S}$, a $\widehat{S}'$.

5 Wybrane problemy klasycznego modelu regresji liniowej

W rozdziale tym zajmiemy się kilkoma wybranymi (bardziej praktycznymi) problemami związanymi z (klasycznym) modelem regresji liniowej. W szczególności omówimy pokrótce, jak weryfikować, czy zbiór danych, którymi dysponujemy, można uznać za spełniający założenia modelu, jak wybrać odpowiedni model, czy jak ocenić jego jakość. Ta część wykładu bazuje głównie na [Far15]. Warto tutaj zaznaczyć, iż niektóre testy – takie jak testowanie normalności, czy warunkowej homoskedastyczności – można znacząco osłabić przy modelu asymptotycznym. Dla uproszczenia nie będziemy jednak poruszać tego zagadnienia, skupiając się tylko na modelu klasycznym; więcej informacji na ten temat można znaleźć w Rozdziale 2 w [Hay00].

5.1 Weryfikacja założeń

Z praktycznego punktu widzenia weryfikacja czy dostępne dane spełniają założenia modelu jest kluczowa. Opiszmy kilka (wybranych) metod, które mogą temu służyć. W praktyce, wiele z popularnych metod to heurystyczne sposoby postępowania, które nie zawsze można powiązać ściśle z testem statystycznym. Przykładem może być tzw. *inspekcja wizualna* wybranych wykresów związanych z modelem, czy subiektywna interpretacja statystyk (np. współczynnika dopasowania). Zaproponowane tutaj metody należy traktować jako zbiór przykładów, a nie gotowe schematy postępowania.

5.1.1 Liniowa zależność (A.1)

Sprawdzenie, czy zależność między zmienną objaśnianą, a zmiennymi objaśniającymi jest liniowa ma kluczowe znaczenie dla modelu. Niespełnienie tego założenia powinno skutkować przekształceniem modelu i/lub zmianą zmiennych objaśniających. Wyniki analizy liniowości często uwzględniają sugestię konkretnych transformacji modelu (np. transformacji eksponencjalnej, jeżeli błędy mają charakter multiplikatywny, a nie addytywny). Istotna jest tutaj również identyfikacja obserwacji, które nie pasują do modelu. Zagadnieniami tymi zajmiemy się dokładniej w dalszej części wykładu, opisując w tym rozdziale tylko ogólny sposób postępowania.

Analiza wykresów dopasowanego modelu to najpopularniejsza metoda sprawdzania liniowości. Najczęściej obejmuje to analizę podstawowych wykresów związanych z oceną jakości dopasowania modelu:

- **Analiza wykresu: zaobserwowane wartości vs. dopasowane wartości**. W dobrze dopasowanym modelu punkty powinny znajdować się blisko diagonali (prostej $y = x$) z symetrycznym rozproszeniem wokół niej. Przy analizie wykresu należy zwrócić uwagę na tzw. wartości odstające (znajdujące się daleko od diagonali), które mogą istotnie zaburzać dopasowanie modelu. W przypadku misspecyfikacji modelu zazwyczaj obejmuje to najmniejsze/największe obserwacje, tzn. te dla których wartość zmiennej objaśnianej jest największa, bądź najmniejsza.

- **Analiza wykresu: residua vs. dopasowane wartości**. W dobrze dopasowanym modelu punktu powinny być symetrycznie rozproszone względem prostej $y = 0$. Nie powinno być widać żadnych trendów (analiza wykresu jest również związana z testowaniem homoskedastyczności błędu).

- **Analiza wykresu: residua vs. wybrane zmienne objaśniające**. Należy sprawdzić, czy nie widać żadnych trendów, bądź grupowania względem zmiennych objaśniających.

Bardziej formalnymi metodami są testy statystyczne badające poprawne zależność funkcyjną. Istnieje wiele metod badających, czy zadaną zależność między zmienną objaśnianą, a zmiennymi objaśniającymi jest poprawna. Oprócz testów badających bezpośrednio dopasowanie liniowe, można spotkać też testy, który porównują obecną specyfikację z różnymi alternatywami (np. poprzez dołączenie wyższych potęg zmiennych objaśniających). Przykładowe metody to:

- **Testy statystyczne badające ogólne dopasowanie.** Przykładami takich testów może być test Harvey'a–Collier'a (badający, czy tzw. *rekurencyjne residua* mają zerową średnią) czy test Rainbow (badający, czy dla średnich wartości zmiennej objaśniającej dopasowanie nie różni się od pełnego dopasowania). W wykładzie tym nie będziemy analizować dokładniej tego typu testów.

- **Testy statystyczne porównujące model z alternatywnymi specyfikacjami.** Przykładami takich testów mogą być testy porównujące wyjściowy model z modelem opartym na transformacji Boxa-Coxa (sprawdzające czy tzw. *przekształcenie potęgowe* nie polepsza jakości modelu), czy test RESET Ramseya (sprawdzający, czy uwzględnienie wyższych potęg dopasowanych wartości z wyjściowego modelu nie polepsza jakości prognozy). Transformację Boxa-Coxa i powiązany test statystyczny omówimy dokładniej w Rozdziale 5.3.

5.1.2 Brak współliniowości (A.2)

Brak ścisłej współliniowości można sprawdzić bardzo łatwo poprzez policzenie rzędu (wyznacznika) macierzy danych. Może jednak zdarzyć się sytuacja, w której zmienne objaśniające są od siebie liniowo zależne, ale szum obecny w danych powoduje, iż rząd macierzy jest pełny – takiej sytuacji chcielibyśmy również uniknąć, gdyż zależność (prawie) liniowa między zmiennymi najczęściej skutkuje brakiem odporności statystycznej modelu. Współczynniki modelu oraz powiązane błędy standardowe są wtedy zazwyczaj bardzo duże i czułe na dodawanie, czy usuwanie obserwacji. Otrzymane wartości ciężko jednoznacznie zinterpretować.

Bezpośrednia analiza zmiennych objaśniających jest często metodą badania współliniowości. Możemy to obejmować nastepujące metody:

- **Analiza graficzna macierzy danych.** Przykładowo, wykres zależności między parami zmiennych objaśniających może dać nam informację o zależności liniowej.

- **Analiza macierzy korelacji między zmiennymi objaśniającymi.** Wartości bliskie ± 1 mogą sugerować problem z modelem.

Oczywiście istnieje również szereg statystyk związanych, które dają informację o występowaniu współliniowości w modelu:

- **Analiza współczynników tolerancji oraz wartości VIF**. Wektor tolerancji T mierzy stopień zależności liniowej pomiędzy każdą zmienną objaśniającą, a wszystkimi pozostałymi zmiennymi objaśniającymi. Definiujemy go jako $T = (1 - R_1^2, \ldots, 1 - R_k^2)$, gdzie R_i^2 to współczynnik determinacji dla modelu, gdzie zmienną objaśnianą jest i-ta kolumna wyjściowej macierzy danych, a zmiennymi objaśniającymi macierz powstała po usunięciu i-tej kolumny z wyjściowej macierzy danych. Ogólna reguła mówi, iż $T < 0.1$ może sugerować problem z współliniowością, natomiast $T < 0.01$ go identyfikuje. Często rozważa się również powiązany wektor VIF (*ang. Variance Inflation Vector*) zdefiniowany jako $\text{VIF} := 1/T$.

