Liniowy model statystyczny

Modelem liniowym nazywamy taki model statystyczny, w którym obserwacje $Y_1,...,Y_n$ mają postać $Y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_{p-1}x_{ip-1}+\varepsilon_i$, $i=1,2,...,n$, gdzie $x_{i,j}$ są ustalonymi liczbami, ε_i są „błędami losowymi", a β_j, $j=0,...,p-1$ są nieznanymi stałymi. Powyższy model można zapisać w klasycznej postaci wektorowo-macierzowej oznaczanej $(\mathbf{Y},\mathbf{X\beta},\sigma^2\mathbf{I})$, symbolizującej model spełniający 3 następujące założenia

1. $\mathbf{Y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$
2. $E(\boldsymbol{\varepsilon})=\mathbf{0}$
3. $V(\boldsymbol{\varepsilon})=\sigma^2\mathbf{I}$

gdzie $\mathbf{Y}=\begin{bmatrix}Y_1\\ \vdots\\ Y_n\end{bmatrix}$, $\mathbf{X}=\begin{bmatrix}1 & x_{11} & \cdots & x_{1\,p-1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{n\,p-1}\end{bmatrix}$, $\boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix}$, $\boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \vdots\\ \beta_{p-1}\end{bmatrix}$.

Model ten przedstawia losowy wektor $\mathbf{Y}$, którego wartość oczekiwana $E(\mathbf{Y})=\mathbf{X\beta}$ jest liniową kombinacją kolumn nielosowej macierzy $\mathbf{X}$, zwanej macierzą planu eksperymentu (w R model.matrix) . Nieznane są parametry $\beta_0,...,\beta_{p-1}$ (współrzędne wektora $\boldsymbol{\beta}$). Losowy wektor zakłóceń $\boldsymbol{\varepsilon}$ ma zerową wartość oczekiwaną a jego składowe są nieskorelowane i mają wspólną (nieznaną) wariancję σ^2.

Zagadnienie oszacowania wektora $\boldsymbol{\beta}$ metodą najmniejszych kwadratów (MNK) polega na wyznaczeniu $\mathbf{b}=\hat{\boldsymbol{\beta}}$ minimalizującego wyrażenie

$$S(\mathbf{b})=\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{p-1}x_{ij}b_j)^2=\|\mathbf{Y}-\mathbf{Xb}\|^2=\langle\mathbf{Y}-\mathbf{Xb},\mathbf{Y}-\mathbf{Xb}\rangle=\langle\mathbf{Y},\mathbf{Y}\rangle-2\langle\mathbf{X}^{\mathrm{T}}\mathbf{Y},\mathbf{b}\rangle+\langle\mathbf{X}^{\mathrm{T}}\mathbf{Xb},\mathbf{b}\rangle$$

gdzie $\|\cdot\|$ oznacza normę a $\langle\cdot,\cdot\rangle$ iloczyn skalarny w R^n. Wyrażenie $S(\mathbf{b})$ jest funkcją kwadratową wektora $\mathbf{b}$.

Rozwiązanie analityczne. WK istnienia ekstremum, to zerowanie się wszystkich pochodnych cząstkowych $\frac{\partial S(\mathbf{b})}{\partial b_i}=0$, $i=0,..,p-1$. W przypadku funkcji kwadratowej ten warunek jest równoważny znikaniu pochodnej Frecheta (macierz pochodnych cząstkowych jest macierzą zerową)

$$\begin{aligned}&S(\mathbf{b}+\Delta\mathbf{b})-S(\mathbf{b})=\\&=\langle\mathbf{Y},\mathbf{Y}\rangle-2\langle\mathbf{X}^{\mathrm{T}}\mathbf{Y},\mathbf{b}+\Delta\mathbf{b}\rangle+\langle\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}+\Delta\mathbf{b}),\mathbf{b}+\Delta\mathbf{b}\rangle-\langle\mathbf{Y},\mathbf{Y}\rangle+2\langle\mathbf{X}^{\mathrm{T}}\mathbf{Y},\mathbf{b}\rangle-\langle\mathbf{X}^{\mathrm{T}}\mathbf{Xb},\mathbf{b}\rangle=\\&=2\langle\mathbf{X}^{\mathrm{T}}\mathbf{Xb}-\mathbf{X}^{\mathrm{T}}\mathbf{Y},\Delta\mathbf{b}\rangle+\langle\mathbf{X}^{\mathrm{T}}\mathbf{X}\Delta\mathbf{b},\Delta\mathbf{b}\rangle\end{aligned}$$

Z nierówności Schwarza mamy $|\langle \mathbf{X}^T\mathbf{X}\Delta\mathbf{b}, \Delta\mathbf{b}\rangle| \le \|\mathbf{X}^T\mathbf{X}\Delta\mathbf{b})\|\|\Delta\mathbf{b}\|$, a z definicji normy operatora liniowego ciągłego $\|\mathbf{X}^T\mathbf{X}\Delta\mathbf{b})\| \le \|\mathbf{X}^T\mathbf{X}\|\|\Delta\mathbf{b}\|$. Wobec tego $\frac{|\langle \mathbf{X}^T\mathbf{X}\Delta\mathbf{b}, \Delta\mathbf{b}\rangle|}{\|\Delta\mathbf{b}\|} \le \|\mathbf{X}^T\mathbf{X}\|\|\Delta\mathbf{b}\| \to 0$, gdy $\Delta\mathbf{b} \to \mathbf{0}$, a macierz $2(\mathbf{X}^T\mathbf{X}\mathbf{b} - \mathbf{X}^T\mathbf{Y})$ jest macierzą pochodnej mocnej. Warunek zerowania się tej macierzy prowadzi do układu równań

(N) $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$,

zwanego układem normalnym, który ma zawsze rozwiązanie (gdyż $\mathrm{Im}\,\mathbf{X}^T = \mathrm{Im}\,\mathbf{X}^T\mathbf{X}$)

Rozwiązanie geometryczne. Jeżeli $\mathbf{Y}=\mathbf{0}$, to rozwiązaniem jest $\mathbf{b} = \mathbf{0}$. Załóżmy że $\mathbf{Y}\neq\mathbf{0}$. Norma $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|$ jest minimalizowana przez takie $\mathbf{b}$, że $\mathbf{X}\mathbf{b}$ jest rzutem wektora $\mathbf{Y}$ na podprzestrzeń generowaną przez kolumny macierzy planu eksperymentu tzn. $\mathrm{span}\{\mathbf{X}\boldsymbol{\beta}: \boldsymbol{\beta}\in R^p\} = \mathrm{Im}\mathbf{X}$, a więc przez $\mathbf{b}$ spełniające układ równań $\langle \mathbf{Y} - \mathbf{X}\mathbf{b}, \mathbf{X}_j\rangle = 0$, $j=0,\ldots,p-1$ (gdzie $\mathbf{X}_j$ oznacza j-tą kolumnę macierzy $\mathbf{X}$), który bezpośrednio prowadzi do układu równań

(N) $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$,

W przypadku, gdy macierz $\mathbf{X}^T\mathbf{X}$ jest nieosobliwa otrzymujemy jednoznaczne rozwiązanie

$\mathbf{b} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$,

które jest estymatorem MNK wektora parametrów $\boldsymbol{\beta}$, natomiast $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$ jest rzutem wektora $\mathbf{Y}$ na podprzestrzeń Im$\mathbf{X}$. Macierz $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ zwana macierzą daszkową (hat matrix- dodaje daszek do $\mathbf{Y}$) jest macierzą idempotentną i symetryczną, więc reprezentuje rzut ortogonalny na Im$\mathbf{X}$.

