Modele liniowe

Na podstawie rozdziałów 3.1 i 3.2 książki *The Elements of Statistical Learning* (Trevor Hastie, Robert Tibshirani, Jerome Friedman)

Mateusz Łyczko

Instytut Matematyczny UWr

24.03.2023

Wprowadzenie

Modele liniowe mogą być używane do predykcji następnych wartości na podstawie zebranych wcześniej danych.
Były one bardzo często stosowane jeszcze przed erą komputerów, ale w obecnych czasach również nierzadko się z nich korzysta mimo istnienia innych metod. Niektóre z powodów są następujące:

- Są one proste, zarówno w strukturze, jak i do analizy,
- Można na nich łatwo zrozumieć, jak dane wejściowe wpływają na dane wyjściowe,
- Czasami dzięki nim uzyskujemy bardziej wiarygodne wyniki niż używając modeli nieliniowych, w szczególności gdy zebranych danych jest mało lub gdy dane są rzadkie (dużo zer w stosunku do innych wartości).

Postać modelu liniowego

Dane są obserwacje $(X_{1,1}, \ldots, X_{1,p-1}, Y_1), \ldots, (X_{n,1}, \ldots, X_{n,p-1}, Y_n)$. Na ich podstawie możemy stworzyć model liniowy. Jego postać w formie wektorowej jest następująca:

$$Y_i = \beta_0 + X_{i,1}\beta_1 + \ldots + X_{i,p-1}\beta_{p-1} + \varepsilon_i = \beta_0 + \left[\sum_{j=1}^{p-1} X_{i,j}\beta_j\right] + \varepsilon_i,$$

gdzie $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. Zapis macierzowy wygląda tak:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma^2 \mathbb{I}_{n\times n}), \quad \text{gdzie}$$

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbb{X} = \begin{pmatrix} 1 & X_{1,1} & \ldots & X_{1,p-1} \\ 1 & X_{2,1} & \ldots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \ldots & X_{n,p-1} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Rola zmiennych w modelu liniowym

Są 4 rodzaje zmiennych w takim modelu:

- W macierzy $\mathbb{X}$ są zmienne objaśniające (zmienne wejścia). Są one deterministyczne (nielosowe) i znamy ich wartości.
- Wektor $\boldsymbol{\beta}$ to wektor parametrów modelu. Nie znamy go i trzeba go oszacować na podstawie danych.
- Wektor $\boldsymbol{\varepsilon}$ to wektor zmiennych losowych.
- W wektorze $\boldsymbol{Y}$ są zmienne objaśniane (zmienne wyjścia). Wartości $Y_1, \ldots, Y_n$ są znane.

Gdy dopasujemy model (tzn. oszacujemy współczynniki β_i), to możemy spróbować wyznaczyć predykcje następnych danych. W tym celu dobieramy nowy rząd w macierzy $\mathbb{X}$ (znane wartości) i za pomocą modelu liniowego szacujemy nową zmienną Y (bo jej nie znamy).

Typy zmiennych objaśniających

Każda z kolumn macierzy $\mathbb{X}$ może pochodzić z różnych źródeł.
Przykłady:

- Zmienne mierzą wielkość badanej cechy (zmienna ilościowa), np. $X_{i,1} \in \mathbb{R}$.
- Zmienne są transformacjami, np. $X_{i,2} = \log \tilde{X}_{i,2}$.
- Zmienne są kategoryczne, tzn. nie przyjmują wartości rzeczywistych, tylko mają stany. Wtedy należy zakodować w pewien sposób takie zmienne. Przykładowo, niech zmienna opisuje wykształcenie i są 3 stany: podstawowe, średnie i wyższe. Wtedy tworzymy nowe zmienne: $X_{i,3}$ to indykator że dana osoba ma wykształcenie średnie, a $X_{i,4}$ to indykator że dana osoba ma wykształcenie wyższe. Czasami postępuje się inaczej: tworzymy jedną zmienną i przypisujemy wartości, np. tutaj odpowiednio 0, 1, 2.
- Zmienne są interakcjami, czyli połączeniem poprzednich zmiennych za pomocą iloczynu, np. $X_{i,5} = X_{i,2} \cdot X_{i,3}$.