- **Analiza wartości własnych macierzy $\boldsymbol{XX'}$.** W idealnym modelu wszystkie zmienne powinny być do siebie ortogonalne, co sugeruje jednostkowe wartości własne macierzy $\boldsymbol{XX'}$. Z

drugiej strony zerowe (bądź małe) wartości własne implikują występowanie współliniowości w modelu. Aby zbadać stopień liniowości między zmiennymi często bada się stosunek największej wartości własnej do najmniejszej. Ogólna reguła jest taka, iż jeżeli pierwiastek z tej liczby jest większy od 30, sugeruje to współliniowość.

Należy również wspomnieć, iż ilość zmiennych objaśniających powinna być (istotnie) mniejsza od ilości obserwacji, aby uniknąć zbytniego dopasowania w modelu.

5.1.3 Poprawna specyfikacja błędów (A.3)–(A.5)

Testowanie egzogeniczności, homoskedastyczności, niezależności liniowej, oraz normalności błędów jest kolejnym kluczowym etapem oceny modelu. W literaturze można znaleźć dziesiątki (jeżeli nie setki) testów poświęconych temu zagadnieniu. Skupimy się na kilku wybranych (kluczowych) aspektach i omówimy powiązane przykładowe metody. Warto zaznaczyć, iż czynnik losowy (ϵ) nie jest bezpośrednio obserwowany – większość testów opiera się na analizie residuów, co nie zawsze prowadzi to takich samych rezultatów. W szczególności wiemy, iż residua (oraz warunkowa wariancja czynnika losowego) zależą od estymatora OLS, czyli pośrednio od macierzy danych. Wpływ ten jest jednak zazwyczaj relatywnie mały, co pozwala na analizę residuów w zastępstwie prawdziwych błędów.

Na początku omówimy metody związane ze sprawdzaniem niezależności błędów (czynników losowych). W wielu modelach obserwacje są zbierane w różnych chwilach czasowych co może implikować zależność (liniową) błędów od czasu. Z założenia (A.3) oraz (A.4) wiemy, iż średnia wartość błędów oraz ich wariancja nie powinna zależeć od czasu. Przy dodatkowym założeniu (A.5) błędy powinny być od siebie niezależne. Analiza tej własności opiera się najczęściej na analizie wykresów czasowych oraz badaniu powiązanych funkcji autokowariancji. Należy też sprawdzić potencjalną zależność pierwszych dwóch momentów błędu od wartości zmiennych objaśniających Do najczęstszych metod sprawdzających **niezależność błędów** należy:

- **Analiza wykresu: residua vs. numer obserwacji (czas).** Należy sprawdzić, czy nie występuje tutaj żaden trend.
- **Analiza wykresu: residua vs. wartości wybranej zmiennej objaśniającej.** Należy sprawdzić, czy nie występuje zależność między residuami, a zmiennymi objaśniającymi. W klasycznym modelu regresji liniowej pierwsze dwa momenty nie powinny zależeć od wartości $\boldsymbol{X}$.
- **Analiza funkcji autokorelacji.** Funkcja autokorelacji mierzy autokorelację procesu stochastycznego. Dla dyskretnego stacjonarnego procesu stochastycznego $(Z_t)_{t\in\mathbb{N}}$ definiujemy $\mathrm{ACF} : \mathbb{N} \to [-1, 1]$ jako $\mathrm{ACF}(k) = \mathrm{Cor}(X_t, X_{t+k})$. Jeżeli błędy są niezależne funkcja ta powinna być stale równa zero (za wyjątkiem $k = 0$). Niezerowe wartości implikują zależność liniową między błędami (w czasie).
- **Testy statystyczne badające autokorelację.** Istnieje wiele testów badających, czy w danym szeregu czasowym istnieje autokorelacja między błędami. Przykładami takich testów mogą być testy: Durbina–Watsona (sprawdzający pewną specyficzną asymptotyczną zależność między kwadratami residuów), Ljunga–Boxa i Boxa-Pierce'a (badające, czy wartości funkcji autokorelacji z próbki są zgodne z przewidywaniami), czy Breuscha–Godfreya (badający postać współczynnika dopasowania dla alternatywnego modelu regresji uwzględniającego prognozy z wyjściowego modelu). Test Durbina–Watsona oraz Ljunga–Boxa omówimy dokładniej w Rozdziale 5.1.4.

Powyższe testy można też uzupełnić o analizę, gdy szereg błędów ϵ zastępujemy przez $|\epsilon|$, czy ϵ^2.

Kolejną istotną cechą modelu, którą należy sprawdzić, jest homoskedastyczność błędów (czynnika losowego). Drugi moment (wariancja) czynnika losowego powinna być stała i niezależna od $\boldsymbol{X}$. Istnieje wiele metod badających **homoskedastyczność czynnika losowego**:

- **Analiza wykresu: residua vs. dopasowane wartości.** Sprawdzamy, czy rozproszenie błędu nie zmienia się w zależności od dopasowanej wartości. Jak wspomnieliśmy, można tutaj również wykryć, czy rozkład residuów nie jest (nieliniową) funkcją dopasowanej wartości, co często ma miejsce w przypadku złego dopasowania modelu.
- **Analiza wykresu: residua vs. numer obserwacji.** W przypadku szeregu czasowego rozproszenie błędu może zależeć od czasu. Dla danych finansowych (np. stóp zwrotu z akcji) często występują zjawisko tzw. skupiania zmienności (*ang. volatility clustering*) np. w okresach kryzysu.
- **Analiza wykresu: residua vs. wartości wybranej zmiennej objaśniające**: należy sprawdzić, czy nie występuje zależność między residuami, a zmiennymi objaśniającymi. W klasycznym modelu regresji liniowej pierwsze dwa momenty nie powinny zależeć od wartości $\boldsymbol{X}$.
- **Testy statystyczne badające homoskedastyczność.** Istnieje wiele testów sprawdzających zarówno warunkową, jak i bezwarunkową homoskedastyczność. Przykładem jest test Breuscha-Pagana (sprawdzający, czy zmienne objaśniające mogą dokładniej prognozować kwadraty residuów wyjściowego modelu), White'a (sprawdzający, czy kwadraty zmiennych objaśniających oraz iloczyny par zmiennych mogą dokładniej prognozować kwadraty residuów wyjściowego modelu), Goldfielda-Quanta (badający wariancję reszt na różnych podzbiorach), czy Harrisona–McCabe'a (badający, czy częściowe wartości RSS są zgodne z oczekiwaniami). Przykładowe dwa z tych testów omówimy dokładniej w Rozdziale 5.1.4.

Ostatnim istotnym założeniem jest założenie (A.5) o normalności błędów. Oprócz standardowych metod graficznych w statystyce istnieje wiele dedykowanych testów, które służą sprawdzeniu, czy dana próbka pochodzi z rozkładu normalnego. Przykładowe (podstawowe) metody sprawdzające **normalność błędów** to:

- **Analiza graficzna rozkładu błędów.** Aby zbadać poprawność założenie o normalności, należy wykonać histogram oraz wykres kwantyl-kwantyl i sprawdzić, czy pokrywają się one z rozkładem normalnym.
- **Testy statystyczne badające normalność.** Uwzględnia to wykonanie podstawowych testów statystycznych badających normalność takich, jak test Kołmogorowa-Smirnova (badający supremum odległości między dystrybuantą empiryczną a prawdziwą), Shapiro-Wilka (sprawdzający czy empiryczny drugi moment jest porównywalny do oczekiwanej ważonej średniej podniesionej do kwadratu), Jarque'a-Bera (sprawdzający grubość ogona rozkładu w oparciu o empiryczny trzeci i czwarty moment), czy Andersona-Darlinga (badający średni ważony kwadrat odległości między dystrybuantą empiryczną, a prawdziwą).