Twierdzenie (Gauss,Markow) W modelu $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ najlepszym liniowym, nieobciążonym (**B**est **L**inear **U**nbiased **E**stimator - BLUE) estymatorem wektora $\boldsymbol{\beta}$ jest wektor $\mathbf{b} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ uzyskany MNK.

Uwaga. Klasa estymatorów liniowych to klasa $\{\mathbf{L}\mathbf{Y}: \mathbf{L}\in\mathcal{M}_{pn}$, gdzie $\mathcal{M}_{pn}$ oznacza zbiór rzeczywistych macierzy typu $(p,n)\}$. Dla danego estymatora liniowego $\mathbf{L}\mathbf{Y}$ określamy tzw. ryzyko macierzowe $\mathbf{M}(\mathbf{L}) = E[(\mathbf{L}\mathbf{Y}-\boldsymbol{\beta})(\mathbf{L}\mathbf{Y}-\boldsymbol{\beta})^T]$. Estymator $\mathbf{L}_1\mathbf{Y}$ jest lepszy od estymatora $\mathbf{L}_2\mathbf{Y}$, jeżeli macierz $\mathbf{M}(\mathbf{L}_2) - \mathbf{M}(\mathbf{L}_1)$ jest macierzą nieujemnie określoną i nie równą tożsamościowo macierzy zerowej.

Dowód. Estymator MNK $\mathbf{b} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ jest oczywiście liniowy tzn. jest liniową funkcją wektora obserwacji. Jest on nieobciążony bo

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

Ponadto $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$, stąd

$$V(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{T} = E\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}\}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} =$$
$$= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\{\sigma^{2}\mathbf{I}\}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^{2}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

Rozważmy dowolny inny estymator liniowy $\tilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{Y}$. Warunek nieobciążoności $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ $\forall\boldsymbol{\beta}$, implikuje warunek $\mathbf{CX}\boldsymbol{\beta}=\boldsymbol{\beta}$, $\forall\boldsymbol{\beta}$, skąd otrzymujemy **warunek nieobciążoności** $\mathbf{CX}=\mathbf{I}$.

$$V(\tilde{\boldsymbol{\beta}}) = E(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} = E\{(\mathbf{C}\mathbf{Y} - \boldsymbol{\beta})(\mathbf{C}\mathbf{Y} - \boldsymbol{\beta})^{\mathrm{T}}\} = E\{(\mathbf{C}\mathbf{Y} - \mathbf{C}\mathbf{X}\boldsymbol{\beta})(\mathbf{C}\mathbf{Y} - \mathbf{C}\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\} =$$
$$= E\{\mathbf{C}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\} = \mathbf{C}E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}\}\mathbf{C}^{\mathrm{T}} = \sigma^{2}\mathbf{C}\mathbf{C}^{\mathrm{T}}$$

Z uwagi na warunek $\mathbf{CX}=\mathbf{I}$ macierz kowariancji $V(\hat{\boldsymbol{\beta}}) = \sigma^{2}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ można zapisać w postaci

$$V(\hat{\boldsymbol{\beta}}) = \sigma^{2}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^{2}\mathbf{C}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}$$

Wobec tego $V(\tilde{\boldsymbol{\beta}}) - V(\hat{\boldsymbol{\beta}}) = \sigma^{2}\mathbf{C}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{C}^{\mathrm{T}}$.

Ale macierz $\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ jest symetryczna i idempotentna, (wobec tego jest to macierz rzutowania ortogonalnego), ma wiec wartości własne równe 0 lub 1, czyli jest nieujemnie określona

Nieujemna określoność $\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ oznacza, że $\mathbf{w}^{\mathrm{T}}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{w}\geq 0$ $\forall\mathbf{w}$, stąd w szczególności dla $\mathbf{w}=\mathbf{C}^{\mathrm{T}}\mathbf{z}$ **mamy** $\mathbf{z}^{\mathrm{T}}\mathbf{C}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{C}^{\mathrm{T}}\mathbf{z}\geq 0$ $\forall\mathbf{z}$ co oznacza nieujemną określoność macierzy $\mathbf{C}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{C}^{\mathrm{T}}$.

Ostatecznie $V(\tilde{\boldsymbol{\beta}}) - V(\hat{\boldsymbol{\beta}})$ jest macierzą nieujemnie określoną.

Konsekwencja nieujemnej określoności $V(\tilde{\boldsymbol{\beta}}) - V(\hat{\boldsymbol{\beta}})$. Rozważmy dowolną liniową kombinacje współrzędnych wektorów $\tilde{\boldsymbol{\beta}}$ i $\hat{\boldsymbol{\beta}}$ tzn. $\mathbf{c}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ i $\mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. Obliczmy różnicę ich wariancji

$$V(\mathbf{c}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}) - V(\mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \mathbf{c}^{\mathrm{T}}V(\tilde{\boldsymbol{\beta}})\mathbf{c} - \mathbf{c}^{\mathrm{T}}V(\hat{\boldsymbol{\beta}})\mathbf{c} = \mathbf{c}^{\mathrm{T}}(V(\tilde{\boldsymbol{\beta}}) - V(\hat{\boldsymbol{\beta}}))\mathbf{c} \geq 0.$$

Wariancja dowolnej liniowej funkcji parametrycznej współrzędnych wektora $\tilde{\boldsymbol{\beta}}$ jest nie mniejsza od wariancji liniowej funkcji parametrycznej współrzędnych wektora $\hat{\boldsymbol{\beta}}$ uzyskanego MNK. W szczególności przyjmując jako z wektor $\mathbf{c}$ *i*- ty wektor bazy kanonicznej w R^{k} otrzymujemy $V(\tilde{\beta}_{i}) \geq V(\hat{\beta}_{i})$ i=1,...,n czyli estymator MNK daje **estymatory liniowe o minimalnej wariancji**.

- Niezwykłą cechą twierdzenia Gaussa Markowa jest ogólność założeń dotyczących rozkładu. Wynik obowiązuje dla dowolnego rozkładu – normalność nie jest tu wymagana. Zakładamy jedynie $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ i $V(\mathbf{Y}) = \sigma^{2}\mathbf{I}$. Jeżeli te założenia nie są spełnione $\hat{\boldsymbol{\beta}}$ może być estymatorem obciążonym lub może mieć większą wariancję niż inny estymator liniowy wektora $\boldsymbol{\beta}$.

- W dowodzie twierdzenia GM korzystaliśmy z założenia nieosobliwości macierzy $\mathbf{X}^T\mathbf{X}$. Twierdzenie jest jednak prawdziwe także, gdy macierz $\mathbf{X}$ nie jest pełnego rzędu (więc $(\mathbf{X}^T\mathbf{X})^{-1}$ nie istnieje i jest zastąpiona przez uogólnioną odwrotność $(\mathbf{X}^T\mathbf{X})^-$).

Def. Liniowa funkcja parametryczna $\mathbf{c}^T\boldsymbol{\beta}$ jest (nieobciążenie) estymowalna, gdy istnieje liniowy nieobciążony estymator $\mathbf{b}^T\mathbf{Y}$ tej funkcji.

WKW estymowalności $E(\mathbf{b}^T\mathbf{Y}) = \mathbf{b}^T E(\mathbf{Y}) = \mathbf{b}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{c}^T\boldsymbol{\beta}$, $\forall\boldsymbol{\beta} \Leftrightarrow \mathbf{X}^T\mathbf{b} = \mathbf{c} \Leftrightarrow \mathbf{c} \in \mathrm{Im}(\mathbf{X}^T)$
(wiersz $\mathbf{c}^T$ jest liniową kombinacją wierszy macierzy $\mathbf{X}$.

