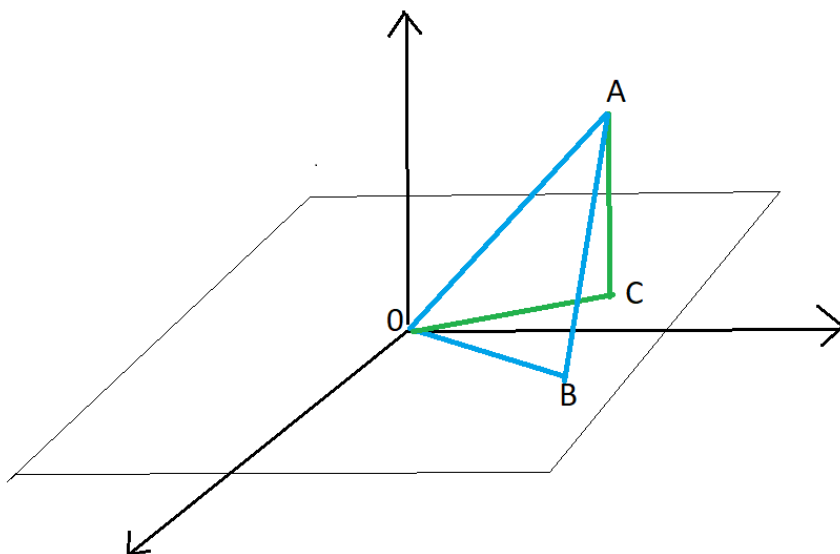


1 Interpretacja geometryczna metody NK.

Interpretacja geometryczna estymatora najmniejszych kwadratów dla modelu regresji liniowej pozwala lepiej zrozumieć jego charakter i własności. O wektorze obserwacji $\mathbf{y}^T = [y_1, y_2, \dots, y_n]$ możemy pomyśleć jako o wektorze łączącym początek układu z punktem A na rysunku 1, a o $y_1, y_2, \dots, y_n$ jako o współrzędnych w n wymiarowej przestrzeni prób. Przestrzeń prób na rysunku 1 jest trójwymiarowa.



Rysunek 1: Interpretacja geometryczna estymatora najmniejszych kwadratów. Na rysunku odcinek $[0, A] = \mathbf{y}$, $[0, B] = \mathbf{X}\boldsymbol{\beta}$, $[0, C] = \mathbf{X}\hat{\boldsymbol{\beta}}$ oraz $[C, A] = \mathbf{y} - \hat{\mathbf{y}}$.

Niech $\mathbf{X}$ będzie macierzą o p kolumnach będących wektorami rozmiaru n . Dla modelu regresji liniowej macierz $\mathbf{X}$ składa się z wektora $\mathbf{1}$ oraz wektorów $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$. Te p wektorów rozpinają p wymiarową podprzestrzeń, przestrzeń estymacji. Przestrzeń estymacji dla $p = 2$ jest pokazana na rysunku 1. Punkty tej przestrzeni estymacji są reprezentowane przez kombinacje liniowe wektorów $\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, a więc każdy punkt przestrzeni estymacji ma postać $\mathbf{X}\boldsymbol{\beta}$. Niech na rysunku 1 wektor $\mathbf{X}\boldsymbol{\beta}$ określa punkt B. Odległość średniokwadratowa od punktu B do punktu A jest określona jako

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Minimalizacja odległości punktu A, zdefiniowanego przez wektor obserwacji $\mathbf{y}$ do przestrzeni estymacji wymaga znalezienia punktu tej przestrzeni, który jest w najmniejszej odległości od punktu A. Tym punktem jest punkt $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ wyznaczony jako rzut ortogonalny wektora $\mathbf{y}$ na podprzestrzeń estymacji. Tak więc, ponieważ $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ jest prostopadły do przestrzeni estymacji mamy

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \quad \text{lub} \quad \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

co rozpoznajemy jako równość definiującą estymator najmniejszych kwadratów.

