Akademia Górniczo-Hutnicza
im. Stanisława Staszica
w Krakowie

Praca licencjacka

Testowanie regresji liniowej przeciwko wypukłej

Grzegorz Mika

Kierunek: Matematyka

Nr albumu: 267543

Promotor
dr Konrad Nosek

Wydział Matematyki Stosowanej

Kraków 2016

Oświadczenie autora

Ja, niżej podpisany Grzegorz Mika oświadczam, że praca ta została napisana samodzielnie i wykorzystywała (poza zdobytą na studiach wiedzą) jedynie wyniki prac zamieszczonych w spisie literatury.

. . . .
(Podpis autora)

Oświadczenie promotora

Oświadczam, że praca spełnia wymogi stawiane pracom licencjackim.

. . . .
(Podpis promotora)

Spis treści

Wstęp

Rozważmy pewien zestaw danych $\{(x_i, y_i)\}_{i=1,2,\ldots,n}$ i spróbujmy dopasować pewną funkcję f do danych według modelu

$$y_i = f(x_i) + \varepsilon_i$$

gdzie zakładamy, że błędy ε_i są niezależnymi zmiennymi losowym o tym samym rozkładzie normalnym.

Najprostszym związkiem między obserwcjami x_i a odpowiedziami y_i jest zależność liniowa, możliwy jest jednak również inny związek między obserwacjami a odpowiedziami, co prowadzi do sformułowania hipotezy

$$H_0\colon f(x) = ax + b \quad vs. \quad H_1\colon f \in \mathcal{F},$$

gdzie $\mathcal{F}$ jest klasą funkcji wypukłych.

W niniejszej pracy postaramy się skonstruować odpowiedni do postawionego problemu test statystyczny. Zaproponowane zostanie rozwiązanie oparte o iloraz wiarogodniści w przypadku modelu regresji z ograniczeniami w postaci nierówności.

W pierwszym rozdziale zostaną omówione podstawowe własności stożków wypukłych traktowanych jako po podzbiór przestrzeni liniowej. Drugi rozdział będzie traktował o konstrukcji estymatora regresji wypukłej jako rzutu wektora danych na wypukły stożek wielościenny. W trzecim rozdziale zostanie wyznaczony rozkład szukanego testu w przypadku ze znaną wariancją błędu obserwacji.

Praca została napisana na podstawie [2], natomiast rozdział o algorytmie baz prymalno- dualnych został napisany w dużym stopniu na podstawie [1].

1 Stożki wypukłe

Poszukiwany test zostanie wyznaczony metodą rzutowania wektora danych na wielościan powstały w wyniku narzuconych ograniczeń liniowym. W tym rozdziale zostaną przedstwione podstawowe definicje i własności wypuklych stożków wielościennych użyteczne w dalszych rozważaniach.

Definicja 1 (**Ortant**). *Ortantem w n-wymiarowej przestrzeni $\mathbb{R}^n$ nazywamy podzbiór powstały przez ograniczenie każdej ze współrzędnych do bycia nieujemną lub niedodatnią, czyli*

$$O = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n \colon \epsilon_i x_i \geqslant 0, |\epsilon_i| = 1, i = 1, 2, \dots, n\} \tag{1.1}$$

Definicja 2 (**Stożek wypukły**). *Niech V będzie przestrzenią wektorową. Stożkiem wypukłym nazywamy przecięcie skończonej ilości półprzestrzeni przestrzeni V.*

Rozważmy n- wymiarową przestrzeń wektorową V. Dowolną półprzestrzeń H przestrzeni V można wyrazić jako

$$H = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n \colon a_1x_1 + a_2x_2 + \dots a_nx_n \geqslant b\} \tag{1.2}$$

gdzie $a_1, a_2, \dots, a_n, b$ są pewnymi, ustalonymi liczbami rzeczywistymi.