5.1.4 Opis wybranych testów statystycznych powiązanych z analizą założeń

Aby zilustrować metody testowania statystycznego założeń w modelu regresji liniowej, omówimy niektóre testy, które zostały wspomniane przy okazji testowania założeń (A.1)–(A.5). W rozdziale tym skupimy się główne na intuicji – większość faktów (takich jak rozkłady statystyk testowych)

będzie podana bez dowodu; warto wspomnieć, iż opis testów związanych z poprawną specyfikacją modelu będzie można znaleźć w Rozdziale 5.4, gdzie omówimy ogólne metody diagnostyczne modelu i porównanie modeli między sobą. W tym rozdziale skupimy się na testach związanych z analizą błędów (residuów) modelu, przedstawiając po dwa przykłady testów sprawdzających niezależność, homoskedastyczność, oraz normalność błędów.

Zacznijmy od opisu przykładowych dwóch statystyk związanych z badaniem autokorelacji:

- **Test Durbina-Watsona**: jest to test statystyczny sprawdzający autokorelację (rzędu 1) błędów. Mając wektor residuów $(e_i)_{i=1}^n$ konstruujemy statystykę testową

$$\mathrm{DW} := \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

  Zakładając, iż korelacja rzędu 1 dla (stacjonarnego) procesu $(e_i)_{i=1}^n$ wynosi ρ, asymptotycznie dostajemy

$$\frac{\sum_{i=2}^n (e_i^2 - 2e_i e_{i-1} + e_{i-1}^2)}{\sum_{i=1}^b e_i^2} \to 2(1-\rho) \quad (n \to \infty).$$

  Przy braku autokorelacji, statystyka DW powinna więc być bliska wartości 2 (dla dużej ilości obserwacji). Wyprowadzenie rozkład statystyki DW opiera się na analizie sum rozkładów χ^2, co wykracza poza material tego wykładu.

- **Test Boxa-Piercea oraz Test Ljunga-Boxa**: są to testy statystyczny sprawdzające występowanie autokorelacji w próbce, dla określonej z górnej granicy opóźnienia $h \in \mathbb{N}$. Mając wektor residuów $(e_i)_{i=1}^n$ oznaczmy przez $\hat{\rho}_k$ standardowy estymator autokorelacji rzędu $k \in \{1, \ldots, h\}$. Statystyka Boxa-Piercea zdefiniowana jest wtedy jako

$$Q_{\mathrm{BP}} := n \sum_{k=1}^h \hat{\rho}_k^2. \tag{5.1}$$

  Można pokazać, iż przy braku autokorelacji asymptotyczny rozkład Q_{BP} to χ^2 o h stopniach swobody. Statystykę Ljunga-Boxa można traktować jako modyfikację statystyki (5.2) daną przez

$$Q_{\mathrm{LB}} := n(n+2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n-k}. \tag{5.2}$$

  Asymptotycznie statystyka ta również dąży do rozkładu χ^2 o h stopniach swobody, przy czym zbieżność ta jest (zwykle) istotnie szybsza.

Następnie przedstawimy opis dwóch przykładowych statystyk związanych z badaniem homoskedastyczności:

- **Test Breuscha-Pagana**: jest to test statystyczny sprawdzający homoskedastyczność (stałą wariancję) błędów oparty o tzw. metodę mnożników Lagrange'a (LM). Statystyka testowa powstaje w oparciu o przeprowadzenie dodatkowej regresji liniowej: po dopasowaniu wyjściowego modelu i otrzymaniu wektora residuów $(\boldsymbol{e}_i)_{i=1}^n$ tworzymy nowy model regresji liniowej w którym zmienną objaśnianą jest kwadrat otrzymanych residuów, a zmiennymi objaśniającymi wyjściowe zmienne modelu, tzn. tworzymy model $\boldsymbol{e}^2 = \tilde{\beta}\boldsymbol{X} + \tilde{\epsilon}$. Wtedy, statystyka testowa wynosi

$$LM_{\mathrm{BP}} = nR^2,$$

gdzie n to wielkość próby, a R^2 to współczynnik dopasowania dodatkowego modelu (zakładamy, iż macierz danych zawiera wyraz wolny). Zakładając brak homoskedastyczności w próbce statystyka LM_{BP} powinna mieć rozkład χ^2 o k stopniach swobody, gdzie k to ilość zmiennych objaśniających. Idea stojąca za testem Breuscha-Pagana jest dosyć prosta. Kwadrat residuów przybliża nam wariancję modelu (zakładając zerową wartość oczekiwaną czynnika losowego). Zmienne objaśniające w dodatkowym modelu badają, czy wariancja błędów zależy w jakimś stopniu od wyjściowych zmiennych objaśniających. Dla dobrego modelu, wartość R^2 powinna więc być mała.

- **Test White'a**: jest to najbardziej popularny test statystyczny sprawdzający homoskedastyczność (stałą wariancję) błędów. Konstrukcja statystyki jest podobna do konstrukcji statystyki w teście Breuscha-Pagana. Różnica polega na tym, iż przy dodatkowej regresji uwzględniamy również kwadraty zmiennych objaśniających. Innymi słowy, tworzymy nowy model regresji liniowej w którym zmienną objaśnianą jest kwadrat otrzymanych residuów, a zmiennymi objaśniającymi zmienne wyjściowe, ich kwadraty, oraz proste iloczyny drugiego rzędu, tzn. wektory zmiennych $\boldsymbol{x}_j$, $\boldsymbol{x}_j^2$ oraz $\boldsymbol{x}_j\boldsymbol{x}_l$ dla $j, l = 1, \ldots, k$, gdzie $j \neq l$. Podobnie jak wcześniej, statystyka testowa wynosi
$$LM_{\mathrm{W}} = nR^2,$$
gdzie n to wielkość próby, a R^2 to współczynnik dopasowania dodatkowego modelu. Zakładając brak homoskedastyczności w próbce statystyka W powinna mieć rozkład χ^2 o h stopniach swobody, gdzie h to ilość estymowanych współczynników w dodatkowym modelu.

Na koniec opiszemy dwa testy związane ze sprawdzeniem normalności błędów:

- **Test Jarque'a–Bera**: test ten sprawdza, czy empiryczna wartość miara grubości ogonów błędów oparta o trzeci oraz czwarty moment zgadza się z oczekiwaniami dla rozkładu normalnego. Statystyka testowa wynosi
$$\mathrm{JB} := \frac{n-k+1}{6}\left(\hat{S}^2 + \frac{1}{4}(\hat{C}-3)^2\right),$$
gdzie $\hat{S} := \hat{\mu}_3/\hat{\sigma}^3$ to estymator parametru skośności, a $\hat{C} := \hat{\mu}_4/\hat{\sigma}^4$ to estymator parametru kurtozy; przez $\hat{\mu}_k$ oznaczamy klasyczny estymator k-tego centralnego momentu, a przez $\hat{\sigma}$ estymator odchylenia standardowego dla wektora residuów e. Zakładając, iż próbka pochodzi z rozkładu normalnego (dla którego $S = 0$ oraz $C = 3$) statystyka JB ma asymptotyczny rozkład χ^2 o dwóch stopniach swobody.