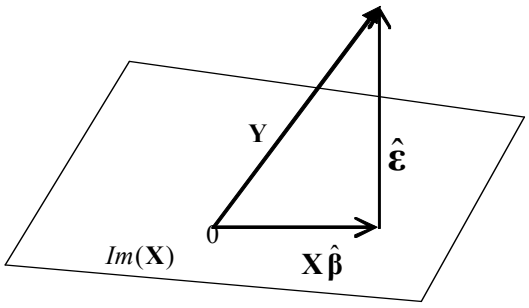
Jeśli macierz $\mathbf{X}$ jest pełnego rzędu to, każda liniowa funkcja parametryczna jest estymowalna. Jeśli $\mathbf{X}$ nie jest pełnego rzędu to w interpretacji twierdzenia GM należy ograniczyć się tylko do nieobciążenia estymowalnych liniowych funkcji parametrycznych

Def. Hipoteza liniowa $\mathbf{G}\boldsymbol{\beta}=\mathbf{0}$ jest testowalna, jeżeli estymowalne są wszystkie liniowe funkcje parametryczne generowane przez wiersze macierzy $\mathbf{G}$

WKW testowalności : $\mathrm{Im}(\mathbf{G}^T) \subset \mathrm{Im}(\mathbf{X}^T)$
(wiersze macierzy **G** są liniowymi kombinacjami wierszy macierzy **X)**

Fakt Niech H będzie przestrzenią Hilberta a P: $H \to H$ operatorem liniowym samosprzeżonym i idempotentnym $P^2=P$. Wówczas P jest operatorem rzutowania na pewną podprzestrzeń przestrzeni H (tzn. rzut na Im(P)).

Szkic dowodu. Dowolny wektor $\boldsymbol{x}$ można zapisać w postaci $\mathbf{x}=P\mathbf{x}+(I-P)\mathbf{x}$. Zauważmy, że $\langle P\mathbf{x}, (I-P)\mathbf{x}\rangle = \{\text{samosprzężoność } P\} = \langle \mathbf{x}, P(I-P)\mathbf{x}\rangle = \langle x, P\mathbf{x} - P^2\mathbf{x}\rangle = 0$, czyli wektory $P\mathbf{x}$ i $(I-P)\mathbf{x}$ są ortogonalne. Z jednoznaczności przedstawienia wektora $\mathbf{x}=P\mathbf{x}+(I-P)\mathbf{x}$ mamy $H = \mathrm{Im}(P) \oplus \mathrm{Ker}(P)$ i $P\mathbf{x}$ jest rzutem ortogonalnym wektora $\mathbf{x}$ na Im(P) a $(I-P)\mathbf{x}$ jest rzutem ortogonalnym na Im($I-P$)=Ker(P).



Estymacja σ^2 . Rozważmy reszty MNK

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = (\mathbf{I}-\mathbf{H})\mathbf{Y} = (\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}$$

Wyznaczmy

$$E(\|\hat{\boldsymbol{\varepsilon}}\|^2) = E\langle(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon},(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\rangle = E\langle(\mathbf{I}-\mathbf{H})^2\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\rangle = E\langle(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\rangle = E(\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}) =$$

$$\{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\text{ jest skalarem}\} = E(tr\{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\}) = \{\text{ tr(}\mathbf{AB}\text{)=tr(}\mathbf{BA}\text{) }\} =$$

$$= E(tr\{(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\}) = \{\text{ przemienność liniowych operacji } E \text{ i } tr\} = tr\{E(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\} =$$

$$= tr\{(\mathbf{I}-\mathbf{H})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}})\} = tr\{(\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I})\} = \sigma^2 tr\{(\mathbf{I}-\mathbf{H})\} = \sigma^2(tr\{\mathbf{I}\}-tr\{\mathbf{H}\}) =$$

$$\sigma^2(tr\{\mathbf{I}\}-tr\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}) = \sigma^2(tr\{\mathbf{I}\}-tr\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\}) = \sigma^2(n-tr\{\mathbf{I}_p\}) = \sigma^2(n-p)$$

Stąd estymator

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n-p}\|\hat{\boldsymbol{\varepsilon}}\|^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\boldsymbol{\varepsilon}}_i^2$$

jest nieobciążonym estymatorem σ^2.

Nieobciążonym estymatorem macierzy kowariancji $\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ estymatora wektorowego $\hat{\boldsymbol{\beta}}$ jest $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

Dokładamy do modelu $(\mathbf{Y},\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ założenie o normalności składnika losowego $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I})$.

Wówczas $\mathbf{Y}\sim N(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ a estymator MNK $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ ma rozkład $N_p(\boldsymbol{\beta},\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$.

Z postaci funkcji gęstości rozkładu wektora $\mathbf{Y}$

$$p(\mathbf{Y};\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})} = (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}}e^{\frac{1}{\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}-\frac{1}{2\sigma^2}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}} =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\langle\mathbf{X}\boldsymbol{\beta},\mathbf{X}\boldsymbol{\beta}\rangle}e^{\frac{1}{\sigma^2}\langle\mathbf{X}^{\mathrm{T}}\mathbf{Y},\boldsymbol{\beta}\rangle-\frac{1}{2\sigma^2}\langle\mathbf{Y},\mathbf{Y}\rangle}$$

Przyjmując $T_i(\mathbf{Y}) = \sum_{j=1}^{n}X_{ji}Y_j = \langle\mathbf{X}_i,\mathbf{Y}\rangle$, i=0,...,p-1 oraz $T_p(\mathbf{Y}) = \sum_{j=1}^{n}Y_j^2 = \langle\mathbf{Y},\mathbf{Y}\rangle$ możemy gęstość zapisać w postaci

$$p(\mathbf{Y};\boldsymbol{\beta},\sigma^2) = C(\boldsymbol{\beta},\sigma^2)e^{\frac{1}{\sigma^2}\sum_{i=0}^{p-1}\beta_i T_i(\mathbf{Y})-\frac{1}{2\sigma^2}T_p(\mathbf{Y})}.$$

Jest to gęstość rozkładu regularnej p+1 parametrowej rodziny wykładniczej a statystyka $(T_0(\mathbf{Y}),\ldots,T_{p-1}(\mathbf{Y}),T_p(\mathbf{Y}))$ jest statystką dostateczną zupełną (więc także minimalną dostateczną)

Estymator MNK $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\begin{bmatrix} T_0(\mathbf{Y}) \\ \vdots \\ T_{p-1}(\mathbf{Y}) \end{bmatrix}$ jest funkcją statystyki dostatecznej zupełnej i jak wiadomo jest nieobciążony, skąd wynika że estymator $\hat{\boldsymbol{\beta}}$ uzyskany MNK jest estymatorem nieobciążonym o minimalnej wariancji wektora $\boldsymbol{\beta}$.

Podobnie $\hat{\sigma}^2 = \frac{1}{n-p}\left\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2 = \frac{1}{n-p}\left(\|\mathbf{Y}\|^2 - \left\|\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2\right) = \frac{1}{n-p}\left(\|\mathbf{Y}\|^2 - \left\|\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}(\mathbf{Y})\right\|^2\right)$, gdzie

$\mathbf{T}(\mathbf{Y}) = \begin{bmatrix} T_0(\mathbf{Y}) \\ \vdots \\ T_{p-1}(\mathbf{Y}) \end{bmatrix}$, jest nieobciążonym estymatorem σ^2 będącym funkcją statystyki dostatecznej zupełnej wiec jest estymatorem nieobciążonym o minimalnej wariancji.