Korzystając z tego przedstawienia możemy dowolny stożek wypukły P zapisać jako

$$K = \bigcap_{i=1}^{m} H_i, \tag{1.3}$$

gdzie

$$H_j = \{\mathbf{x} \in \mathbb{R}^n \colon \sum_{i=1}^{n} a_i^j x_j \geqslant b^j\}. \tag{1.4}$$

Stąd możemy zapisać, że

$$K = \left\{ (x_1, x_2, \dots, x_n) \in \mathbb{R}^n \colon \left\{ \begin{array}{l} \sum_{i=1}^n a_i^1 x_i \geqslant b^1 \\ \sum_{i=1}^n a_i^2 x_i \geqslant b^2 \\ \vdots \\ \sum_{i=1}^n a_i^m x_i \geqslant b^m \end{array} \right. \right\} \tag{1.5}$$

co będziemy zapisywać skrótowo jako

$$K = \{\mathbf{x} \in \mathbb{R}^n \colon \mathbf{A}\mathbf{x} \geqslant \mathbf{b}\} \tag{1.6}$$

gdzie

$$\mathbf{A} = \begin{bmatrix} a_1^1 & a_2^1 & \dots & a_n^1 \\ a_1^2 & a_2^2 & \dots & a_m^2 \\ \vdots & \vdots & \dots & \vdots \\ a_1^n & a_2^n & \dots & a_m^n \end{bmatrix}, \mathbf{b}^T = (b^1, b^2, \dots, b^m), \mathbf{x} = (x_1, x_2, \dots, x_n)^T. \tag{1.7}$$

Symbolem $\langle \cdot, \cdot \rangle$ będziemy oznaczać iloczyn skalarny w przestrzeni wektorowej V. Oznaczmy przez $\boldsymbol{\gamma}_i$ kolejne wiersze macierzy $-\mathbf{A}$. Bez straty ogólności możemy założyć ponadto, że tworzą one układ wektorów liniowo niezależnych, gdyż w przeciwnym wypadku któreś ograniczenie stanowiłoby kombinację pozostałych oraz że $m \leqslant n$. Wtedy stożek K możemy też zapisać w sposób

$$K = \{\mathbf{x} \in \mathbb{R}^n \colon \langle \mathbf{x}, \boldsymbol{\gamma}_i \rangle \leqslant 0, i = 1, 2, \dots, m\} \tag{1.8}$$

Uzupełniając zbiór wektorów $\{\boldsymbol{\gamma}_i\}$ do bazy przestrzni $\mathbb{R}^n$ o wektory ortogonalne i definując bazę dualną złożoną z wektorów $\boldsymbol{\beta_i}$ w następujący sposób

$$\boldsymbol{\beta}_i^T \boldsymbol{\gamma}_j = \begin{cases} -1, & i = j \\ 0, & i \neq j \end{cases} \tag{1.9}$$

możemy zapisać równoważne przedstawienie stożka K

$$K = \{\mathbf{x} \in \mathbb{R}^n \colon \mathbf{x} = \sum_{i=1}^{m} b_i \boldsymbol{\beta}_i + \sum_{i=m+1}^{n} c_i \boldsymbol{\beta}_i, b_i \geqslant 0, c_i \in \mathbb{R}\} \tag{1.10}$$

Twierdzenie 1. *Przedstawienia*

$$K = \{\boldsymbol{x} \in \mathbb{R}^n \colon \langle \boldsymbol{x}, \boldsymbol{\gamma}_i \rangle \leqslant 0, i = 1, 2, \dots, m\} \tag{1.11}$$

$$K = \{\boldsymbol{x} \in \mathbb{R}^n \colon \boldsymbol{x} = \sum_{i=1}^{m} b_i \boldsymbol{\beta}_i + \sum_{i=m+1}^{n} c_i \boldsymbol{\beta}_i, b_i \geqslant 0, c_i \in \mathbb{R}\} \tag{1.12}$$

są równoważne.

Dowód. Wektory $\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i, i = 1, 2, \dots, n$ spełniają zależność $\boldsymbol{\beta}_i^T \boldsymbol{\gamma}_i = -1$ oraz $\boldsymbol{\beta}_i \boldsymbol{\gamma}_j = 0,\ i \neq j$. Oznaczając rzez $\mathbf{B}$, $\mathbf{C}$ macierze, których kolumnami są odpowiednio wektory β_i, γ_i, związek ten możemy przedstawić jako $\mathbf{B}^T\mathbf{C} = -\mathbf{I}$. Ze związku $\mathbf{B}^T\mathbf{C} = -\mathbf{I}$ dostajemy, że $\mathbf{C}^T\mathbf{x} = -\mathbf{B}^{-1}\mathbf{x}$. Wyrażenia $\langle \mathbf{x}, \boldsymbol{\gamma}_i \rangle, i = 1, 2, \dots, m$ są pierwszymi m współrzędnymi $\mathbf{C}^T\mathbf{x}$. Zatem wektor $\mathbf{x}$ wyrażony w bazie złożonej z wektorów $\boldsymbol{\beta}_i$ ma pierwsze m współrzędnych nieujemnych wtedy i tylko wtedy, gdy $\langle \mathbf{x}, \boldsymbol{\gamma}_i \rangle \leqslant 0, i = 1, 2, \dots, m$, co dowodzi równoważności przedstawień. □

2 Regresja wypukła

Podobnie jak w przypadku zwykłego estymatora regresji liniowej, który jest rzutem wektora danych na pewną mniej wymiarową podprzestrzeń, tak w przypadku estymatora regresji wypukłej jest on rzutem na pewen wielościan wypukły powstały w wyniku stosownych ograniczeń.