- **Test Shapiro–Wilka**: jest to standardowy test wykorzystywany do testowania normalności danych. Statystyka testowa dana jest przez
$$W = \frac{\left(\sum_{i=1}^n a_i e_{(i)}\right)^2}{\sum_{i=1}^n (e_i - \overline{e})^2},$$
gdzie $e_{(j)}$ odpowiada j-tej statystyce pozycyjnej z próbki, a (a_i) to pewien (ustalony) ciąg liczb zależny od n, charakterystyczny dla rozkładu normalnego.

Uwaga 5.1 (Wyciąganie wniosków z testów statystycznych)**.** Należy wyraźnie zaznaczyć, iż pozytywny wynik testów (duże p-value) nie prowadzi do automatycznej akceptacji modelu. Brak podstaw

do odrzucenia hipotezy alternatywnej wobec hipotezy zerowej nie powinien skutkować przyjęciem hipotezy zerowej (szczególnie jeżeli nie znamy mocy testu). Zazwyczaj testy sprawdzają jedną konkretną własność modelu (np. gruby ogon rozkładu reszt w przypadku testu normalności Jarque-Bera) i sprowadzają ją do analizy wartości wybranej statystyki testowej (jednej liczby). Wszechstronna analiza modelu jest zazwyczaj dużo ważniejsza, niż automatyczne przeprowadzenie szeregu testów statystycznych. Co więcej, negatywny wynik testu również nie musi prowadzić do automatycznego odrzucenia modelu; więcej na temat pułapek związanych z testowaniem statystycznym można znaleźć np. w [Nuz14].

5.2 Identyfikacja nietypowych obserwacji

Z praktycznego punktu widzenia przy dopasowaniu modelu regresji linowej mamy do czynienia z dwoma podstawowymi problemami: pojedynczymi obserwacjami, które nie pasują do modelu bądź istotnie wpływają na dopasowanie modelu. W tym rozdziale pokażemy, jak dokonać analizy modelu pod kątem tego typu danych. Ponieważ metoda najmniejszych kwadratów opiera się na pomiarze kwadratu odległości od dopasowania, niewielka ilość obserwacji odstających od reszty może istotnie wpłynąć na dopasowanie modelu.

5.2.1 Dźwignia

Jedną z podstawowych miar czułości modelu na daną obserwacje jest tzw. dźwignia.

Definicja 5.2 (Dźwignia)**. Dźwignią** (ang. *leverage*) i-tej obserwacji (dla $i = 1, 2, \ldots, n$) nazywamy i-ty wyraz diagonali macierzy projekcji $\boldsymbol{P}$, tzn. wartość $h_i := \boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

Łatwo pokazać, iż dla wektora residuów $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ przy standardowej estymacji OLS zachodzi

$$\mathrm{Var}[e_i|\boldsymbol{X}] = (1 - h_i)\sigma^2, \quad i = 1, 2, \ldots, n. \tag{5.3}$$

Mówi nam to, iż im większa wartość h_i tym mniejsza tolerancja błędu między zmienną objaśnianą, a jej dopasowaniem. Obserwacje o dużej dźwigni mają więc istotny wpływ na model niejako *przyciągając* krzywą regresji. Mówiąc bardziej formalnie, można wykazać, iż

$$h_i = \frac{\partial \hat{y}_i}{\partial y_i},$$

co daje nam interpretacji dźwigni w terminach czułości dopasowania na zmianę wartości zmiennej objaśnianej. Łatwo wykazać, iż $0 \leq h_i \leq 1$ oraz $\sum_{i=1}^{n} h_i = k$, co sugeruje, iż średnia wartość h_i powinna wynosić k/n. Ogólna zasada mówi, iż wszystkie obserwacje dla których dźwignia jest większa niż $2k/n$ powinny zostać poddane analizie w celu weryfikacji poprawnego dopasowania modelu. Duże wartości dźwigni oczywiście nie muszą świadczyć o złym dopasowaniu modelu - dają nam one tylko informacje, które obserwacje mogą istotnie wpływać na model.

Z (5.3) widzimy, iż wariancja residuów (w przeciwieństwie do wariancji nieznanych błędów w klasycznym modelu regresji liniowej) zależy istotnie od wartości h_i. Często dokonuje się standaryzacji wektora residuów rozpatrując wektor $(r_i)_{i=1}^n$ dany przez

$$r_i := \frac{e_i}{\sqrt{s^2(1 - h_i)}}, \tag{5.4}$$

który powinien mieć jednostkową wariancję i niskie wartości dla funkcji autokorelacji. Duże wartości r_i mogą sugerować, iż dana obserwacja nie pasuje do modelu.

5.2.2 Obserwacje wpływowe

Analiza obserwacji wpływowych (ang. *influential observation*) opiera się na analizie dopasowania modelu przy usuniętych pojedynczych obserwacjach. Dla $i = 1, 2, \ldots, n$, niech $\boldsymbol{b}^{(i)}$ będzie estymatorem OLS parametru β powstałym po usunięciu i-tej obserwacji (wiersza) z macierzy danych, a $\hat{y}^{(i)}$ powiązanym wektorem dopasowań. Aby długości wektorów oraz indeksy dla prognoz $\hat{y}$ oraz $\hat{y}^{(i)}$ były zgodne dodefiniowujemy $\hat{y}_i^{(i)} := \hat{y}_i$. Podstawowe metody badania wpływu obserwacji to:

- **Badanie odległości między wektorami współczynników liniowych**, tzn. między wyjściowym estymatorem $\boldsymbol{b}$, a $\boldsymbol{b}^{(i)}$. Odległość ta wynosi

$$\boldsymbol{b}^{(i)} - \boldsymbol{b} = -\left(\frac{1}{1-h_i}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i, \tag{5.5}$$

  gdzie $\boldsymbol{x}_i$ to i-ty wiersz wyjściowej macierzy $\boldsymbol{X}$, e_i to wartość residua dla i-tej obserwacji, a h_i to dźwignia i-tej obserwacji. Z (5.5) widać, iż duże wartości residuów, bądź duża dźwignia może powodować odstępstwa między wartościami $\boldsymbol{b}^{(i)}$, a $\boldsymbol{b}$. Analiza $\boldsymbol{b}^{(i)} - \boldsymbol{b}$ dla wszystkich obserwacji może być jednak czasochłonna i trudna do oceny, gdyż do przeanalizowania jest $n \times k$ wartości.

- **Badania różnic, między prognozami modeli**, tzn. wartości $\hat{y}$ (bez i-tej obserwacji) oraz $\hat{y}^{(i)}$. Podstawową miarą wpływu jest tzw. statystyka Cook'a D_i dana przez

$$D_i := \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_j^{(i)}\right)^2}{k \cdot s^2}, \tag{5.6}$$

  gdzie s^2 to estymator wariancji błędu (2.22) wyjściowego modelu. W dużym skrócie chcemy oszacować odległość między dopasowaniami modelu po usunięciu i-tej obserwacji. Statystykę (5.5) można równoważnie przedstawić jako

$$D_i = \frac{e_i^2}{k \cdot s^2}\left(\frac{h_i}{(1-h_i)^2}\right) = \frac{r_i^2}{k}\left(\frac{h_i}{1-h_i}\right),$$

  co obrazuje nam wpływ zestandaryzowanych residuów oraz dźwigni na wartość D_i. Analiza, które wartości statystyki D_i (zazwyczaj podawane jako funkcja od h_i) prowadzą do zbiorów krytycznych wykracza poza materiał tego wykładu.

