

1 Twierdzenie Gaussa-Markowa.

Estymatorami parametrów b_1 i b_2 równania regresji postaci

$$y = \beta_1 + \beta_2 x,$$

są

$$b_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad (1)$$

$$b_1 = \bar{y} - b_2 \bar{x}. \quad (2)$$

Wzory (1) i (2) podają postać estymatorów najmniejszych kwadratów parametrów b_1 i b_2 prostej regresji. Jak wiemy taka postać estymatorów nie jest jednak odpowiednio do badania teoretycznych właściwości estymatorów najmniejszych kwadratów. W tej części wykładu skorzystamy z równoważnej postaci estymatora b_2 , wygodniejszej do analizy, a mianowicie

$$b_2 = \sum_{i=1}^N w_i y_i,$$

gdzie

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^N (x_i - \bar{x})^2}.$$

Zauważmy, że

$$b_2 = \beta_2 + \sum w_i e_i,$$

gdzie e_i to błąd losowy w modelu regresji liniowej

$$y_i = \beta_1 + \beta_2 x_i + e_i.$$

Co możemy powiedzieć o estymatorach najmniejszych kwadratów b_1 i b_2 ? Estymatory (1) i (2), wyznaczone metodą najmniejszych kwadratów można, można użyć do oszacowania nieznanych parametrów β_1 i β_2 modelu regresji liniowej, bez względu na to jakie okazują się dane (nie musimy przyjmować założenia o rozkładzie normalnym błędów). Estymatory najmniejszych kwadratów są estymatorami liniowymi. Zarówno b_1 , jak i b_2 można zapisać jako średnie ważone wartości y_i . Jeśli założenia (1) – (8) podane w wykładzie są spełnione to estymatory najmniejszych kwadratów są nieobciążone. To oznacza, że

$$E(b_1) = \beta_1 \quad \text{oraz} \quad E(b_2) = \beta_2.$$

Znamy też postać wariancji estymatorów b_1 i b_2 oraz ich kowariancję. Ponadto, argumentowaliśmy, że sensownie jest wybrać do estymacji parametrów modelu nieobciążonych estymatorów o jak najmniejszej wariancji, ponieważ oznacza to, że mamy większą szansę na uzyskanie oszacowania zbliżonego do rzeczywistych wartości parametrów.

1.1 Słynne i użyteczne twierdzenie.

Teraz przedstawimy i omówimy słynne twierdzenie Gaussa-Markowa. Zanim je wysłowimy przypomnijmy podstawowe założenia modelu regresji liniowej.

Założenie 1. Liniowy model regresji. Model regresji ma **parametry liniowe**, czyli następującą postać

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Założenie 2. Zerowa wartość oczekiwana błędu. Pod warunkiem wartości X , wartość oczekiwana błędu u_i jest równa zero. Formalnie oznacza to, że

$$E(u_i|X_i) = 0$$

Założenie 3. Homoscedastyczność. Pod warunkiem wartości X , wariancja błędu u_i jest taka sama dla wszystkich obserwacji. Formalnie oznacza to, że

$$\begin{aligned} \text{Var}(u_i|X_i) &= E[u_i - E(u_i|X_i)]^2 \\ &= E(u_i^2|X_i) \\ &= \sigma^2. \end{aligned}$$

Założenie 4. Brak autokorelacji między błędami. Pod warunkiem wartości dwóch obserwacji zmiennej X , X_i i X_j , dla $i \neq j$, korelacja między dwoma błędami u_i i u_j wynosi zero. Symbolicznie,

$$\begin{aligned} \text{Cov}(u_i, u_j|X_i, X_j) &= E\{[u_i - E(u_i)]|X_i\} \{[u_j - E(u_j)]|X_j\} \\ &= E(u_i|X_i)(u_j|X_j) \\ &= 0. \end{aligned}$$

Założenie 5. Zerowa wartość kowariancji między u_i a X_i lub $E(u_i X_i) = 0$. Formalnie,

$$\begin{aligned} \text{Cov}(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)] &= E[u_i(X_i - E(X_i))], \quad \text{ponieważ } E(u_i) = 0 \\ &= E(u_i X_i) - E(X_i)E(u_i), \\ &= E(u_i X_i), \quad \text{ponieważ } E(u_i) = 0 \\ &= 0. \end{aligned}$$

Twierdzenie (Gaussa-Markowa).

*Przy założeniach (1)–(5) liniowego modelu regresji, estymatory b_1 i b_2 mają najmniejszą wariancję spośród wszystkich liniowych i nieobciążonych estymatorów β_1 i β_2 . Są to **najlepsze liniowe nieobciążone estymatory** b_1 i b_2 .*

Najpierw wyjaśnijmy, co mówi twierdzenie Gaussa-Markowa, a czego nie mówi.