Wniosek. Przy założeniu normalności w modelu $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ MNK daje estymatory optymalne (nieobciążone o minimalnej wariancji) w szerszej (niż klasa estymatorów liniowych nieobciążonych o skończonej wariancji) klasie estymatorów nieobciążonych o skończonej wariancji .

Metoda największej wiarygodności w gaussowskim modelu $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

Przy założeniu $\boldsymbol{\varepsilon}$~N($\mathbf{0}$,$\sigma^2\mathbf{I}$) funkcja wiarygodności ma postać

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2}$$

a jej podwojony logarytm

$$2l(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = -n\ln(2\pi\sigma^2) - \frac{1}{\sigma^2}\left\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right\|^2 = -n\ln(2\pi\sigma^2) - \frac{1}{\sigma^2}RSS(\boldsymbol{\beta})$$

Widać, że maksymalizacja $2l(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})$ względem $\boldsymbol{\beta}$ jest równoważna minimalizacji $RSS(\boldsymbol{\beta})$, czyli estymator NW i MNK parametru wektorowego $\boldsymbol{\beta}$ są identyczne.

Natomiast estymator NW parametru σ^2 uzyskujemy z jako rozwiązanie ostatniego p-tego z równań wiarygodności (pierwsze p-1 równań to oczywiście układ (N))

$$-\frac{n}{\sigma^2} + \frac{1}{\sigma^4}RSS(\hat{\boldsymbol{\beta}}) = 0,$$

skąd otrzymujemy $\frac{1}{n}\left\|\hat{\boldsymbol{\varepsilon}}\right\|^2 = \frac{RSS(\hat{\boldsymbol{\beta}})}{n}$ =ENW[σ^2].

Wniosek. W gaussowskim modelu liniowym ($\mathbf{Y}$,$\mathbf{X}\boldsymbol{\beta}$,$\sigma^2\mathbf{I}$) metody MNK, NW i estymacja nieobciążona o minimalnej wariancji prowadzą do tych samych estymatorów wektora $\boldsymbol{\beta}$.

Niespodzianka. Jeżeli dla estymacji wektora $\boldsymbol{\beta}$ przyjmiemy kwadratową funkcję straty $\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2$ to dla

- dla $p \le 2$ estymator MNK jest dopuszczalny

- dla p >2 estymator MNK jest niedopuszczalny – lepszy przy pewnym λ>0 jest estymator grzbietowy „ridge" postaci $(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

W gaussowskim modelu liniowym $(\mathbf{Y},\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ estymator uzyskany $\hat{\boldsymbol{\beta}}$ ma rozkład $N_p\left(\boldsymbol{\beta},\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right)$, co umożliwia konstrukcję testów istotności dla współczynników regresji (testy t-Studenta i Walda) oraz konstrukcję przedziałów ufności dla regresji (wartości średniej) i prognozy (przyszłej wartości) $Y_{\mathrm{progn}}(\mathbf{x}_0)$ przy warunku $\mathbf{x}=\mathbf{x}_0$. I tak

$$\hat{\beta}_i \sim N(\beta_i,\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{ii}) \Leftrightarrow \frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{ii}}} \sim N(0,1)$$

gdzie $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{ii}$ oznacza i-ty element diagonalny macierzy $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$,

$$\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2}=\frac{n-p}{\sigma^2}\hat{\sigma}^2=(\tfrac{1}{\sigma}\boldsymbol{\varepsilon})^T(\mathbf{I}-\mathbf{H})(\tfrac{1}{\sigma}\boldsymbol{\varepsilon})\sim\chi^2_{n-p}$$

Ponadto zmienne losowe

$$\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{ii}}} \qquad \text{i} \qquad \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2}=\frac{n-p}{\sigma^2}\hat{\sigma}^2=(\tfrac{1}{\sigma}\boldsymbol{\varepsilon})^T(\mathbf{I}-\mathbf{H})(\tfrac{1}{\sigma}\boldsymbol{\varepsilon})$$

są niezależne bo $\frac{1}{\sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\frac{1}{\sigma}\boldsymbol{\varepsilon})$ a macierze $\mathbf{B}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ i $\mathbf{A}=(\mathbf{I}-\mathbf{H})$ spełniają warunek $\mathbf{BA}=\mathbf{0}$,(zob. dodatek Formy liniowe i kwadratowe wektorów normalnych).

Wobec tego $\frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{ii}}}=\frac{\hat{\beta}_i-\beta_i}{SE_{\hat{\beta}_i}}\sim t_{n-p}$. Stąd natychmiast można podać konstrukcję testu istotności dla parametru β_i jak i przedziału ufności dla tego parametru.

Test Walda może być użyty do testowania hipotezy o jednoczesnym znikaniu kilku współczynników. Podzielmy wektor $\boldsymbol{\beta}$ na dwie składowe $\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}$ odpowiednio o wymiarach p_1 i p_2 i rozważmy hipotezę zerową H_0: $\boldsymbol{\beta}_2=\mathbf{0}$. Statystyka Walda jest formą kwadratową

$$W=\hat{\boldsymbol{\beta}}_2^T\mathbf{V}^{-1}(\hat{\boldsymbol{\beta}}_2)\hat{\boldsymbol{\beta}}_2,$$

gdzie $\hat{\boldsymbol{\beta}}_2$ jest estymatorem NW $\boldsymbol{\beta}_2$ a $\mathbf{V}(\hat{\boldsymbol{\beta}}_2)$ jest macierzą kowariancyjną wektora $\hat{\boldsymbol{\beta}}_2$ (zależną od nieznanego parametru σ^2). Zastępując nieznany parametr σ^2 jego estymatorem $\hat{\sigma}^2=\frac{RSS}{n-p}$ otrzymujemy $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_2)$ i statystyka $\frac{\hat{\boldsymbol{\beta}}_2^T\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\beta}}_2)\hat{\boldsymbol{\beta}}_2}{p_2}$ ma rozkład $F_{p_2,n-p}$, co jest podstawą konstrukcji testu z prawostronnym obszarem krytycznym.

Rozważane testy są szczególnym przypadkiem ogólnego testu hipotezy liniowej.

Hipotezy liniowe

Model: $\mathbf{Y}=\boldsymbol{\varphi}+\boldsymbol{\varepsilon}$

gdzie $\mathbf{Y} \in R^n$ jest wektorem obserwacji

$\boldsymbol{\varphi}\in\Omega\subset R^n$ jest wektorem średnich, o którym wiadomo, że należy do pewnej **właściwej podprzestrzeni liniowej Ω przestrzeni R^n**, tzn. $\Omega\subset R^n$ i $\dim(\Omega)=p<n$

$\boldsymbol{\varepsilon}\in R^n$ jest losowym wektorem błędów o rozkładzie $\boldsymbol{\varepsilon}\sim N_n(\mathbf{0},\sigma^2\mathbf{I})$

Uwaga: $\mathbf{Y}\sim N_n(\boldsymbol{\varphi},\sigma^2\mathbf{I})$

Interesuje nas problem testowania hipotezy

H_0: $\boldsymbol{\varphi}\in\omega\subset\Omega$ przeciwko H_1: $\boldsymbol{\varphi}\in\Omega-\omega$,

gdzie ω jest pewną **właściwą podprzestrzenią liniową** przestrzeni Ω,

czyli $\omega\subset\Omega$ i $\dim(\omega)=p-r<\dim(\Omega)$.