5.3 Podstawowe transformacje modelu

W podrozdziale 5.1.1 wspomnieliśmy, że analiza założenia liniowej zależności (A.1) często opiera się na porównaniu wyjściowego modelu z alternatywnymi specyfikacjami. Omówmy teraz przykłady podstawowych metod transformacji modelu, które obejmują zarówno transformacje zmiennej objaśniającej jak i zmiennych objaśnianych. Kwestia wyboru właściwego modelu z danej rodziny modeli będzie omówiona w kolejnym podrozdziale.

5.3.1 Transformacja zmiennej objaśniającej

Jak wspomnieliśmy w Uwadze 1.6 transformacja zmiennej objaśniającej ma istotny wpływ na to, jaki typ błędu chcemy uwzględnić w wyjściowym modelu i często powiązana jest też z transformacją zmiennych objaśnianych. Zasadniczy podział rozróżnia dwa typy błędów: addytywne (bezwzględne) oraz multiplikatywne (relatywne):

- **Błąd addytywny**, to błąd którego wielkość (wariancja) nie powinna zależeć od (wyjściowych) wartości zmiennych objaśniających. Odpowiada to podstawowemu modelowi regresji liniowej

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}. \tag{5.7}$$

- **Błąd multiplikatywny**, to błąd którego wielkość zależy (ściśle) od wielkości (wyjściowych) zmiennych objaśniających. Zakładając, że $\boldsymbol{y}, \boldsymbol{\epsilon} > 0$, odpowiada to modelowi

$$\boldsymbol{y} = e^{\boldsymbol{X}\beta} \cdot \boldsymbol{\epsilon}. \tag{5.8}$$

Transformacja logarytmiczna pozwala na sprowadzenie modelu (5.8) do postaci modelu regresji liniowej (5.7). Obkładając obie strony (5.8) przez logarytm i wprowadzając pomocnicze oznaczenia $\tilde{\boldsymbol{y}} := \log(\boldsymbol{y})$ oraz $\tilde{\boldsymbol{\epsilon}} := \log(\boldsymbol{\epsilon})$ dostajemy

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\beta + \tilde{\boldsymbol{\epsilon}}.$$

Mając wyjściowy wektor danych $(\boldsymbol{y}, \boldsymbol{X})$ nie wiemy, który z modeli będzie dawał lepsze dopasowanie. Często dokonuje się transformacji samej zmiennej $\boldsymbol{y}$ aby zbadać (wstępnie) dopasowanie.[15] Jedną z popularniejszych metod transformacji jest **transformacja Boxa-Coxa**. Zakładając nieujemność zmiennych objaśniających ($\boldsymbol{y} > 0$) wprowadzamy rodzinę funkcji $g_\lambda : \mathbb{R}^+ \to \mathbb{R}$, indeksowanych parametrem $\lambda \in \mathbb{R}$, daną przez

$$g_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log x, & \lambda = 0. \end{cases}$$

Dla każdej wartości $\lambda \in \mathbb{R}$ dokonujemy dopasowania do modelu

$$g_\lambda(\boldsymbol{y}) = \boldsymbol{X}\beta + \epsilon,$$

a następnie sprawdzamy, kiedy jest ono najlepsze; warto zwrócić uwagę, że obejmuje to zarówno transformacje wielomianowe, jak i transformację logarytmiczną. Najczęściej dopasowanie optymalnego parametru $\lambda \in \mathbb{R}$ opiera się na metodzie największej wiarygodności przy założeniu normalności. Z transformacją Boxa-Coxa związany jest też test statystyczny badający, czy dana specyfikacja (najczęściej $\lambda_0 = 1$) jest poprawna, tzn. czy zachodzi $\lambda = \lambda_0$, gdzie λ to prawdziwa wartość najlepszego dopasowania. Definiujemy statystykę $K := 2(L(\hat{\lambda}) - L(\lambda_0))$, gdzie L to funkcja wiarygodności zadana przez

$$L(\lambda) := -\frac{n}{2} \log \mathrm{RSS}_\lambda + (\lambda - 1) \sum_{i=1}^{n} \log y_i + \frac{n \log n}{2},$$

RSS_λ to suma kwadratów residuów przy transformacji $g_\lambda(\cdot)$, a $\hat{\lambda}$ oznacza najlepiej dopasowaną wartość λ (metodą ML). Przy założeniu normalności, rozkład K powinien w przybliżeniu odpowiadać rozkładowi χ^2 o jednym stopniu swobody.

5.3.2 Transformacje zmiennych objaśnianych

Strukturalna postać modelu zakłada liniową relację między zmienną objaśnianą, a wyjściową macierzą danych $\boldsymbol{X}$. Poprzez transformacje $\boldsymbol{X}$ możemy jednak uwzględnić nieliniowe typy zależności. Do najpopularniejszych metod należy:

[15] Jeżeli występują ujemne wartości w wektorze $\boldsymbol{y}$ to zawsze możemy przesunąć wszystkie obserwacje o jakąś ustaloną stałą.

- **Dołączenie wyższych potęg zmiennych objaśniających w modelu**. Mając dany n-wymiarowy wektor $\boldsymbol{x}_i$ (odpowiadający i-tej kolumnie macierzy $\boldsymbol{X}$), do macierzy informacji możemy dołączyć dodatkowe kolumny $\boldsymbol{x}_i^2$, $\boldsymbol{x}_i^3$, itd. Możemy rozważać też iloczyny (po współrzędnych) różnych kolumn. Rozważając współczynniki dopasowania, czy kryteria wyboru (np. AIC lub BIC) można rozstrzygnąć, czy przekształcenie modelu prowadzi do jego polepszenia. Pomocne są tutaj również standardowe wykresy badające jakość dopasowania. Zostanie to omówione dokładniej w podrozdziale 5.4.

- **Stworzenie zmiennych kategorycznych w oparciu o wyjściowe zmienne**. Zamiast rozważać bezpośrednio zmienne objaśniające (np. o rozkładzie ciągłym) możemy przyporządkować je do kilku zbiorów, a następnie użyć zmiennych charakterystycznych tych zbiorów w regresji. Jest to szczególnie pomocne przy dużych zbiorach danych. Przykładem może być wartość funkcji identyfikującej, czy dana zmienna jest nieujemna.

- **Segmentacja zmiennej** (ang. *piecewise regression*). Dokonujemy segmentacji zmiennej w oparciu o jej wartości i następnie dołączamy każdy segment osobno do modelu. Opiszmy, jak procedura ta może wyglądać w przypadku podziału zmiennej na dwa spójne i dopełniające się podzbiory. Mając dany wyjściowy zbiór obserwacji $\boldsymbol{x}_j$ oraz punkt segmentujący s tworzymy dwie nowe zmienne $\boldsymbol{x}^1$ oraz $\boldsymbol{x}^2$ zadane przez

$$x_i^1 = \begin{cases} x_{ij} - s, & x_{ij} > s \\ 0, & x_{ij} \leq s \end{cases} \qquad \text{oraz} \qquad x_i^2 = \begin{cases} 0, & x_{ij} > s \\ s - x_{ij}, & x_{ij} \leq s, \end{cases}$$

  dla $i = 1, 2, \ldots, n$, które następnie możemy dołączyć do modelu zamiast zmiennej $\boldsymbol{x}_j$.