Zbiór nad którym będziemy minimalizować kwadrat błędu powstaje w sposób następujący. Przypuśmy, że wartości x są różne między sobą i uporządkowane rosnąco oraz niech $\theta_i = f(x_i), i = 1, 2, \dots, n$. Rozważając kawałkami liniowe przybliżenie funkcji regresji z węzłami w punktach x_i, wymóg wypukłości może zostać zapisany jako zbiór ograniczeń w postaci nierówności linowych następującej postaci:

$$\theta_i(x_{i+2} - x_{i+1}) - \theta_{i+1}(x_{i+2} - x_i) + \theta_{i+2}(x_{i+1} - x_i) \geqslant 0, i = 1, 2, \dots, n-2 \tag{2.1}$$

Zgodnie z definicją 1 możemy zbiór tych ograniczeń zapisać jako

$$K = \{\mathbf{A}\theta \geqslant 0\} \tag{2.2}$$

gdzie $\mathbf{A}$ jest rzeczywistą macierzą wymiaru $(n-2) \times n$.

W tym momencie problem znalezienia estymatora regresji wypukłej przyjmuje postać

$$\text{minimalizuj } ||\mathbf{y} - \boldsymbol{\theta}||^2 \text{ po } \boldsymbol{\theta} \in K, \tag{2.3}$$

gdzie $||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}}$.

Niech $B = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ oznacza bazę kanoniczną przestrzeni $\mathbb{R}^n$. Oznaczmy przez $\boldsymbol{\gamma}_i = -\mathbf{e}_i^T\mathbf{A}^T$, $i = 1, 2, \dots, n-2$. Wtedy zbiór K możemy zapisać jako $K = \{\boldsymbol{\theta} \in \mathbb{R}^n \colon -\mathbf{e}_i^T\mathbf{A}^T\boldsymbol{\theta} \leqslant 0, i = 1, 2, \dots, n-2\} = \{\boldsymbol{\theta} \in \mathbb{R}^k \colon \langle\boldsymbol{\gamma}_i, \boldsymbol{\theta}\rangle \leqslant 0, i = 1, 2, \dots, n-2\}$.

Z określenia macierzy $\mathbf{A}$ oraz wektorów $\boldsymbol{\gamma}_i, i = 1, 2, \dots, n-2$, widać, że tworzą one układ wektorów liniowo niezależnych. Zatem zbiór $B'_\gamma = \{\boldsymbol{\gamma}_i, i = 1, 2, \dots, n-2\}$ można uzupełnić do bazy B_γ przestrzeni $\mathbb{R}^n$ o wektory $\boldsymbol{\gamma}_{n-1}, \boldsymbol{\gamma}_n$ tak, żeby były one ortogonalne do wszytkich wektorów z bazy B'_γ. Łatwo sprawdzić, że warunek ten spełniają wektory $\boldsymbol{\gamma}_{n-1} = \mathbf{1}$ oraz $\boldsymbol{\gamma}_n = (x - \bar{x}\mathbf{1})$, gdzie $x = (x_1, x_2, \dots, x_n)$, $\bar{x}$ oznacza wartość średnią, $\mathbf{1} = (1, 1, \dots, 1)^T$, a norma $|| \cdot ||$ jest normą zadaną wcześniej.

Teraz możemy zdefiniować bazę $B_\beta = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_n\}$ dualną do bazy B_γ w następujący sposób:

$$\boldsymbol{\beta}_i^T\boldsymbol{\gamma}_j = \begin{cases} -1, & i = j \\ 0, & i \neq j \end{cases} \tag{2.4}$$

Oznaczając przez $\mathbf{B}$ i $\mathbf{C}$ macierze, których kolumnami są odpowiednio wektory $\boldsymbol{\beta}_i$ i $\boldsymbol{\gamma}_i$ związek między nimi możemy wyrazić jako

$$\mathbf{B}^T\mathbf{C} = -\mathbf{I}, \tag{2.5}$$

gdzie $\mathbf{I}$ oznacza macierz jednostkową.