1. Estymatory b_1 i b_2 są „najlepsze” w porównaniu z podobnymi estymatorami, które są liniowe i nieobciążone. Twierdzenie to nie mówi, że b_1 i b_2 są najlepsze spośród wszystkich możliwych estymatorów.
2. Estymatory b_1 i b_2 są najlepsze w swojej klasie, ponieważ mają najmniejszą zmienność. Porównując dwa estymatory liniowe i nieobciążone, zwykle chcemy wybrać ten, który ma mniejszą wariancję, ponieważ ta reguła szacowania daje nam wyższą wartość prawdopodobieństwa uzyskania oszacowania zbliżonego do prawdziwej wartości parametru.
3. Aby teza twierdzenia Gaussa-Markowa mogła być prawdziwa, założenia (1) – (5) muszą być spełnione. Jeśli którekolwiek z tych założeń nie jest spełnione, wówczas b_1 i b_2 nie są konieczne najlepszymi liniowymi nieobciążonymi estymatorami b_1 i b_2 .
4. Twierdzenie Gaussa-Markowa nie zależy od założenia normalności.
5. W prostym modelu regresji liniowej, jeśli chcemy zastosować liniowy i nieobciążony estymator mamy gotowe rozwiązanie w postaci estymatorów b_1 i b_2 . To wyjaśnia, dlaczego badamy te estymatory i dlaczego są tak szeroko stosowane w badaniach, nie tylko w dziedzinie ekonomii, ale także we wszystkich naukach społecznych, medycznych, przyrodniczych i technicznych.
6. Twierdzenie Gaussa-Markowa dotyczy estymatorów wyznaczonych metodą najmniejszych kwadratów.
7. W twierdzeniu Gaussa-Markowa rozważamy liniowe estymatory nieobciążone. Założenie liniowości jest bardzo pożądane z punktu widzenia obliczeniowego oraz określenia własności statystycznych estymatora. Własność nieobciążoności może być pożądana w przypadku gdy naszym celem jest wyznaczenie dokładnej postaci modelu. W tym przypadku estymator obciążony błędnie wyznacza rzeczywistą wartość parametrów modelu. Jednak gdy naszym celem jest prognoza na podstawie wyestymowanego modelu, może się okazać, że estymator obciążony ale o mniejszej wariancji niż estymator nieobciążony pozwala nam dokonać predykcji obciążonej mniejszym błędem. Twierdzenie Gaussa-Markowa podaje postać najlepszego liniowego estymatora nieobciążonego ale nie gwarantuje, że nie będziemy w stanie wyznaczyć estymatora obciążonego ale o mniejszej wariancji.

1.2 Dowód twierdzenia Gaussa-Markowa.

Udowodnimy tezę twierdzenia Gaussa-Markowa dla estymatora b_2 parametru β_2 , wyznaczonego metodą najmniejszych kwadratów. Nasz cel to wykazanie, że w klasie estymatorów liniowych i nieobciążonych estymator b_2 ma najmniejszą wariancję.

Niech

$$b_2^* = \sum k_i y_i,$$

gdzie k_i są stałymi, będzie innym estymatorem liniowym β_2 .

Do ułatwienia porównania z estymatorem najmniejszych kwadratów b_2 , przedstawmy k_i w postaci,

$$k_i = w_i + c_i,$$

gdzie

$$w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}.$$

Biorąc pod uwagę taką postać b_2^* i korzystając z postaci i właściwości w_i otrzymamy

$$\begin{aligned}
 b_2^* &= \sum k_i y_i = \sum (w_i + c_i) y_i = \sum (w_i + c_i) (\beta_1 + \beta_2 x_i + e_i) \\
 &= \sum (w_i + c_i) \beta_1 + \sum (w_i + c_i) \beta_2 x_i + \sum (w_i + c_i) e_i \\
 &= \beta_1 \sum w_i + \beta_1 \sum c_i + \beta_2 \sum w_i x_i + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \\
 &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i
 \end{aligned}$$

Teraz nakładając wartość oczekiwaną i korzystając z faktu, że zakładamy $E(e_i) = 0$ otrzymamy

$$\begin{aligned}
 E(b_2^*) &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) E(e_i) \\
 &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i
 \end{aligned}$$

Zauważmy, że aby estymator liniowy $b^* = \sum k_i y_i$, był nieobciążony, muszą być spełnione następujące równości

$$\sum c_i = 0 \quad \text{oraz} \quad \sum c_i x_i = 0. \quad (3)$$

Zakładamy więc, że warunki (3) są spełnione i użyjemy ich w celu uproszczenia wyrażenie na b_2^*

$$b_2^* = \sum k_i y_i = \beta_2 + \sum (w_i + c_i) e_i \quad (4)$$

Biorąc teraz pod uwagę własności wariancji oraz równość

$$\sum c_i w_i = \sum \left[\frac{c_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{\sum (x_i - \bar{x})^2} \sum c_i x_i - \frac{\bar{x}}{\sum (x_i - \bar{x})^2} \sum c_i = 0$$

możemy wyznaczyć wariancję liniowego estymatora nieobciążonego b_2^*

$$\begin{aligned}
 \text{var}(b_2^*) &= \text{var}[\beta_2 + \sum (w_i + c_i) e_i] = \sum (w_i + c_i)^2 \text{var}(e_i) \\
 &= \sigma^2 \sum (w_i + c_i)^2 = \sigma^2 \sum w_i^2 + \sigma^2 \sum c_i^2 \\
 &= \text{var}(b_2) + \sigma^2 \sum c_i^2 \\
 &\geq \text{var}(b_2)
 \end{aligned}$$

Ostatna nierówność wynika z oczywistego faktu, że $\sum c_i^2 \geq 0$ i pokazuje, że w rodzinie liniowych i nieobciążonych estymatorów nie ma estymatora parametru β_2 o wariancji mniejszej niż estymator najmniejszych kwadratów b_2 , każdy z alternatywnych estymatorów ma wariancję, która jest większa lub równa estymator najmniejszych kwadratów b_2 . Zauważy, że aby

$$\text{Var}(b_2^*) = \text{Var}(b_2)$$

musi być spełniona równość $\sigma^2 \sum c_i = 0$, a ponieważ $\sigma^2 > 0$, więc $\sum c_i^2 = 0$ co oznacza, że wszystkie współczynniki c_i muszą być równe zero, $c_i = 0$. W tym przypadku przypadku $b_2^* = b_2$. Zatem nie ma innego liniowego i nieobciążonego estymatora β_2 lepszego niż b_2 , co dowodzi twierdzenia Gaussa-Markowa.