Związek liniowy

Przypomnijmy postać modelu liniowego:

$$Y_i = \beta_0 + X_{i,1}\beta_1 + \ldots + X_{i,p-1}\beta_{p-1} + \varepsilon_i.$$

Nałóżmy wartość oczekiwaną na obie strony. Wtedy

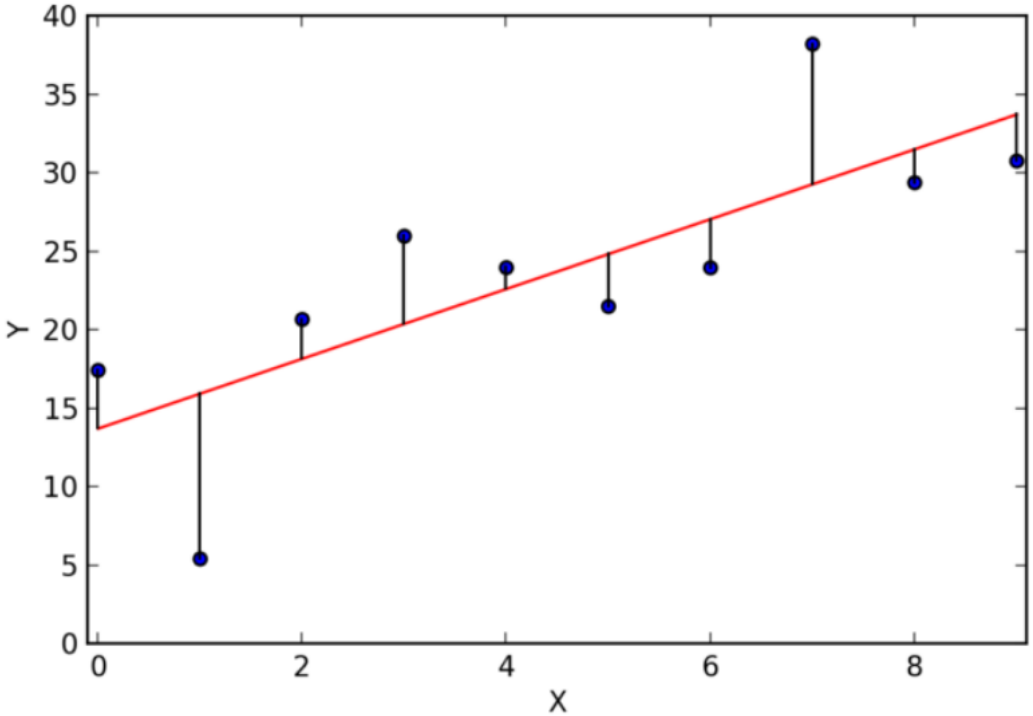
$$E[Y_i] = \beta_0 + X_{i,1}\beta_1 + \ldots + X_{i,p-1}\beta_{p-1}.$$

Widzimy, że model liniowy zakłada, iż funkcja regresji $E[Y]$ jest liniowa albo że model liniowy jest rozsądnym przybliżeniem pewnej relacji.