Uwaga. Zbiory ω i Ω są **podprzestrzeniami liniowymi**, czyli nie mogą to być zbiory ograniczone np. kule w R^n.

Po wyborze parametryzacji (bazy w przestrzeni R^n) rozważany model liniowy można zapisać w postaci gaussowskiego modelu liniowego

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I})$$

X jest tu nielosową macierzą wymiaru (n,p) zwaną **macierzą planu eksperymentu** o kolumnach $\mathbf{x}_i$, $i=0,...,p-1$ przy czym $\mathbf{x}_0=\mathbf{1}$, liniowo niezależnych tzn. $rz(\mathbf{X})=p$.

Wektor $\boldsymbol{\varphi}=E(\mathbf{Y})\in\Omega\subset R^n$ przy czym $\Omega=\text{span}\{\mathbf{x}_i;\ i=0,...,p-1\}=\text{Im}(\mathbf{X})$

Testować będziemy hipotezę liniową

H_0: $\mathbf{G}\boldsymbol{\beta}=\mathbf{0}$ przeciwko H_1: $\mathbf{G}\boldsymbol{\beta}\neq\mathbf{0}$,

gdzie **G** jest macierzą (r,p) $r<p$ pełnego rzędu tzn. $rz(\mathbf{G})=r$, która specyfikuje podprzestrzeń $\omega=\text{Im}(\mathbf{X}_{|Ker\mathbf{G}})$ ($\dim(\omega)=p-r$) będącą obrazem obcięcia przekształcenia **X** do jądra przekształcenia **G**.

Test hipotezy liniowej oparty na ilorazie wiarygodności (wariant I)

Na początku skonstruujemy estymator największej wiarygodności parametrów (φ,σ^2) w modelu liniowym

$$\mathbf{Y}=\varphi+\boldsymbol{\varepsilon},$$

gdzie $\varphi\in M\subset R^n$ i $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I})$ a ***M*** jest podprzestrzenią liniową przestrzeni R^n.

Funkcja wiarygodności jest postaci

$$L(\varphi,\sigma^2;\mathbf{Y})=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\|\mathbf{Y}-\varphi\|^2}$$

a jej logarytm

$$l(\varphi,\sigma^2;\mathbf{Y}) = \ln L = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\left\|\mathbf{Y}-\varphi\right\|^2 .$$

Oznaczmy przez P_M projekcję ortogonalną na podprzestrzeń $M \subset R^n$. Tym samym symbolem będziemy oznaczać reprezentację macierzową tej projekcji (rzutowania). Macierz P_M jest idempotentna ($P_M^2 = P_M$ - to gwarantuje, że jest to macierz rzutowania) i symetryczna ($P_M^T = P_M$ - ten warunek zapewnia, że projekcja jest ortogonalna). Stąd

$l(\varphi,\sigma^2;\mathbf{Y}) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\left\|\mathbf{Y}-P_M\mathbf{Y}+P_M\mathbf{Y}-\varphi\right\|^2$, ale

$\left\|\mathbf{Y}-P_M\mathbf{Y}+P_M\mathbf{Y}-\varphi\right\|^2 = \left\langle \mathbf{Y}-P_M\mathbf{Y}+P_M\mathbf{Y}-\varphi, \mathbf{Y}-P_M\mathbf{Y}+P_M\mathbf{Y}-\varphi\right\rangle = \left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2 + \left\|P_M\mathbf{Y}-\varphi\right\|^2$

bo $\left\langle \mathbf{Y}-P_M\mathbf{Y}, P_M\mathbf{Y}-\varphi\right\rangle = 0$ gdyż $\mathbf{Y}-P_M\mathbf{Y} \in M^{\perp}$ a $P_M\mathbf{Y}-\varphi \in M$.

Wobec tego $l(\varphi,\sigma^2;\mathbf{Y}) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2 - \tfrac{1}{2\sigma^2}\left\|P_M\mathbf{Y}-\varphi\right\|^2$

W powyższej funkcji wektor φ występuje tylko w jednym miejscu, więc łatwo widać, że $\forall\varphi$ $l(\varphi,\sigma^2;\mathbf{Y}) \le -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2$ i równość występuje tylko dla $\varphi = P_M\mathbf{Y}$ więc $\hat{\varphi} = P_M\mathbf{Y} = ENW[\varphi]$. $\hat{\sigma}^2 = \arg\max l(\hat{\varphi},\sigma^2;\mathbf{Y}) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2$ wyznaczamy szukając ekstremum funkcji różniczkowalnej jednej zmiennej. Oczywiście $\hat{\sigma}^2 = \tfrac{1}{n}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2$.

Wykażemy, że rzeczywiście $(\hat{\varphi},\hat{\sigma}^2) = (\mathbf{Y}-P_M\mathbf{Y}, \tfrac{1}{n}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2) = ENW(\varphi,\sigma^2)$. Rozważmy dowolny inny estymator $(\tilde{\varphi},\tilde{\sigma}^2)$

$$\begin{aligned} l(\hat{\varphi},\hat{\sigma}^2;\mathbf{Y}) - l(\tilde{\varphi},\tilde{\sigma}^2;\mathbf{Y}) &= -\tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln\hat{\sigma}^2 - \tfrac{1}{2\hat{\sigma}^2}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2 \\ &\quad - (-\tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln\tilde{\sigma}^2) - \tfrac{1}{2\tilde{\sigma}^2}\left\|\mathbf{Y}-P_M\mathbf{Y}\right\|^2 - \tfrac{1}{2\tilde{\sigma}^2}\left\|P_M\mathbf{Y}-\varphi\right\|^2) \\ &= \tfrac{n}{2}(-\ln\tfrac{\hat{\sigma}^2}{\tilde{\sigma}^2} - 1 + \tfrac{\hat{\sigma}^2}{\tilde{\sigma}^2}) + \tfrac{1}{2\tilde{\sigma}^2}\left\|P_M\mathbf{Y}-\varphi\right\|^2 \end{aligned}$$

Z nierówności $\ln x \le x-1$, która jest konsekwencją wklęsłości funkcji $\ln x$ otrzymujemy nierówność $-\ln x - 1 + x \ge 0$, która jest prawdziwa dla wszystkich $x > 0$, a równość zachodzi jedynie dla $x = 1$.

Widać wiec że $l(\hat{\varphi},\hat{\sigma}^2;\mathbf{Y}) - l(\varphi,\sigma^2;\mathbf{Y})$ jest sumą dwóch nieujemnych składników $\tfrac{n}{2}(-\ln\tfrac{\hat{\sigma}^2}{\sigma^2} - 1 + \tfrac{\hat{\sigma}^2}{\sigma^2})$ i $\tfrac{1}{2\sigma^2}\left\|P_M\mathbf{Y}-\varphi\right\|^2$ więc $l(\hat{\varphi},\hat{\sigma}^2;\mathbf{Y}) - l(\varphi,\sigma^2;\mathbf{Y}) \ge 0$, a wartość 0 różnica ta osiąga, gdy oba składniki równocześnie się zerują, czyli $\varphi = P_M\mathbf{Y}$ a $\tfrac{\hat{\sigma}^2}{\tilde{\sigma}^2} = 1$.