- **Ortogonalizacja zmiennych, redukcja wymiaru**. W praktycznych zastosowaniach, gdy dysponujemy bardzo dużą ilością danych, dokonuje się ich redukcji i/lub ortogonalizacji. Przykładem takiej metody może być stworzenie nowej macierzy danych w oparciu o analizę składowych głównych (PCA - Principal Component Analysis).

Oczywiście istnieje wiele innych metod, które dokonują nieliniowych przekształceń wyjściowej macierzy $\boldsymbol{X}$. Zazwyczaj ich poprawność sprawdza się graficznie, oraz w oparciu o formalne kryteria selekcji modelu. Dla dużych zbiorów danych, efektywne algorytmy doboru zmiennych objaśniających (ang. *feature selection, variable selection*) są często kluczowym elementem dopasowania modelu. Należy przy tym pamiętać, że dokonując każdej transformacji, czy poszukując odpowiedniej funkcji, możemy zmniejszyć (pośrednio) ilość stopni swobody w modelu.

5.4 Wybór właściwego modelu

Mając dane wiele różnych specyfikacji modelu regresji liniowej możemy zadać sobie pytanie, która jest najlepsza. Istnieją algorytmy pozwalające na redukcję ilości zmiennych objaśniających, bądź dołączanie zmiennych, które wydają się najlepiej wzbogacać model. W tym rozdzialem omówimy kilka podstawowych metod związanych z tego typu problemami. Dwie podstawowe metody dobierające lepszy model to:

- **Porównanie dwóch określonych specyfikacji za pomocą testu statystycznego** (hipoteza zerowa zakłada jedną postać, a hipoteza alternatywna drugą). Podstawowe podział obejmuje testy związane z:

  - **Modelami zagnieżdżonymi** (ang. *nested models*). Obejmuje to dwa modele, w których jeden jest rozszerzeniem/zawężeniem drugiego. Testy tego typu, oparte o t i F statystyki były już omówione w Rozdziale 2.8.
  - **Modelami niezagnieżdżonymi** (ang. *non-nested models*). Mając daną ustaloną zmienną objaśniającą y możemy porównać dwa ogólne modele: $y = \boldsymbol{X}\beta + \epsilon$ oraz $y = \tilde{\boldsymbol{X}}\tilde{\beta} + \tilde{\epsilon}$. Przykładową metoda może być stworzenie *supermodelu*

$$y = (1 - \lambda)\boldsymbol{X}\beta + \lambda\tilde{\boldsymbol{X}}\tilde{\beta} + \bar{\epsilon},$$

    i przetestowanie hipotezy $\lambda = 1$ wobec alternatywy $\lambda \neq 1$. Problemem tutaj jest jednak fakt, iż estymacja parametru λ nie powinna być przeprowadzona niezależnie od estymacji $\tilde{\beta}$ oraz β. Dokładna analiza tego oraz innych podobnych testów wykracza poza materiał tego wykładu. Przykładowy dokładniejszy opis J-testu Davidsona-MacKinnona można znaleźć w [Gre18, Rozdział 5.6].

- **Metody dobierające model w oparciu o określone kryterium wyboru**. Aby wybrać najlepszy model często rozważa się tzw. kryterium wyboru, które określa nam za pomocą jednej liczby, jak dobrze (relatywnie, w odniesieniu do innych dopasowań), dopasowany jest nasz model. Podstawowymi kryteriami są:

  - **Skorygowany współczynnik** R^2 (ang. *Adjusted* R^2). Podstawową miarą dopasowania modelu, przedstawioną w Rozdziale 2.3 jest współczynnik determinacji R^2. Wiemy jednak, że dołączenie kolejnej zmiennej do modelu może tylko zwiekszyć wartość tego współczynnika, co ogranicza jego użycie. Aby temu zaradzić wprowadza się jego skorygowaną wartość

$$\tilde{R}^2 := 1 - \frac{n-1}{n-k}(1 - R^2),$$

    który wprowadza karę za zużycie zbyt dużej liczby zmiennych, które nie polepszają dopasowania (a obniżają liczby stopni swobody w modelu).
  - **Kryterium informacyjne Akaike** (ang. *Akaike Information Criterion*). Jest ono zadane przez

$$\mathrm{AIC} := n \cdot \ln\left(\frac{RSS}{n}\right) + 2k.$$

    Kryterium to ma na celu minimalizacje tzw. odległości Kullbacka-Leiblera między zadanym modelem, a prawdziwym modelem.
  - **Kryterium informacyjne Bayesa** (ang. *Bayesian Information Criterion*). Jest ono zadane przez

$$\mathrm{BIC} := n \cdot \ln\left(\frac{RSS}{n}\right) + k \ln n.$$

    Kryterium to jest powiązane z kryterium AIC, ale opiera się na innym rozkładzie początkowym.

  Więcej informacji na tematy tych kryteriów można znaleźć w [Far15, Rozdział 10.3] oraz [Gre18, Rozdział 5.8.1].

Powyższe metody umożliwiają stworzenia wielu procedur służących do konstrukcji optymalnego modelu, czy sprawdzenia jego poprawności. Omówmy kilka przykładów:

- **Testy statystyczne sprawdzające, czy wyjściowy model można rozszerzyć do lepszego modelu**. Metody te skupiają się na tym, czy wyjściowy model można rozszerzyć do innego lepszego modelu modyfikując macierz danych, co często jest związane ze sprawdzeniem poprawność założenia (A.1). Powiązane testy statystyczne często badają potencjalne alternatywy (np. dołączenie wyższych potęg zmiennych objaśniających), które mogą prowadzić do lepszego dopasowania – przykładowe metody (np. test RESET Ramsey'a) omówiliśmy pokrótce w Rozdziale 5.1.1.

- **Procedury pojedynczej eliminacji zmiennych** (ang *backward elimination*). Mając dany model z dużą ilością zmiennych objaśniających, eliminujemy pojedynczo zmienne w oparciu o zadany algorytm. Procedura to może być powiązane zarówno z testem statystycznym (np. eliminacja oparta o minimalną wartość p-value dla t-statystyk dla wszystkich współczynników), jak i kryterium doboru (np. eliminacja oparta o najmniejszą wartość AIC, o ile ta spada).

- **Procedury pojedynczego doboru zmiennych** (ang. *forward selection*). Startując z prostego modelu, dodajemy do niego kolejne zmienny w oparciu o zadaną procedurę.

W literaturze można znaleźć tez procedury hybrydowe, które łączą eliminację, z doborem zmiennej (ang. *stepwise regression*). Więcej informacji na temat procedur oraz ich implementacji w środowisku R można znaleźć w [Far15, Rozdział 10]. Warto zaznaczyć, że mając dwa modele, często nie da się stwierdzić, który jest lepszy (albo który będzie lepszy w przyszłości, przy napływie nowych danych). Istnieją wiele metod opartych na wnioskowaniu Bayesowskim, które pozwalają na połączeniu wielu modeli i określenie ich poprawności (tzn. prawdopodobieństwa, który jest lepszy). Więcej informacji na ten temat można znaleźć np. w [Gre18, Section 5.8.4].