1.1 Macierz projekcji.

Kiedy estymujemy parametry modelu regresji liniowej, po prostu mapujemy zmienną zależną $\mathbf{y}$ na wektor dopasowanych wartości $\mathbf{X}\hat{\boldsymbol{\beta}}$ i wektor reszt $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Geometrycznie te odwzorowania są rzutami ortogonalnymi, które przenoszą każdy punkt przestrzeni $\mathbb{R}^n$ na punkt w podprzestrzeni $\mathbb{E} \subset \mathbb{R}^n$, pozostawiając wszystkie punkty w tej podprzestrzeni bez zmian. Z tego powodu, podprzestrzeń ta nazywana jest niezmienniczą podprzestrzenią projekcji. Rzut ortogonalny odwzorowuje dowolny punkt na najbliższy punkt podprzestrzeni. Jeśli punkt znajduje się już w niezmienniczej podprzestrzeni, jest odwzorowywany na siebie. Algebraicznie, rzut ortogonalny na podprzestrzeń może być przedstawiony jako przekształcenie wektora przez macierz projekcji. W przypadku metody najmniejszych kwadratów dwie macierze projekcji, które dają odpowiednio wektor dopasowanych wartości i wektor reszt mają postać

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T,$$

$$\mathbf{M}_{\mathbf{X}} = \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T,$$

gdzie $\mathbf{I}$ jest macierzą identycznościową $n \times n$.

Aby to zobaczyć, przypomnijmy sobie postać estymatora najmniejszych kwadratów parametru $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Widzimy, że

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{P}_{\mathbf{X}}(\mathbf{y}). \quad (1)$$

Dlatego pierwsza macierz projekcji $\mathbf{P}_{\mathbf{X}}$, odpowiada rzutowi na $\mathcal{S}(\mathbf{X})$, podprzestrzeń rozpiętą przez $\mathbf{X}$. Dla dowolnego wektora $\mathbf{y}$, $\mathbf{P}_{\mathbf{X}}(\mathbf{y})$ zawsze jest wektorem z podprzestrzeni $\mathcal{S}(\mathbf{X})$, ponieważ

$$\mathbf{P}_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}).$$

Ponieważ $\mathbf{P}_{\mathbf{X}}(\mathbf{y})$ ma postać $\mathbf{X}\mathbf{b}$ dla $\mathbf{b} = \hat{\boldsymbol{\beta}}$, jest to liniowa kombinacja kolumny $\mathbf{X}$, a zatem należy do $\mathcal{S}(\mathbf{X})$.

Z (1) jasno wynika, że zastosowanie rzutu $\mathbf{P}_{\mathbf{X}}$ do wektora $\mathbf{y}$ daje wyniku wektor dopasowanych wartości. Podobnie, gdy $\mathbf{M}_{\mathbf{X}}$, drugą z dwóch macierzy projekcji w (1), jest zastosowany do $\mathbf{y}$, daje wektor reszt:

$$\mathbf{M}_{\mathbf{X}}(\mathbf{y}) = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{u}}.$$

1.2 Zadania

Zadanie 1. Rozważmy dwa modele regresji

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

$$y = \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + u,$$

gdzie $z_1 = x_1 - 2x_2$, $z_2 = x_2 + 4x_3$, oraz $z_3 = 2x_1 - 3x_2 + 5x_3$. Niech $X = [x_1 \ x_2 \ x_3]$ i $Z = [z_1 \ z_2 \ z_3]$ będą macierzami o kolumnach odpowiednio x_1, x_2, x_3 i z_1, z_2, z_3 . Pokaż, że kolumny macierzy Z mogą być wyrażone jako liniowe kombinacje kolumn macierzy X , to znaczy, że dla pewnej macierzy A rozmiaru 3×3 , mamy równość $Z = XA$.

1. Znajdź elementy macierzy A .

2. Pokaż, że macierz A jest odwracalna, pokazując, że kolumny macierzy X są kombinacjami liniowymi kolumn Z .
3. Podaj postać macierzy A^{-1} .
4. Pokaż, że oba modele regresji dają w wyniku te same wartości dopasowanych parametrów i te same wartości reszt.
5. Podaj w jaki sposób oszacowanie najmniejszych kwadratów $\hat{\beta}_i$ jest powiązane z oszacowaniem najmniejszych kwadratów OLS $\hat{\alpha}_i$, dla $i = 1, 2, 3$.

Zadanie 2. Rozważ następującą regresję liniową:

$$y = X_1\beta_1 + X_2\beta_2 + u,$$

gdzie y to macierz $n \times 1$, natomiast X_1 to macierz $n \times k_1$, a X_2 to macierz $n \times k_2$. Niech $\hat{\beta}_1$ i $\hat{\beta}_2$ będą estymatorami najmniejszych kwadratów parametrów tego modelu. Teraz rozważmy następujące modele regresji:

1. $y = X_2\beta_2 + u$,
2. $P_1y = X_2\beta_2 + u$,
3. $P_1y = P_1X_2\beta_2 + u$,
4. $P_Xy = X_1\beta_1 + X_2\beta_2 + u$,
5. $P_Xy = X_2\beta_2 + u$,
6. $M_1y = X_2\beta_2 + u$,
7. $M_1y = M_1X_2\beta_2 + u$,
8. $M_1y = X_1\beta_1 + M_1X_2\beta_2 + u$,
9. $M_1y = M_1X_1\beta_1 + M_1X_2\beta_2 + u$,
10. $P_Xy = M_1X_2\beta_2 + u$,

Tutaj P_1 oznacza rzut ortogonalny na podprzestrzeń rozpiętą przez X_1 , a $M_1 = I - P_1$.