Konstrukcja testu hipotezy liniowej oparta na ilorazie wiarygodności

$$C=\{\mathbf{Y}:\lambda(\mathbf{Y})=\frac{L(\hat{\varphi},\hat{\sigma}^2;\mathbf{Y})}{L(\dot{\varphi},\dot{\sigma}^2;\mathbf{Y})}>c_1\}=\{\mathbf{Y}:\left(\frac{\dot{\sigma}^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}}>c_1\}=\{\mathbf{Y}:\frac{\dot{\sigma}^2}{\hat{\sigma}^2}>c_2\}=\{\mathbf{Y}:\frac{n-p}{r}\frac{\dot{\sigma}^2-\hat{\sigma}^2}{\hat{\sigma}^2}>c\}$$

Oznaczając

$RSS_\omega=R_1=\|\mathbf{Y}-P_\omega\mathbf{Y}\|^2=\|\dot{\boldsymbol{\varepsilon}}\|^2$ resztkowa suma kwadratów przy prawdziwości H_0

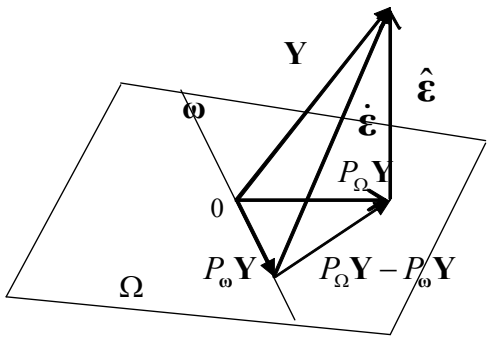
(całkowita suma kwadratów SS_{total})

$RSS_\Omega=R_0=\|\mathbf{Y}-P_\Omega\mathbf{Y}\|^2=\|\hat{\boldsymbol{\varepsilon}}\|^2$ resztkowa suma kwadratów w modelu bez ograniczeń

(w problemie *k*-prób suma kwadratów wewnątrz grup SS_{within})

otrzymujemy $C=\{\mathbf{Y}:F=\frac{n-p}{r}\frac{RSS_\omega-RSS_\Omega}{RSS_\Omega}>c\}$.

Interpretacja geometryczna



$RSS_\omega-RSS_\Omega=\|\dot{\boldsymbol{\varepsilon}}\|^2-\|\hat{\boldsymbol{\varepsilon}}\|^2=\|P_\Omega\mathbf{Y}-P_{\boldsymbol{\omega}}\mathbf{Y}\|^2=\|P_{\boldsymbol{\omega}^{\perp\Omega}}\mathbf{Y}\|^2$ (w problemie *k*-prób suma kwadratów pomiędzy grupami (modelami) $SS_{between}$)

Przedstawmy $\Omega=\omega\oplus\omega^{\perp\Omega}$ a następnie $R^n=\omega\oplus\omega^{\perp\Omega}\oplus\Omega^\perp$, gdzie $\omega^{\perp\Omega}$ oznacza dopełnienie ortogonalne ω do Ω.

Przekształcenie (oraz macierz reprezentująca to przekształcenie) $P_\Omega-P_{\boldsymbol{\omega}}=P_{\boldsymbol{\omega}^{\perp\Omega}}$ jest projekcją na dopełnienie ortogonalne przestrzeni ω do przestrzeni Ω. Wektor $\frac{1}{\sigma}\boldsymbol{\varepsilon}$ ma rozkład $N_n(\mathbf{0},\mathbf{I})$,więc przy prawdziwości H_0:$\boldsymbol{\varphi}\in\omega$ wektor $\frac{1}{\sigma}(P_\Omega-P_{\boldsymbol{\omega}})\mathbf{Y}=\frac{1}{\sigma}P_{\boldsymbol{\omega}^{\perp\Omega}}\mathbf{Y}=\frac{1}{\sigma}P_{\boldsymbol{\omega}^{\perp\Omega}}(\boldsymbol{\varphi}+\boldsymbol{\varepsilon})=P_{\boldsymbol{\omega}^{\perp\Omega}}(\frac{1}{\sigma}\boldsymbol{\varepsilon})$ jest obrazem wektora $\frac{1}{\sigma}\boldsymbol{\varepsilon}\sim N_n(\mathbf{0},\mathbf{I})$ w projekcji $P_{\boldsymbol{\omega}^{\perp\Omega}}$ rzędu *r*.

Podobnie wektor $\frac{1}{\sigma}(\mathbf{I}-P_\Omega)\mathbf{Y}=\frac{1}{\sigma}P_{\Omega^\perp}\mathbf{Y}=\frac{1}{\sigma}P_{\Omega^\perp}(\boldsymbol{\varphi}+\boldsymbol{\varepsilon})=P_{\Omega^\perp}(\frac{1}{\sigma}\boldsymbol{\varepsilon})$ jest obrazem wektora $\frac{1}{\sigma}\boldsymbol{\varepsilon}\sim N_n(\mathbf{0},\mathbf{I})$ w projekcji $P_{\Omega^\perp}$ rzędu *n-p*. Z uwagi na fakt że złożenie (iloczyn macierzy) $P_{\Omega^\perp}$ $P_{\boldsymbol{\omega}^{\perp\Omega}}$ jest przekształceniem zerowym wektory $\frac{1}{\sigma}(P_\Omega-P_{\boldsymbol{\omega}})\mathbf{Y}$ i $\frac{1}{\sigma}(\mathbf{I}-P_\Omega)\mathbf{Y}$ są niezależne a formy kwadratowe

$\left\|\frac{1}{\sigma}(P_\Omega - P_\omega)\mathbf{Y}\right\|^2$ i $\left\|\frac{1}{\sigma}(\mathbf{I} - P_\Omega)\mathbf{Y}\right\|^2$ mają przy prawdziwości H_0 rozkłady odpowiednio χ^2_r i χ^2_{n-p} (zob. uzupełnienie- Formy liniowe i kwadratowe wektorów normalnych), więc

statystyka $$F = \frac{n-p}{r}\frac{RSS_\omega - RSS_\Omega}{RSS_\Omega} = \frac{n-p}{r}\frac{\left\|(P_\Omega - P_\omega)\mathbf{Y}\right\|^2}{\left\|\mathbf{Y} - P_\Omega\mathbf{Y}\right\|^2}$$

ma przy prawdziwości H_0 rozkład Snedecora $F_{r,n-p}$.

Uwaga. W przypadku, gdy nie jest prawdziwa hipoteza H_0: $\varphi \in \omega$, wektor $\frac{1}{\sigma}P_{\omega^{\perp\Omega}}\mathbf{Y}$ ma rozkład $N_n(\frac{1}{\sigma}P_{\omega^{\perp\Omega}}\varphi, \mathbf{I})$ a forma kwadratowa $\left\|\frac{1}{\sigma}P_{\omega^{\perp\Omega}}\mathbf{Y}\right\|^2$ ma niecentralny rozkład $\chi^2_{r,\delta}$ z parametrem niecentralności $\delta = \left\|\frac{1}{\sigma}P_{\omega^{\perp\Omega}}\varphi\right\|$. Fakt ten umożliwia wyznaczenie tzw. obserwowanej mocy testu hipotezy liniowej

Podsumowaniem jest następująca tabelka ANOVA

Suma kwadratów SS	Stopnie swobody df	Średni kwadrat MS	Iloraz F
$RSS_\omega - RSS_\Omega$	r	$\frac{1}{r}(RSS_\omega - RSS_\Omega)$	$F = \frac{n-p}{r}\frac{RSS_\omega - RSS_\Omega}{RSS_\Omega}$
RSS_Ω	n-p	$\frac{1}{n-p}\ RSS_\Omega$	
RSS_ω	n-p+r		