Oprócz wyboru między modelami, dobrze jest na końcu przeprowadzić ogólną ocenę modelu. Często uwzględnia to wszystkie aspekty wymienione w tym rozdziale, m.in. weryfikacje założeń, czy analizę nietypowych obserwacji. Często dokonuje się również dalszej analizy modelu w oparciu o różne techniki walidacyjne, takie jak

- **Walidacja krzyżowa** (ang. *cross–validation*). Przykładowy algorytm polega na podziale obserwacje na dwa zbiory: na jednym zbiorze (uczącym) dopasowujemy parametry; na drugim zbiorze (testowym) sprawdzamy jakość uzyskanej prognozy, np. poprzez analizę wartości RMSE (ang. *root-mean-square error*), tzn.

$$\sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{m}},$$

  gdzie $i = 1, 2, \ldots, m$ oznacza indeks obserwacji w zbiorze testowym, a $\hat{y}_i$ oznacza prognozę modelu (uzyskaną na podstawie zbioru uczącego i i-go wiersza macierzy danych zbioru testowego. Następnie możemy zamienić ze sobą zbiory i powtórzyć algorytm.

- **Sprawdzenie odporności modelu**, np. poprzez resampling obserwacji i inne metody bootstrapowe.

- **Testowanie wsteczne** (ang. *backtesting*). Używane zazwyczaj do modeli uwzględniających czas – analiza zachowania modelu w przeszłości.

Niektóre z tych technik wymagają podziału zbioru obserwacji na dwa podzbiory - jeden służący do estymacji, a drugi do oceny jakości prognozy modelu. W przypadku niesymetrycznym, gdy zbiór testowy służy tylko do sprawdzania jakości prognozy, typowym podejściem jest podział w stosunku 80/20.

A Wybrane fakty ze statystyki oraz algebry liniowej

W rozdziale tym przypomnimy kilka wybranych podstawowych faktów związanych ze statystyką i algebrą liniową, które zostały użyte podczas dowodów. Zacznijmy od definicji podstawowych pojęć (własności), które są nam potrzebne w podczas kursu.

Definicja A.1 (Podstawowe własności macierzy)**.** Niech $\boldsymbol{A}$ oraz $\boldsymbol{B}$ będą kwadratowymi macierzami rzeczywistymi o wymiarze $n \times n$, gdzie $n \in \mathbb{N}$. Mówimy, że

- $\boldsymbol{A}$ jest **symetryczna** jeżeli $\boldsymbol{A}' = \boldsymbol{A}$.
- $\boldsymbol{A}$ jest **idempotentna** jeżeli $\boldsymbol{A}^2 = \boldsymbol{A}$.
- $\boldsymbol{A}$ jest **nieosobliwa (odwracalna)**, jeżeli istnieje $(n \times n)$ macierz $\boldsymbol{C}$ taka, że $\boldsymbol{AC} = \boldsymbol{CA} = \mathrm{I}_n$
- $\boldsymbol{A}$ jest **dodatnio półokreślona**, jeżeli $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$ dla dowolnego wektora $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$.
- $\boldsymbol{A}$ jest **dodatnio określona**, jeżeli $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ dla dowolnego wektora $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$.
- $\boldsymbol{A} \geq \boldsymbol{B}$, jeżeli macierz $\boldsymbol{A} - \boldsymbol{B}$ jest dodatnio półokreślona.

W dowodach przedstawionych w trakcie wykładu często korzystamy z własności macierzy $\boldsymbol{X}\boldsymbol{X}'$. Poniższa propozycją wprowadza podsumowanie wybranych własności.

Propozycja A.2. Niech $\boldsymbol{X}$ oraz $\boldsymbol{A}$ będą macierzami rzeczywistymi o wymiarach $n \times k$ oraz $n \times n$, gdzie $n > k$ oraz $n, k \in \mathbb{N}$. Wtedy

1) $\boldsymbol{X}'\boldsymbol{X}$ jest symetryczna.

2) $\boldsymbol{X}'\boldsymbol{X}$ jest dodatnio pół-określona. Gdy $\mathrm{rank}(X) = k$, to $\boldsymbol{X}'\boldsymbol{X}$ jest dodatnio określona.

3) $\mathrm{rank}(\boldsymbol{X}'\boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X})$.

Dowód.
1) Dla dowolnych macierzy $\boldsymbol{B}, \boldsymbol{C}$ zachodzi $(\boldsymbol{BC})' = \boldsymbol{C}'\boldsymbol{B}'$ oraz $(\boldsymbol{B}')' = \boldsymbol{B}$, co daje $(\boldsymbol{X}'\boldsymbol{X})' = \boldsymbol{X}'\boldsymbol{X}$.

2) Dla $\boldsymbol{x} \in \mathbb{R}^k \setminus \{0\}$ dostajemy $\boldsymbol{x}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{x} = (\boldsymbol{X}\boldsymbol{x})'(\boldsymbol{X}\boldsymbol{x}) = \langle \boldsymbol{X}\boldsymbol{x}, \boldsymbol{X}\boldsymbol{x} \rangle \geq 0$, gdzie $\langle \cdot, \cdot \rangle$ to standardowy iloczyn wewnętrzny (skalarny) na przestrzeni $\mathbb{R}^k$ dla którego zachodzi $\langle z, z \rangle = z'z = \sum_{i=1}^{k} z_i^2$, gdzie $z = (z_1, \ldots, z_k) \in \mathbb{R}^k$. Dostajemy więc dodatnią półokreśloność. Zauważając, że $\langle \boldsymbol{X}\boldsymbol{x}, \boldsymbol{X}\boldsymbol{x} \rangle = 0$ wtedy i tylko wtedy, gdy $\boldsymbol{X}\boldsymbol{x} = 0$, dostajemy drugą część tezy.

3) Pokażmy, że dla dowolnego $x \in \mathbb{R}^k \setminus \{0\}$ zachodzi $\boldsymbol{X}\boldsymbol{x} = 0$ wtedy i tylko wtedy, gdy $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{x} = 0$. Równość wymiaru jądra odwzorowania da nam równość rzędu macierzy. Załóżmy, $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{x} = 0$. Mnożąc obie strony przez $\boldsymbol{x}'$ dostajemy warunek $\langle \boldsymbol{X}\boldsymbol{x}, \boldsymbol{X}\boldsymbol{x} \rangle = 0$, który implikuje $\boldsymbol{X}\boldsymbol{x} = 0$ dla dowolnego $\boldsymbol{x} \in \mathbb{R}^k \setminus \{0\}$. Implikacja w drugą stronę jest oczywista. □

Aby poznać rozkłady statystyk testowych, dobrze jest też przypomnieć niektóre fakty ze statystyki, które wiążą rozkład normalny z rozkładem χ^2.

Propozycja A.3. Niech $\boldsymbol{X}$ oraz $\boldsymbol{A}$ będą macierzami rzeczywistymi o wymiarach $n \times k$ oraz $n \times n$, gdzie $n > k$ oraz $n, k \in \mathbb{N}$. Wtedy

1) Jeżeli A jest symetryczna, idempotentna oraz $X \sim N(0, \mathrm{I}_n)$ to $X'AX \sim \chi^2(\mathrm{rank}(A))$.

2) Jeżeli A jest idempotentna to $rank(A) = \mathrm{tr}(A)$.