Niech E oznacza podprzestrzeń przestrzeni $\mathbb{R}^n$ rozpiętą przez wektory $\boldsymbol{\beta}_{n-1}, \boldsymbol{\beta}_n$, natomiast $\mathcal{L}(K)$ oznacza przestrzeń rozpiętą przez wektory $\boldsymbol{\beta}_i, i = 1, 2, \ldots, n-2$. Przestrzenie E oraz $\mathcal{L}(K)$ są do siebie ortogonalne, zatem wektor obserwacji $\mathbf{y}$ możemy zapisać jako sumę $\mathbf{y}_E + \mathbf{z}$, gdzie $\mathbf{y}_E$ i $\mathbf{z}$ są rzutami wektora $\mathbf{y}$ odpowiednio na podprzestrzeń E oraz $\mathcal{L}(K)$.

Przykład 1. *Prześledźmy powyższe rozważania na przykładzie dla przypadku czterowymiarowego i równoodległych punktów x_i odległych o 1.*

Macierz ograniczeń G przybiera wtedy postać

$$G = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix} \tag{2.6}$$

Zatem stożek powstały z ograniczeń jest postaci

$$K = \{\theta \in \mathbb{R}^4 \colon \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix} \theta \geqslant 0\} \tag{2.7}$$

Baza wektorów B_γ jest postaci

$$B_\gamma = \Big((-1, 2, -1, 0), (0, -1, 2, -1), (1, 1, 1, 1), \Big(-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}\Big)\Big) \tag{2.8}$$

Wektory β_i spełniające warunek $\langle \beta_i, \gamma_j \rangle = \left\{\begin{smallmatrix} -1, & i=j \\ 0, & i \neq j \end{smallmatrix}\right.$ przybierają następująca postać

$$B_\beta = ((3, -4, -1, 2), (2, -1, -4, 3), (-1, -1, -1, -1), (3, 1, -1, -3)) \tag{2.9}$$

Przestrzenie na które będziemy rzutować wektor danych przybierają postać

$$E = \{t_1(-1, -1, -1, 1) + t_2(3, 1, -1, -3), t_1, t_2 \in \mathbb{R}\} \tag{2.10}$$

$$\mathcal{L}(K) = \{t_1(3, -4, -1, 2) + t_2(2, -1, -4, 3), t_1, t_2 \in \mathbb{R}\} \quad \diamondsuit$$

Zadanie znalezienia rzutu wektora danych na stożek K sprowadza się w tym momencie do znalezienia rzutu jego składowych na stożek K. Wszytkie elementy podprzestrzeni E należą do stożka K, więc rzut wektora $\mathbf{y}$ na stożek K jest tym samym co jego rzut na podprzestrzeń E i wyraża się wzorem

$$\mathbf{y}_E = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.11}$$

gdzie macierz $\mathbf{X}$ jest podmacierzą macierzy $\mathbf{A}$ złożoną z pierwszych $n-2$ kolumn. Pozostaje zagadnienie znalezienia rzutu $\mathbf{z}$ na stożek K. Sprowadza się ono do znalezienia rzutu $\mathbf{z}$ na stożek

$$K' = K \cap \mathcal{L}(K) = \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{\theta} = \sum_{i=1}^{n-2} b_i \boldsymbol{\beta}_i, b_i \geqslant 0\}. \tag{2.12}$$

W pracy [1] zostało pokazane, że przestrzeń $\mathcal{L}(K)$ może zostać podzielona na 2^{n-2} rozłącznych regionów w taki sposób, że każdy z nich może byś opisany jako nieujemny ortant w bazie $B_J = \{\boldsymbol{\beta}_i, i \in J, \boldsymbol{\gamma}_i, i \in L \setminus J\}$, gdzie J jest pewnym podzbiorem zbioru $L = \{1, 2, \ldots, n-2\}$. Zatem każdy element z należacy do $\mathcal{L}(K)$ może być przedstawiony w następujący sposób

$$z = \sum_{i \in J} b_i \boldsymbol{\beta}_i + \sum_{i \in L \setminus J} c_i \boldsymbol{\gamma}_i, \;\; b_i > 0, c_i \geqslant 0 \tag{2.13}$$