Dodatek- Formy liniowe i kwadratowe wektorów normalnych

- Niech $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie wektorem losowym o rozkładzie $N(\mathbf{0}, \mathbf{I})$ i niech $\mathbf{A}$ będzie macierzą $n \times n$ symetryczną, idempotentną rzędu r. Wówczas forma kwadratowa $\mathbf{X}'\mathbf{A}\mathbf{X}$ ma rozkład χ^2_r.
- Niech $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie wektorem losowym o rozkładzie $N(\mathbf{0}, \mathbf{I})$, $\mathbf{A}$ będzie macierzą $n \times n$ symetryczną, idempotentną rzędu r a $\mathbf{B}$ dowolną macierzą $m \times n$ taką, że $\mathbf{BA} = \mathbf{0}$. Wówczas funkcja liniowa (wektor losowy) $\mathbf{BX}$ ma rozkład niezależny od rozkładu formy kwadratowej $\mathbf{X}'\mathbf{A}\mathbf{X} = \langle \mathbf{X}, \mathbf{A}\mathbf{X} \rangle$.
- Niech $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie wektorem losowym o rozkładzie $N(\mathbf{0}, \mathbf{I})$, $\mathbf{A}$ będzie macierzą $n \times n$ symetryczną, idempotentną rzędu r a $\mathbf{B}$ macierzą $n \times n$ symetryczną, idempotentną rzędu $s \le n - r$. Załóżmy ponadto, że $\mathbf{BA} = \mathbf{0}$. Wówczas forma kwadratowa $\mathbf{X}'\mathbf{A}\mathbf{X} = \langle \mathbf{X}, \mathbf{A}\mathbf{X} \rangle$ ma rozkład niezależny od rozkładu formy kwadratowej $\mathbf{X}'\mathbf{B}\mathbf{X} = \langle \mathbf{X}, \mathbf{B}\mathbf{X} \rangle$.

Dowody powyższych faktów

- Niech $\mathbf{X}=\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie wektorem losowym o rozkładzie $N(\mathbf{0},\mathbf{I})$ i niech $\mathbf{A}$ będzie macierzą $n\times n$ symetryczną, idempotentną rzędu *r*. Wówczas forma kwadratowa $\mathbf{X}'\mathbf{A}\mathbf{X}$ ma rozkład χ_r^2.

Macierz $\mathbf{A}$ jako macierz symetryczna ma rzeczywiste wartości własne a jako idempotentna ma wartości własne równe 0 i 1 przy czym 1 jest *r*-krotną wartoscią własną a 0 *n-r* krotną wartością własną. Wektory własne odpowiadające różnym wartościom własnym macierzy symetrycznej są ortogonalne. Podprzestrzeń własna odpowiadająca wartości własnej 1 jest wymiaru *r* istnieje więc *r* liniowo niezależnych wektorów własnych odpowiadających wartości własnej 1. Przeprowadzając procedurę ortonormalizacji Grama Schmidta można wybrać w przestrzeni własnej $X_{\lambda=1}$ bazę ortonormalną. To samo dotyczy przestrzeni własnej $X_{\lambda=0}$. Reasumując z kolejnych ortonormalnych wektorów własnych przestrzeni $X_{\lambda=1}$ i $X_{\lambda=0}$ tworzymy nową bazę przestrzeni R^n i tworzymy macierz ortogonalną $\mathbf{C}$, której kolumny są wektorami zbudowanej nowej bazy ortonormalnej. Wówczas

$$\mathbf{C}'\mathbf{A}\mathbf{C}=\mathbf{\Lambda}=\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Niech $\mathbf{Y}=\mathbf{C}'\mathbf{X}\Leftrightarrow\mathbf{X}=\mathbf{C}\mathbf{Y}$. Oczywiście $\mathbf{Y}=\mathbf{C}'\mathbf{X}\sim N(\mathbf{0},\mathbf{I})$.

Stąd $\mathbf{X}'\mathbf{A}\mathbf{X}=\mathbf{Y}'\mathbf{C}'\mathbf{A}\mathbf{C}\mathbf{Y}=\mathbf{Y}'\mathbf{\Lambda}\mathbf{Y}$. Zapiszmy wektor $\mathbf{Y}=\mathbf{C}'\mathbf{X}$ w postaci $\mathbf{Y}=\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}$, gdzie

$$\mathbf{Y}_1=\begin{bmatrix} Y_1 \\ \vdots \\ Y_r \end{bmatrix}\ \mathbf{Y}_2=\begin{bmatrix} Y_{r+1} \\ \vdots \\ Y_n \end{bmatrix}.$$

$$\mathbf{X}'\mathbf{A}\mathbf{X}=\mathbf{Y}'\mathbf{C}'\mathbf{A}\mathbf{C}\mathbf{Y}=\mathbf{Y}_1'\mathbf{Y}_1=\sum_{i=1}^{r}Y_i^2\sim\chi_r^2.$$

- Niech $\mathbf{X}=\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie wektorem losowym o rozkładzie $N(\mathbf{0},\mathbf{I})$, $\mathbf{A}$ będzie macierzą $n\times n$ symetryczną, idempotentną rzędu *r* a $\mathbf{B}$ dowolną macierzą $m\times n$ taką, że $\mathbf{B}\mathbf{A}=\mathbf{0}$. Wówczas funkcja liniowa $\mathbf{B}\mathbf{X}$ ma rozkład niezależny od rozkładu formy kwadratowej $\mathbf{X}'\mathbf{A}\mathbf{X}=\langle\mathbf{X},\mathbf{A}\mathbf{X}\rangle$.

Niech $\mathbf{C}$ będzie macierzą ortogonalną sprowadzającą macierz $\mathbf{A}$ do postaci diagonalnej

$\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{\Lambda} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. Przy oznaczeniach z poprzedniego punku mamy

$$\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{Y}'\mathbf{C}'\mathbf{A}\mathbf{C}\mathbf{Y} = \mathbf{Y}'\mathbf{\Lambda}\mathbf{Y} = \mathbf{Y}_1'\mathbf{Y}_1 = \sum_{i=1}^{r} Y_i^2 \sim \chi_r^2$$

Niech $\mathbf{F} = \mathbf{B}\mathbf{C}$ .Wówczas $\mathbf{F}\mathbf{\Lambda} = \mathbf{B}\mathbf{C}\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{B}\mathbf{A}\mathbf{C} = \mathbf{0}$ bo $\mathbf{B}\mathbf{A} = \mathbf{0}$. Dokonując podziału macierzy $\mathbf{F} = \mathbf{B}\mathbf{C}$ wymiaru $m \times n$ na bloki $\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 \end{bmatrix}$, gdzie $\mathbf{F}_1$ jest macierzą $m \times r$ a $\mathbf{F}_2$ macierzą $m \times (n-r)$ otrzymujemy $\mathbf{F}\mathbf{\Lambda} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 \end{bmatrix} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{0} \end{bmatrix} = \mathbf{0}$, skąd $\mathbf{F}_1 = \mathbf{0}$ czyli $\mathbf{F} = \begin{bmatrix} \mathbf{0} & \mathbf{F}_2 \end{bmatrix}$. Funkcję liniową $\mathbf{B}\mathbf{X}$ zapisujemy w postaci

$$\mathbf{B}\mathbf{X} = \mathbf{B}\mathbf{C}\mathbf{C}'\mathbf{X} = \mathbf{F}\mathbf{Y} = \begin{bmatrix} \mathbf{0} & \mathbf{F}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \mathbf{F}_2\mathbf{Y}_2 .$$

Wyraziliśmy funkcję liniową $\mathbf{B}\mathbf{X}$ jako funkcję liniową wektora $\mathbf{Y}_2$ niezależnego od wektora $\mathbf{Y}_1$. Stąd $\mathbf{B}\mathbf{X}$ i $\mathbf{X}'\mathbf{A}\mathbf{X}$ są niezależne. Oczywiście $\mathbf{B}\mathbf{X} \sim N(\mathbf{0}, \mathbf{F}_2\mathbf{F}_2')$.