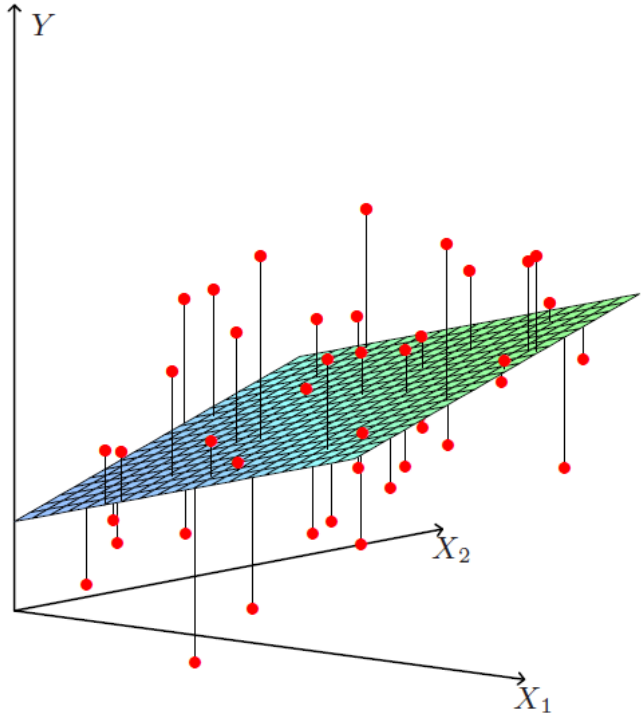
Uwaga

W ogólności funkcją regresji jest $E[Y|X]$. Założyliśmy tutaj, że X-sy są deterministyczne, a w takim przypadku $E[Y|X] = E[Y]$.

Wizualizacja funkcji regresji



(A)



(B)

(A): Dane to $(X_1, Y_1), \ldots, (X_n, Y_n)$, czerwony obiekt to prosta regresji $E(Y) = \beta_0 + \beta_1 X$.

(B): Dane to $(X_{1,1}, X_{1,2}, Y_1), \ldots, (X_{n,1}, X_{n,2}, Y_n)$, kolorowy obiekt to płaszczyzna regresji $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Estymacja parametrów

Najbardziej popularną metodą estymacji jest metoda najmniejszych kwadratów (Ordinary Least Squares). Zakłada ona, że chcemy zminimalizować sumę kwadratów residuów, czyli RSS (Residual Sum of Squares):

$$RSS = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \sum_{j=1}^{p-1} X_{i,j}\beta_j) \right]^2 .$$

Można pokazać, że minimalizacja tego wyrażenia prowadzi do uzyskania poniższego wektora.

Fakt

Estymatorem dla modelu liniowego uzyskanym za pomocą metody najmniejszych kwadratów jest

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}.$$

Wyznaczanie predykcji

Przypomnijmy, że $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Gdy mamy estymator $\widehat{\boldsymbol{\beta}}$, to możemy wyznaczyć predykcje. Mogą one być dla zbioru danych:

$$\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$$

lub dla nowych obserwacji ($\boldsymbol{X}$ to tutaj nowy wiersz macierzy):

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

Niech $\mathbb{H} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$. Wtedy

$$\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y} = \mathbb{H}\boldsymbol{Y}.$$

Macierz $\mathbb{H}$ nazywa jest macierzą rzutu, a wektor $\widehat{\boldsymbol{Y}}$ jest rzutem ortogonalnym wektora $\boldsymbol{Y}$ na podprzestrzeń rozpiętą przez kolumny macierzy $\mathbb{X}$.

Kiedy wektor parametrów jest jednoznaczny, a kiedy nie?

Wektor $\widehat{\boldsymbol{\beta}}$ jest wyznaczony jednoznacznie, gdy macierz $\mathbb{X}$ jest pełnego rzędu (ma rząd p). Wtedy macierz $\mathbb{X}^T\mathbb{X}$ jest odwracalna. Wektor parametrów jest niejednoznaczny, gdy $rank(\mathbb{X}) < p$. Może się tak stać, gdy:

- Dwie kolumny mają współczynnik korelacji równy 1 lub -1, np. $\mathbb{X}_1 = 4\mathbb{X}_2$ ($\mathbb{X}_i$ to tutaj i-ta kolumna macierzy $\mathbb{X}$).
- Liczba obserwacji n jest mniejsza niż liczba zmiennych objaśniających p.

W takich przypadkach dopasowane wartości $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$ wciąż są rzutem ortogonalnym $\boldsymbol{Y}$ na podprzestrzeń rozpiętą przez kolumny $\mathbb{X}$. Jednak niejednoznaczność nie jest tutaj pożądana. Aby temu zaradzić, możemy, odpowiednio:

- Usunąć jedną z dwóch kolumn, dla których korelacja co do modułu wynosi 1.
- Skorzystać np. z kryteriów informacyjnych (m.in. AIC, BIC) lub metod regularyzacji (m.in. Ridge, Lasso).