Dla dowolnego zbioru $J \subset L$ B_J jest bazą przestrzeni $\mathcal{L}(K)$, ponadto $\boldsymbol{\beta_i}, i \in J$ oraz $\boldsymbol{\gamma}_i, i \in L \setminus J$ są wzajemnie ortogonalne, zatem rzutem z na K' jest wektor postaci

$$z_{K'} = \sum_{i \in J} b_i \boldsymbol{\beta}_i, b_i > 0 \tag{2.14}$$

Podsumowując, dowolny wektor $\mathbf{y}$ z przestrzeni $\mathbb{R}^n$ można przedstawić w następującej postaci

$$\mathbf{y} = \mathbf{z} + \mathbf{y}_E = \sum_{i \in J} b_i \boldsymbol{\beta}_i + \sum_{i \in L \setminus J} c_i \boldsymbol{\gamma}_i + d_1 \boldsymbol{\gamma}_{n-1} + d_2 \boldsymbol{\gamma}_n, \;\; b_i > 0, c_i \geqslant 0, d_1, d_2 \in \mathbb{R} \tag{2.15}$$

Wtedy rzut tego wektora na stożek

$$K = \{\mathbf{A}\boldsymbol{\theta} \geqslant 0\} \tag{2.16}$$

jest postaci

$$\hat{\boldsymbol{\theta}} = \sum_{i \in J} b_i \boldsymbol{\beta}_i + d_1 \boldsymbol{\gamma}_{n-1} + d_2 \boldsymbol{\gamma}_n. \tag{2.17}$$

Natomiast wektor błędu $\hat{\boldsymbol{\rho}} = \mathbf{y} - \hat{\boldsymbol{\theta}}$ jest postaci

$$\hat{\boldsymbol{\rho}} = \sum_{i \in L \setminus J} c_i \boldsymbol{\gamma_i}. \tag{2.18}$$

3 Test statystyczny i jego rozkład

Na początek wprowadzimy kilka oznaczeń i udowodnimy cztery lematy z których skorzystamy w dalszej części rozważań.

$$C_{L\setminus J} = \{\boldsymbol{y} \in \mathbb{R}^n : y = \sum_{i\in L\setminus J} b_i\boldsymbol{\gamma}_i + \sum_{i\in J} c_i\boldsymbol{\beta}_i + d_1\boldsymbol{\gamma}_{n-1} + d_2\boldsymbol{\gamma}_n, b_i > 0, c_i \geqslant 0, d_1, d_2 \in \mathbb{R}\} \tag{3.1}$$

$$S_{L\setminus J} = \operatorname{span}\{\boldsymbol{\gamma}_i, i \in L \setminus J\} \tag{3.2}$$

$$d = |L \setminus J| = n - 2 - |J| \tag{3.3}$$

Niech

$$\boldsymbol{A}_{L\setminus J} \tag{3.4}$$

oznacza macierz wymiaru $(n-2) \times n$ taką, że pierwsze d wierszy to wektory $-\boldsymbol{\gamma}_i, i \in L \setminus J$ natomiast pozostałe $n-2-d$ wierszy to wektory $-\boldsymbol{\beta}_i, i \in J$.

Lemat 1. *Niech $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n)^T \sim N_n(0, \boldsymbol{I})$ oraz niech $\boldsymbol{A}$ będzie rzeczywistą macierzą wymiaru $m \times n$. Wtedy rozkładem warunkowym $||\boldsymbol{Z}||^2$ pod warunkiem $\boldsymbol{AZ} \geqslant 0$ jest χ^2_n, o ile zbiór $\{\boldsymbol{Z} : \boldsymbol{AZ} \geqslant 0\}$ jest niepusty.*

Dowód. Chcemy pokazać, że $P(||\boldsymbol{Z}||^2 \leqslant a|\boldsymbol{AZ} \geqslant 0) = \chi^2_n(a)$. W tym celu zapiszmy wektor $\boldsymbol{Z}$ we współrzędnych biegunowych

$$\begin{array}{rcl} Z_1 &=& r\cos\phi_1\cos\phi_2\cos\phi_3\ldots\cos\phi_{n-1} \\ Z_2 &=& r\sin\phi_1\cos\phi_2\cos\phi_3\ldots\cos\phi_{n-1} \\ Z_3 &=& r\sin\phi_2\cos\phi_3\ldots\cos\phi_{n-1} \\ Z_4 &=& r\sin\phi_3\ldots\cos\phi_{n-1} \\ \vdots && \\ Z_n &=& r\sin\phi_{n-1}, \end{array} \tag{3.5}$$