- (a) Dla których z powyższych modeli regresji estymatory β_2 będą takie same jak dla oryginalnego modelu?
- (b) Dla których modeli reszty będą takie same.

Zadanie 3. Rozważmy dwa modele regresji liniowej, jeden z ograniczeniami ($\gamma = 0$), a drugi bez ograniczeń:

$$\begin{aligned} y &= X\beta + u, \\ y &= X\beta + Z\gamma + u. \end{aligned}$$

Pokaż, że w przypadku regresorów wzajemnie ortogonalnych, gdy $X^T Z = O$, estymatory dla β w obu modelach są identyczne.

Zadanie 4. Rozważmy dwa modele regresji liniowej, jeden z ograniczeniami ($\gamma = 0$), a drugi bez ograniczeń:

$$\begin{aligned}y &= X\beta + u, \quad u \sim IID(0, \sigma^2 I) \\y &= X\beta + Z\gamma + u, \quad \sim IID(0, \sigma^2 I).\end{aligned}$$

Pokaż, że różnica między estymatorem $\tilde{\beta}$ w modelu bez ograniczeń oraz estymatorem $\hat{\beta}$ w modelu z ograniczeniami jest równa:

$$\tilde{\beta} - \hat{\beta} = (X^T M_Z X)^{-1} X^T M_Z M_X(y).$$

Wskazówka: Aby udowodnić tę równość, najłatwiej jest przemnożyć różnicę przez $X^T M_Z X$.

Zadanie 5. Niech b_1 i b_2 będą zmiennymi losowymi o skończonych wariancjach. Korzystając z faktu, że macierz kowariancji wektora $\mathbf{b} = [b_1 \ b_2]$ jest dodatnio określona udowodnij nierówność

$$[\text{Cov}(b_1 \ b_2)]^2 \leq \text{Var}(b_1)\text{Var}(b_2).$$

1.3 Zadania z wykorzystaniem komputera.

Zadanie 6. Wykorzystamy dane dotyczące konsumpcji i osobistych dochodów w Kanadzie w okresie 1947 do 1996 zamieszczone w pliku konsumpcja.data. Na podstawie tych danych wykonaj estymację parametrów modelu

$$c_t = \beta_1 + \beta_2 y_t + u_t, \quad u_t \sim NID(0, \sigma^2),$$

gdzie $c_t = \log C_t$ jest logarytmem konsumpcji, a $y_t = \log Y_t$ jest logarytmem dochodu. Następnie użyj estymatorów parametrów β_1 , β_2 oraz σ aby wysymulować 200 niezależnych obserwacji c_t . Na podstawie pierwszych 3 wysymulowanych danych wyestymuj β_1 , β_2 oraz σ i zachowaj w pliku ich wartość. Powtórz tę operację zwiększając liczbę wysymulowanych danych $n = 4, 5, \dots, 200$. Wykreśl wartość otrzymanych estymatorów jako funkcji od n . Zaobserwuj jak zmieniają się te wartości. Powtórz całe ćwiczenie z innym zestawem wysymulowanych danych. Które cechy wykresów oszacowań parametrów są wspólne dla obu eksperymentów, a które są różne?

Zadanie 6a. Narysuj wykres dystrybucyjny empirycznej rozkładu reszt estymacji metodą najmniejszych kwadratów wykonanej na podstawie jednego z zestawów danych symulowanych dla całego okresu próby, z zadania 6. Na tym samym wykresie narysuj dystrybucję rozkładu normalnego $N(0, \sigma^2)$, gdzie σ^2 oznacza teraz wariancję zastosowaną do symulacji logarytmu konsumpcji.

Pokaż, że rozkład empiryczny i rozkład normalny użyty do symulacji mają te same średnie, ale różne wariancje. Jak można zmodyfikować reszty tak aby rozkład empiryczny dla zmodyfikowanych reszt miał taką samą wartość wariancji jak wykorzystany do symulacji rozkład normalny.