3) Dla dowolnego m-wymiarowego wektora losowego $X \sim N(\mu, \Sigma)$, gdzie Σ jest niezdegenerowana, zachodzi $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(m)$

Dowód.
1) Ponieważ A jest symetryczna, można ją przedstawić w postaci $A = Q\Lambda Q'$, gdzie Λ jest macierzą diagonalną, natomiast Q jest macierzą ortogonalną (spełniającą $Q' = Q^{-1}$). Zdefiniujmy wektor losowy $V := Q'X$. Wektor V ma wielowymiarowy rozkład normalny (jako kombinacja afiniczna X) o średniej $\mathbb{E}[V] = Q'\mathbb{E}[X] = 0$ oraz macierzy wariancji $\mathrm{Var}[V] = Q'\,\mathrm{Var}[X]Q = Q'Q = \mathrm{I}_n$. Łatwo również zauważyć, że $X = (Q')^{-1}V = QV$, skąd dostajemy $X'AX = V'Q'AQV = V'\Lambda V$. Ponieważ A jest idempotentna, to wiemy, że wyrazy na przekątnej macierzy Λ są równe zero lub jeden. Łącząc to z własnością $V \sim N(0, \mathrm{I}_n)$ oraz zauważając, że ilość niezerowych wyrazów na przekątnej wyznacza rząd macierzy równy $\mathrm{rank}(A)$, dostajemy równość

$$X'AX = V'\Lambda V = \sum_{i=1}^{\mathrm{rank}(A)} V_i^2,$$

gdzie (V_i) jest ciągiem niezależnych zmiennych losowych o standardowym rozkładzie normalnym. To już implikuje $X'AX \sim \chi^2(\mathrm{rank}(A))$.

2) Ponieważ A jest idempotentna, można ją przedstawić w postaci $A = Q\Lambda Q^{-1}$, gdzie Λ jest macierzą diagonalną o wartościach zero oraz jeden, przy czym ilośc jedynek równa jest rzędowi macierzy A. Korzystając z faktu, że ślad macierzy jest niezmienniczy względem przesunięć cyklicznych dostajemy $\mathrm{tr}(A) = \mathrm{tr}(Q\Lambda Q^{-1}) = \mathrm{tr}(Q^{-1}Q\Lambda) = \mathrm{tr}(\Lambda) = \mathrm{rank}(A)$.

3) Ponieważ Σ jest symetryczna i dodatnio określona, można ją przedstawić w postaci $\Sigma = Q\Lambda Q'$, gdzie Λ jest macierzą diagonalną o dodatnich wyrazach na przekątnej, a Q jest macierzą ortogonalną (spełniającą $Q' = Q^{-1}$). Niech $V := H(X-\mu)$, gdzie $H := Q'\sqrt{\Lambda^{-1}}Q$. Korzystając z symetryczności H łatwo zauważyć, że

$$H'H = \left(Q'\sqrt{\Lambda^{-1}}Q\right)\left(Q'\sqrt{\Lambda^{-1}}Q\right) = Q'\Lambda^{-1}Q = \Sigma^{-1}$$

oraz V ma wielowymiarowy rozkład normalny o średniej $\mathbb{E}[V] = H\mathbb{E}[X-\mu] = 0$ i macierzy wariancji

$$\mathrm{Var}[V] = H\,\mathrm{Var}[V]H' = H\Sigma H' = \left(Q\sqrt{\Lambda^{-1}}Q'\right)Q\Lambda Q'\left(Q\sqrt{\Lambda^{-1}}Q'\right) = \mathrm{I}_m;$$

Stąd dostajemy

$$(X-\mu)'\Sigma^{-1}(X-\mu) = (X-\mu)'H'H(X-\mu) = V'V = \sum_{i=1}^{m} V_i^2,$$

gdzie (V_i) jest ciągiem niezależnych zmiennych losowych o standardowym rozkładzie normalnym, co implikuje $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(m)$ i kończy dowód. □

B Notacja

n	liczba obserwacji
k	liczba zmiennych objaśniających
i	(zazwyczaj) indeks oznaczający numer obserwacji (wiersz $\boldsymbol{X}$)
j	(zazwyczaj) indeks oznaczający numer zmiennej objaśniającej (kolumnę $\boldsymbol{X}$)
α	Współczynnik (wyraz) wolny regresji liniowej
β	$(1 \times k)$ wektor współczynników regresji liniowej
$\boldsymbol{y}$	$(n \times 1)$ wektor (losowy) zmiennej do regresji; zmienna objaśniana
$\boldsymbol{X}$	$(k \times n)$ macierz (losowa) danych regresji liniowej; zmienne objaśniające
$\boldsymbol{\epsilon}$	$(k \times 1)$ wektor (losowy) błędu w regresji liniowej; czynnik losowy (stochastyczny)
$\boldsymbol{b}$	estymator OLS parametru β
$\boldsymbol{e}$	residua $(y - \hat{y})$ dla estymatora $\boldsymbol{b}$, tzn. $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$
$\hat{\boldsymbol{y}}$	prognoza modelu regresji liniowej dla ustalonego estymatora $\hat{\beta}$, tzn. $\hat{\boldsymbol{y}} = \hat{\beta}\boldsymbol{X}$
s^2	estymator OLS parametru σ^2, tzn. $s^2 = \mathrm{RSS}\,/(n-k)$
RSS	suma kwadratów residuów, czasami używa się też oznaczeń SRR lub RSE.
RSE	błąd standardowego czynnika losowego, $\sqrt{s^2}$
R^2	współczynnik determinacji
$\overline{z}$	średnia z próbki z
S_{xx}	średnia z próbki $(x_i x_i')_{i=1}^n$
s_{xy}	średnia z próbki $(x_i y_i)_{i=1}^n$
Σ_{xx}	Macierz kowariancji dla stacjonarnego szeregu (x_i), tzn. $\mathbb{E}[x_i x_i']$
s	Macierzy kowariancji dla stacjonarnego szeregu $g_i = x_i \epsilon_i$, tzn. $\mathbb{E}[g_i g_i']$
$\boldsymbol{g}$	proces MDS dla modelu asymptotycznego, tzn. $(x_i \epsilon_i)$
h_i	dźwignia i-tej obserwacji
$\boldsymbol{P}$	macierz projekcji, $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
$\boldsymbol{M}$	anihilator, $\boldsymbol{M} = I_n - \boldsymbol{P}$
t_j	statystyka testowa t-studenta dla t-testu (dla j-go współczynnika)
F	statystyka Fishera-Snedecora dla F-testu
(r_i)	wektor zestandaryzowanych residuów dany przez $e_i/\sqrt{s^2(1-h_i)}$,

---

x'	transpozycja wektora x, czasami również x^T
I_n	n-wymiarowa macierz jednostkowa
$\mathrm{rank}(A)$	rząd macierzy A
$\mathrm{tr}(A)$	ślad macierzy A

Literatura

[Ame85] T. Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[Bil08] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[Far15] J. J. Faraway. *Linear models with R*. Chapman and Hall/CRC, 2015.

[Gre18] W. H. Greene. *Econometric analysis (8th Ed.)*. Pearson, 2018.

[Hay00] F. Hayashi. *Econometrics*. Princeton University Press, 2000.

[Nuz14] R. Nuzzo. Scientific method: statistical errors. *Nature News*, 506(7487):150, 2014.

[Rao73] C. R. Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.