gdzie $r \in (0, \infty), \phi_i \in [0, 2\pi), i = 1, 2, \ldots, n-1$.
Wtedy

$$||\boldsymbol{Z}||^2 = r^2 \tag{3.6}$$

oraz

$$\boldsymbol{AZ} \geqslant 0 \iff \boldsymbol{A}\begin{bmatrix} \cos\phi_1\cos\phi_2\cos\phi_3\ldots\cos\phi_{n-1} \\ \sin\phi_1\cos\phi_2\cos\phi_3\ldots\cos\phi_{n-1} \\ \sin\phi_2\cos\phi_3\ldots\cos\phi_{n-1} \\ \sin\phi_3\ldots\cos\phi_{n-1} \\ \vdots \\ \sin\phi_{n-1} \end{bmatrix} \geqslant 0. \tag{3.7}$$

Widzimy zatem, że wartość $||\boldsymbol{Z}||^2$ zależy jedynie od wartości r natomiast warunek $\boldsymbol{AZ} \geqslant 0$ dotyczy jedynie kąta, który jest niezależny od promienia r, a zatem $P(||\boldsymbol{Z}||^2 \leqslant a|\boldsymbol{AZ} \geqslant 0) = \chi^2_n(a)$. □

Lemat 2. *Niech $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n)^T \sim N(0, \boldsymbol{I})$ oraz niech $\hat{\boldsymbol{Z}}$ będzie rzutem $\boldsymbol{Z}$ na przestrzeń liniową S wymiaru $d < n$. Ponadto niech $\boldsymbol{A}$ będzie rzeczywistą macierzą wymiaru $m \times n$ taką, że każdy jej wiersz jest ortogonalny do przestrzeni S. Wtedy rozkładem warunkowym $||\hat{\boldsymbol{Z}}||^2$ pod warunkiem $\boldsymbol{AZ} \geqslant 0$ jest χ^2_d, o ile zbiór $\{\boldsymbol{AZ} \geqslant 0\}$ jest niepusty.*

Dowód. Niech $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ będą wzjamnie ortonormalnymi wektorami w $\mathbb{R}^n$ takimi, że wektory $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$ rozpinają przestrzeń S. Wektor $\boldsymbol{Z}$ możemy zapisać jako $\boldsymbol{Z} = \sum_{i=1}^n a_i \boldsymbol{v}_i$, gdzie $a_i = \langle \boldsymbol{v}_i, \boldsymbol{z} \rangle$. Stąd $a_i, i = 1, 2, \ldots, n$ są niezależnymi zmiennymi losowymi o standardowym rozkładzie normalnym oraz $\hat{\boldsymbol{Z}} = \sum_{i=1}^d a_i \boldsymbol{v}_i$. Wtedy $||\hat{\boldsymbol{Z}}||^2 = a_1^2 + a_2^2 + \cdots + a_d^2$ co ma gęstość χ^2_d. Niech teraz $\boldsymbol{V}$ oznacza macierz taką, której poszczególne kolumny są kolejno wektorami $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ oraz niech $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$. Macierz $\boldsymbol{V}$ możemy zapisać jako $\boldsymbol{V} = [\boldsymbol{V}_1|\boldsymbol{V}_2]$, gdzie $\boldsymbol{V}_1$ jest macierzą wymiaru $n \times d$, oznaczmy też przez $\boldsymbol{a}^1$ wektor $(a_1, a_2, \ldots, a_d)^T$ a przez $\boldsymbol{a}^2$ wektor $(a_{d+1}, \ldots, a_n)^T$. Wtedy $\boldsymbol{Z} = \boldsymbol{Va} = \boldsymbol{V}_1\boldsymbol{a}^1 + \boldsymbol{V}_2\boldsymbol{a}^2$ a warunek $\boldsymbol{AZ} \geqslant 0$ możemy zapisać jako $\boldsymbol{AV}_1\boldsymbol{a}^1 + \boldsymbol{AV}_2\boldsymbol{a}^2 \geqslant 0$. Zauważmy, że z założeń oraz konstrukcji macierzy $\boldsymbol{V}$ dostajemy, że $\boldsymbol{AV}_1 = 0$ oraz $\boldsymbol{a}^1, \boldsymbol{a}^2$ są niezależne. Zatem warunek $\boldsymbol{AZ} \geqslant 0$ nie wpływa na gęstość $||\hat{\boldsymbol{Z}}||^2$. □