Rozkład wektora parametrów

Fakt

Wektor $\widehat{\boldsymbol{\beta}}$ ma następujący rozkład:

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}).$$

Oznacza to, że jest on nieobciążony, a wariancje poszczególnych składowych wektora są wyznaczone na odpowiednich miejscach przekątnej macierzy $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.
Wariancję błędów losowych σ^2 zazwyczaj estymuje się w ten sposób:

$$\widehat{\sigma}^2 = s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$

Wtedy jest on nieobciążonym estymatorem σ^2. Gdy w mianowniku jest n, to wtedy jest on obciążony.

Testowanie istotności jednego parametru

Niech j będzie ustalone. Przeprowadzamy następujący test:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0.$$

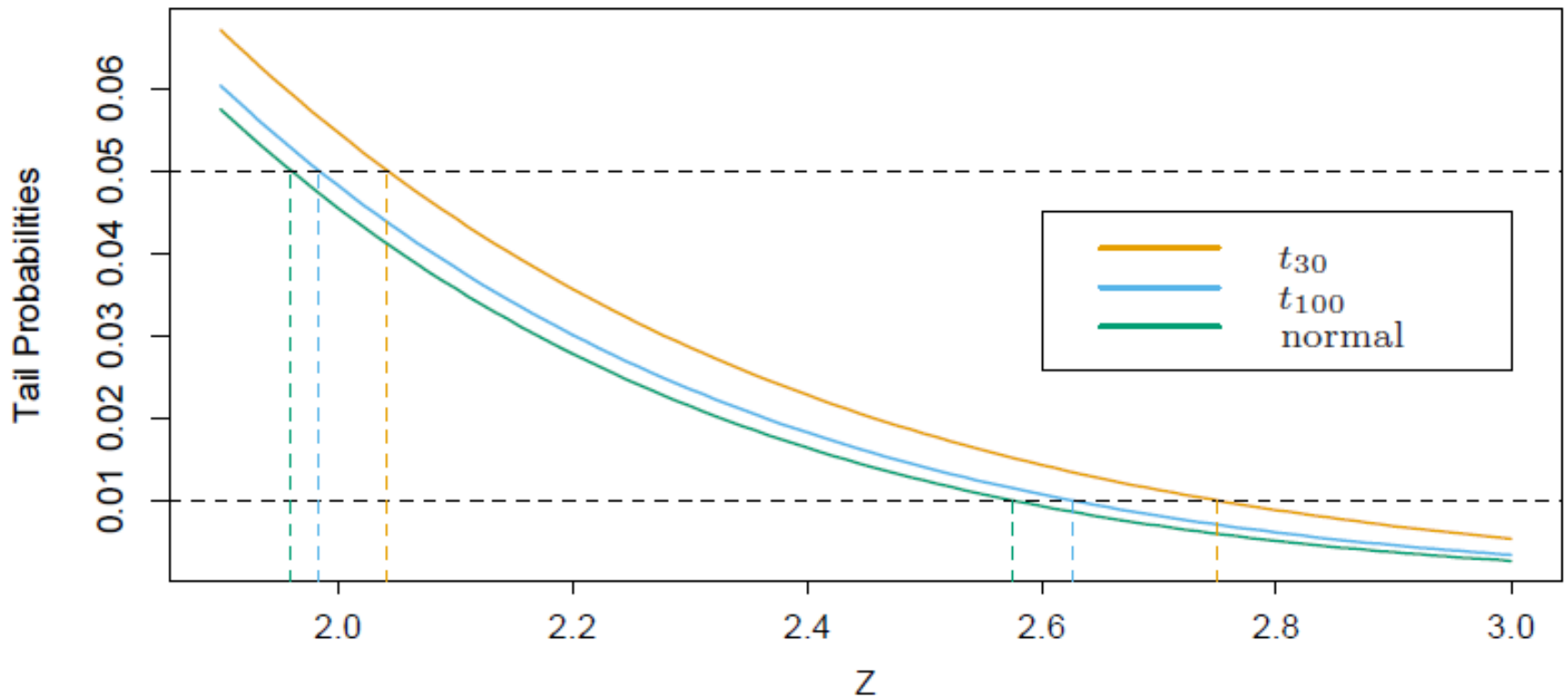
Statystyka testowa dla tego problemu jest następująca:

$$z_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{[(\mathbb{X}^T\mathbb{X})^{-1}]_{j+1,j+1}}}.$$

Przy hipotezie zerowej ma ona rozkład Studenta z $n - p$ stopniami swobody. Odrzucamy hipotezę zerową dla dużych (co do modułu) wartości z_j.
Gdy σ jest znana, to w statystyce testowej $\widehat{\sigma}$ zastępujemy przez σ. Przy hipotezie zerowej ma ona wtedy standardowy rozkład normalny.

Zastąpienie rozkładu Studenta rozkładem normalnym

Gdy nie znamy σ, a $n - p$ jest duże, to wtedy również można użyć rozkładu $N(0, 1)$, ponieważ różnice w kwantylach są w takich przypadkach zaniedbywalne. Pokazuje to poniższy rysunek:



Testowanie istotności wielu parametrów jednocześnie

Niech J będzie podzbiorem indeksów zmiennych objaśniających, tzn. $J \subseteq \{1, \ldots, p-1\}$. Przeprowadzamy teraz następujący test:

$$H_0 : \underset{j \in J}{\forall} \beta_j = 0 \quad vs \quad H_1 : \underset{j \in J}{\exists} \beta_j \neq 0.$$

Mamy więc 2 modele: w pierwszym nie ma zmiennych o indeksach z J, drugi jest pełny (choć pewne zmienne mogą, ale nie muszą być równe 0). Wtedy do testowania używamy następującej statystyki:

$$F = \frac{RSS_0 - RSS_1/(p_1 - p_0)}{RSS_1/(n - p_1)},$$

gdzie indeks dolny 0 oznacza, że dana liczba jest związana z hipotezą zerową, a 1 – z alternatywną ($RSS_1 = RSS$, $p_1 = p$).
Przy założeniu H_0 powyższa statystyka pochodzi z rozkładu $F(p_1 - p_0, n - p_1)$. Hipotezę zerową odrzucamy dla dużych wartości F.

Przykład: Rak prostaty

Dane pochodzą z badania przeprowadzonego w 1989 roku. Znajduje się tam 1 zmienna objaśniana i 8 zmiennych objaśniających.
Zmienna objaśniana: Logarytm ze swoistego antygenu gruczołu krokowego – log of prostate-specific antigen (`lpsa`)
Zmienne objaśniające:

- Logarytm z wielkości raka – log cancer volume (`lcavol`),
- Logarytm z wagi prostaty – log prostate weight (`lweight`),
- Wiek – age (`age`),
- 5 innych: `lbph`, `svi` (zmienna kategoryczna o 2 stanach), `lcp`, `gleason` (zmienna kategoryczna o 4 stanach), `pgg45`.

Tutaj zmienne kategoryczne nie są rozbijane na indykatory, lecz każdy stan ma przypisaną pewną liczbę.

Tabelka do przykładu

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

Wartość krytyczna dla testu $\beta_j = 0$ vs $\beta_j \neq 0$ gdy $\alpha = 0.05$: około 1.96.
Wartość statystyki F dla testu:
age=0 i lcp=0 i gleason=0 i pgg45=0 vs min. 1 z tych zmiennych $\neq 0$
to 1.67, p-wartość: 0.17.

Wykres do przykładu (scatterplot matrix)

Liniowość estymatora najmniejszych kwadratów

Estymator OLS jest liniowy. W przypadku modeli liniowych oznacza to, że $\widehat{\beta}_j = c_{1,j} Y_1 + \ldots + c_{n,j} Y_n$, gdzie $c_{i,j}$ nie zależy od β_j, ale może zależeć od X-sów. Liniowość estymatora nie ma więc nic wspólnego z liniowością modelu. Przykład: w prostej regresji liniowej

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{j=1}^n (X_j - \bar{X})^2} =$$

$$\sum_{i=1}^n \frac{X_i - \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2} Y_i = \sum_{i=1}^n c_i Y_i = c_1 Y_1 + \ldots + c_n Y_n,$$

gdzie $c_i = \frac{X_i - \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}$. Widzimy więc, że jest on liniowy.