Lemat 3. *Niech $\boldsymbol{y} \in C_J$ dla pewngo zbioru $J \subset L = \{1, 2, \ldots, n-2\}$ oraz niech $a, b \in \mathbb{R}$. Wtedy $\boldsymbol{y}' = \boldsymbol{y} + a\boldsymbol{\gamma}_{n-1} + b\boldsymbol{\gamma}_n \in C_J$ oraz wektory błędów $\boldsymbol{\rho} = \boldsymbol{y} - \hat{\boldsymbol{\theta}}$ i $\boldsymbol{\rho}' = \boldsymbol{y}' - \hat{\boldsymbol{\theta}}'$ są sobie równe.*

Dowód. Jeśli $\boldsymbol{y} \in C_J$ to $\boldsymbol{y}$ możemy zapisać jako $\boldsymbol{y} = \sum_{i\in J} b_i\boldsymbol{\beta}_i + \sum_{i\in L\setminus J} c_i\boldsymbol{\gamma}_i + d_1\boldsymbol{\gamma}_{n-1} + d_2\boldsymbol{\gamma}_n,\ b_i > 0, c_i \geqslant 0, d_1, d_2 \in \mathbb{R}$. Wtedy $\boldsymbol{y}' = \sum_{i\in J} b_i\boldsymbol{\beta}_i + \sum_{i\in L\setminus J} c_i\boldsymbol{\gamma}_i + (d_1 + a)\boldsymbol{\gamma}_{n-1} + (d_2 + b)\boldsymbol{\gamma}_n,\ b_i > 0, c_i \geqslant 0, d_1, d_2 \in \mathbb{R}$. Oczywiście $d_1 + a, d_2 + b \in \mathbb{R}$ zatem $\boldsymbol{y}' \in C_J$.

Wektor $\boldsymbol{\rho}$ jest postaci $\boldsymbol{\rho} = \sum_{i\in L\setminus J} c_i\boldsymbol{\gamma}_i$. Z postaci wktora $\boldsymbol{y}'$ widzimy jednak, że $\boldsymbol{\rho}' = \sum_{i\in L\setminus J} c_i\boldsymbol{\gamma}_i = \boldsymbol{\rho}$. □

Lemat 4. *Wektory losowe $\boldsymbol{y} - \hat{\boldsymbol{\theta}}$ i $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{y}}$ są niezależne.*

Dowód. Zauważmy, że $\langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{y}} - \hat{\boldsymbol{\theta}} \rangle = \langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{y}} \rangle + \langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}} \rangle = 0 + 0 = 0$. □

Dla ułatwienia rozważań założymy, że wariancja w zaproponowanym modelu σ^2 jest znana.

Teraz możemy przystąpić do wyliczania rozkładu testu opartego o iloraz wiarogodności hipotezy

$$H_0\colon f(x) = ax + b \ vs. \ H_1\colon f \in \mathcal{F} \tag{3.8}$$

gdzie $\mathcal{F}$ jest klasą funkcji wypukłych.

Niech $\hat{\boldsymbol{y}}$ oznacza estymator regresji liniowej, czyli rzut wektora danych $\boldsymbol{y}$ na przestrzeń span$\{\boldsymbol{\gamma}_{n-1}, \boldsymbol{\gamma}_n\}$. Ponadto oznaczmy przez $R_0 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ oraz $R_1 = \sum_{i=1}^n (y_i - \hat{\theta}_i)^2$. Wtedy poszukiwany test przyjmuje postać

$$M = \frac{R_0 - R_1}{\sigma^2}. \tag{3.9}$$

W celu znalezienia rozkładu testu M, gdy prawdziwa jest hipoteza zerowa potrzebne będzie znalezienie wymiaru modelu i liczby stopni swobody błędu dla modelu regresji wypukłej. Z postaci rzutu $\hat{\boldsymbol{\theta}}$ wektora $\boldsymbol{y}$ można przypuszczać, że wymiar modelu wynosi $n-d$ oraz liczba stopni swobody błędu wynosi d. Jednak zbiór J jest losowy, różne wartości wektora błędu $\boldsymbol{\varepsilon}$ mogą umieścić wektor danych $\boldsymbol{y}$ w różnych zbiorach C_J. Wyliczenie rozkładu testu zaczniemy w następujący sposób

$$P(M \leqslant a) = \sum_{J \in \mathcal{P}(L)} P(M \leqslant a, \boldsymbol{y} \in C_J = \sum_{J \in \mathcal{P}(L)} P(M \leqslant a | \boldsymbol{y} \in C_J) P(\boldsymbol{y} \in C_J)(3.10)$$

gdzie $\mathcal{P}(L)$ oznacza zbiór potęgowy zbioru L.