Twierdzenie Gaussa-Markowa

Mówi ono, że estymator uzyskany za pomocą OLS dla $\boldsymbol{\beta}$ ma najmniejszą wariancję w klasie liniowych nieobciążonych estymatorów, czyli jest tam najlepszy. Jest on więc BLUE (Best Linear Unbiased Estimator). Przyjrzyjmy się temu twierdzeniu nieco bliżej, a dla ułatwienia rozważmy 1-wymiarowy wektor $\boldsymbol{a}^T\boldsymbol{\beta}$ zamiast p-wymiarowego wektora $\boldsymbol{\beta}$:

- Estymator $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ jest liniowy, ponieważ

$$\boldsymbol{a}^T\widehat{\boldsymbol{\beta}} = \boldsymbol{a}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y} := \boldsymbol{c}^T\boldsymbol{Y} = c_1Y_1 + \ldots + c_nY_n.$$

- Estymator $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ jest nieobciążony, gdyż

$$E[\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}] = E[\boldsymbol{a}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}] = \boldsymbol{a}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TE[\boldsymbol{Y}] =$$

$$\boldsymbol{a}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\boldsymbol{\beta} = \boldsymbol{a}^T\mathbb{I}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}.$$

- Zgodnie z 2 powyższymi punktami, estymator $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ jest w klasie liniowych nieobciążonych estymatorów. Z twierdzenia wynika, że jeśli weźmiemy inny liniowy estymator $\boldsymbol{d}^T\boldsymbol{Y}$ taki, że $E[\boldsymbol{d}^T\boldsymbol{Y}] = \boldsymbol{a}^T\boldsymbol{\beta}$, to

$$Var[\boldsymbol{a}^T\boldsymbol{\beta}] \leqslant Var[\boldsymbol{d}^T\boldsymbol{Y}].$$

Czy ograniczanie się do nieobciążoności to dobry pomysł?

Ponownie dla ułatwienia rozważmy $\boldsymbol{a}^T\boldsymbol{\beta}$. Widzieliśmy, że estymator $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ uzyskany za pomocą OLS jest nieobciążony. Nie zawsze jednak jest to najrozsądniejsze rozwiązanie. Często jakość estymatora $\widehat{\theta}$ parametru θ mierzy się za pomocą błędu średniokwadratowego (MSE – Mean Squared Error):

$$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + [E(\widehat{\theta}) - \theta]^2 \quad (\text{"}MSE = Variance + Bias^2\text{"}).$$

Tutaj za θ możemy przyjąć $\boldsymbol{a}^T\boldsymbol{\beta}$, a za $\widehat{\theta}$ możemy przyjąć $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$.
Mogą istnieć estymatory, które są lekko obciążone, ale za to mają dużą mniejszą wariancję niż estymatory nieobciążone. W konsekwencji MSE może być mniejsze. Przykładem estymatora obciążonego dla regresji liniowej jest estymator uzyskany za pomocą Ridge (regresji grzbietowej).

Ciekawostka

W 2022 roku Portnoy pokazał interesującą własność. Okazuje się, że w przypadku modeli liniowych nieobciążoność implikuje liniowość, czyli brak liniowości implikuje obciążoność. Innymi słowy, dla regresji liniowej zbiór nieliniowych nieobciążonych estymatorów jest pusty. Oznacza więc to, że tutaj w wyrażeniu BLUE (Best Linear Unbiased Estimator) słowo "linear" ("liniowy") jest zbędne. Można więc napisać, że estymator OLS jest BUE (Best Unbiased Estimator).

Bibliografia

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*
2. `https://www.toppr.com/guides/fundamentals-of-business-mathematics-and-statistics/correlation-and-regression/regression-lines/`
3. `https://statisticalhorizons.com/is-ols-blue-or-bue/`
4. `https://arxiv.org/pdf/2212.14185v1.pdf`