Z lematu 3. możemy bez straty ogólności założyć, że $f(x) = 0$ i $\boldsymbol{y} = \boldsymbol{\varepsilon}$. Dla dowolnej realizacji wektora $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ oznaczmy przez $L \setminus J$ taki zbiór indeksów, że $\boldsymbol{\varepsilon} \in C_{L \setminus J}$. Niech $\hat{\boldsymbol{\varepsilon}}$ będzie rzutem wektora $\boldsymbol{\varepsilon}$ na przestrzeń $S_{L \setminus J}$. Zauważmy, że macierz $\boldsymbol{A}_{L \setminus J}$ można zapisać jako $[\boldsymbol{A}^1 | \boldsymbol{A}^2]$, gdzie macierz $\boldsymbol{A}^1$ jest wymiaru $d \times n$. Zatem kolumny macierzy $\boldsymbol{A}^1$ rozpinają $S_{L \setminus J}$, natomiast kolumny macierzy $\boldsymbol{A}^2$ są ortogonalne do przestrzeni $S_{L \setminus J}$. Dodatkowo, gdy $\boldsymbol{\varepsilon} \in C_J$, zachodzi $\boldsymbol{A}^1 \boldsymbol{\varepsilon} \geqslant 0$ oraz $\boldsymbol{A}^2 \boldsymbol{\varepsilon} \geqslant 0$. Stąd na mocy lematu 2. dostajemy, że rozkładem warunkowym $\frac{||\hat{\boldsymbol{\varepsilon}}||^2}{\sigma^2}$ przy zadanym J jest χ^2_d. Jako że $R_1 = ||\hat{\boldsymbol{\rho}}||^2$ a przy założeniu prawdziwości hipotezy zerowej $||\hat{\boldsymbol{\rho}}||^2$ jest równa $\frac{||\hat{\boldsymbol{\varepsilon}}||^2}{\sigma^2}$ przy ustalonym zbiorze J możemy napisać następujący wniosek

Wniosek 1. *Jeśli hipoteza zerowa $\boldsymbol{\theta} \in$ span$\{\boldsymbol{\gamma}_{n-1}, \boldsymbol{\gamma}_n\}$ jest prawdziwa to rozkładem warunkowym $\frac{R_1}{\sigma^2}$ przy ustalonym $\boldsymbol{y} \in C_J$ jest χ^2_d, gdzie $d = |L \setminus J|$.*

Zmienna losowa $\frac{R_0}{\sigma^2}$ ma oczywiście rozkład χ^2_{n-2}. Niech D będzie zmienną losową reprezentującą liczność zbioru $L \setminus J$. Z Wniosku 1. mamy, że rozkładem warunkowym $\frac{R_1}{\sigma^2}$ pod warunkiem $D = d$ jest χ^2_d. Z lematu 4. dostajemy zatem, że rozkładem warunkowym M jest χ^2_{n-d-2} pod warunkiem $D = d$. Stąd możemy zapisać następujący wniosek

Wniosek 2. *Przy założeniu prawdziwości hipotezy zerowej postawionego problemu mamy*

$$P(M \leqslant a) = \sum_{d=0}^{n-2} P(\chi^2_{n-d-2} \leqslant a) P(D = d), \tag{3.11}$$

gdzie $\chi^2_0 \equiv 0$.

Wartości prawdopodobieństw $P(D = d), d = 0, 1, \ldots, n-2$ jest wyliczane na podstawie względnych objętości zbiorów $C_J, J \in \mathcal{P}(L)$. Prawdopodobieństwo, że $\boldsymbol{y} \in C_J$, gdy hipoteza zerowa jest prawdziwa, jest równoważne prawdopodobieństwu, że wektor losowy o n- wymiarowym standardowym rozkładzie normalny wpada do zbioru C_J.

Kropka nie oznacza końca zdania.
Ona daje możliwość coraz to lepszej kontynuacji.

Bibliografia

[1] Fraser D.A.S., Massam H., *A Mixed Primal- Dual Bases Algorithm for Regression under Inequality Constraints. Application to Concave regression*, Scand J. Statist, **16** 65-74, 1989

[2] Meyer Mary C., *A test for linear vs convex regression function using shape-restricted regression*, Stanford University, Technical Report No. 2001-20, sierpień 2001