- Niech $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ będzie wektorem losowym o rozkładzie $N(\mathbf{0}, \mathbf{I})$, $\mathbf{A}$ będzie macierzą $n \times n$ symetryczną, idempotentną rzędu *r* a $\mathbf{B}$ macierzą $n \times n$ symetryczną, idempotentną rzędu $s \le n - r$. Załóżmy ponadto, że $\mathbf{B}\mathbf{A} = \mathbf{0}$.Wówczas forma kwadratowa $\mathbf{X}'\mathbf{A}\mathbf{X} = \langle \mathbf{X}, \mathbf{A}\mathbf{X} \rangle$ ma rozkład niezależny od rozkładu formy kwadratowej $\mathbf{X}'\mathbf{B}\mathbf{X} = \langle \mathbf{X}, \mathbf{B}\mathbf{X} \rangle$.

Niech $\mathbf{C}$ będzie macierzą ortogonalną sprowadzającą macierz $\mathbf{A}$ do postaci diagonalnej $\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{\Lambda} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. Przy oznaczeniach z poprzedniego punku mamy

$$\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{Y}'\mathbf{C}'\mathbf{A}\mathbf{C}\mathbf{Y} = \mathbf{Y}_1'\mathbf{Y}_1 = \sum_{i=1}^{r} Y_i^2$$

Niech $\mathbf{G} = \mathbf{C}'\mathbf{B}\mathbf{C}$. Oczywiście $\mathbf{G}$ jest macierzą symetryczną i idempotentną oraz $\mathbf{G}\mathbf{\Lambda} = \mathbf{C}'\mathbf{B}\mathbf{C}\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{C}'\mathbf{B}\mathbf{A}\mathbf{C} = \mathbf{0}$ gdyż $\mathbf{B}\mathbf{A} = \mathbf{0}$. Zapisując macierz $\mathbf{G}$ w postaci blokowej $\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 \\ \mathbf{G}_2' & \mathbf{G}_3 \end{bmatrix}$, gdzie $\mathbf{G}_1$ ma wymiary $r \times r$ a $\mathbf{G}_3$ ma wymiary $(n-r) \times (n-r)$ otrzymujemy $\mathbf{G}\mathbf{\Lambda} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 \\ \mathbf{G}_2' & \mathbf{G}_3 \end{bmatrix} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{0} \\ \mathbf{G}_2' & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ z czego wynika, że $\mathbf{G}_1 = \mathbf{0}$ i $\mathbf{G}_2' = \mathbf{0}$ więc także $\mathbf{G}_2 = \mathbf{0}$. Wobec powyższego

$$\mathbf{X}'\mathbf{B}\mathbf{X} = \mathbf{X}'\mathbf{C}\mathbf{C}'\mathbf{B}\mathbf{C}\mathbf{C}'\mathbf{X} = \mathbf{Y}'\mathbf{G}\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1' & \mathbf{Y}_2' \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_3 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \mathbf{Y}_2'\mathbf{G}_3\mathbf{Y}_2$$ jest formą kwadratową ostatnich *n-r* współrzędnych wektora $\mathbf{Y} = \mathbf{C}'\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$,wiec niezależną od formy $\mathbf{X}'\mathbf{A}\mathbf{X}$ zależnej od pierwszych *r* współrzędnych wektora $\mathbf{Y} = \mathbf{C}'\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$.

Zauważmy, że $\mathbf{G}^2 = \mathbf{C}'\mathbf{B}\mathbf{C}\mathbf{C}'\mathbf{B}\mathbf{C} = \mathbf{C}'\mathbf{B}\mathbf{B}\mathbf{C} = \mathbf{C}'\mathbf{B}\mathbf{C} = \mathbf{G}$ i $\mathbf{G} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_3 \end{bmatrix}$ stąd $\mathbf{G}_3^2 = \mathbf{G}_3$ i $\operatorname{rank}(\mathbf{G}) = \operatorname{rank}(\mathbf{G}_3) = s$ $(s \le n - r)$ więc $\mathbf{X}'\mathbf{B}\mathbf{X} \sim \chi_s^2$.

Odległość Mahalanobisa

Wersja populacyjna. Rozważmy wektor losowy $\mathbf{X}=\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$ o rozkładzie Q, o wektorze wartości oczekiwanych $\mathbf{m}=\begin{bmatrix} m_1 \\ \vdots \\ m_k \end{bmatrix}$ i macierzy kowariancyjnej $\mathbf{\Sigma}=\begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_k^2 \end{bmatrix}$.

Odległość Mahalanobisa punktu $\mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$ od rozkładu Q (od jego centrum) jest zdefiniowana wzorem $d_M(\mathbf{x},Q)=\sqrt{(\mathbf{x}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}$.

Odległość ta jest wielowymiarowym uogólnieniem idei mierzenia odległości punktu od „środka" rozkładu w jednostkach odchylenia standardowego z uwzględnieniem skorelowania zmiennych.

Mając dane dwa punkty $\mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$ i $\mathbf{y}=\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}$ ich odległość względem rozkładu Q określamy wzorem $d_M(\mathbf{x},\mathbf{y},Q)=\sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})}$.

Jeśli $\mathbf{X}=\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}\sim N(\mathbf{m},\mathbf{\Sigma})$, to $f_k(\mathbf{x})=\frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}=\frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}}e^{-\frac{1}{2}d_M^2(\mathbf{x},Q)}$.

Warstwice funkcji gęstości rozkładu normalnego są izoliniami odległości Mahalanobisa od wektora wartości oczekiwanych $\mathbf{m}$.

Wersja próbkowa. Jeśli dysponujemy próbą prostą z rozkładu Q możemy zdefiniować próbkową wersję odległości Mahalanobisa $d_M(\mathbf{x},Q)=\sqrt{(\mathbf{x}-\bar{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}})}$ i odpowiednio $d_M(\mathbf{x},\mathbf{y},Q)=\sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$, gdzie $\mathbf{S}=\frac{1}{n-1}\left[\sum_{s=1}^{n}(X_{is}-\bar{X}_i)(X_{js}-\bar{X}_j)\right]$ jest próbkową macierzą kowariancji.

Z uwagi na możliwość dekomopozycji macierzy $\mathbf{\Sigma}^{-1}=\mathbf{W}^T\mathbf{W}$ (dekompozycja Choleskiego lub użycie pierwiastka symetrycznego –z tw. spektralnego) można uważać odległość Mahalanobisa jako normę euklidesową $d_M(\mathbf{x},\mathbf{y},Q)=\sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})}=\|\mathbf{W}(\mathbf{x}-\mathbf{y})\|$ wektora $\mathbf{x}-\mathbf{y}$ poddanego transformacji wybielającej $\mathbf{W}(\mathbf{x}-\mathbf{y})$ tzn. usuwającej skorelowanie współrzędnych. Rzeczywiście

$$\mathbf{\Sigma}=\mathbf{W}^{-1}(\mathbf{W}^T)^{-1} \text{ a } V(\mathbf{WX})=\mathbf{W}V(\mathbf{X})\mathbf{W}^T=\mathbf{W}\mathbf{W}^{-1}(\mathbf{W}^T)^{-1}\mathbf{W}^T=\mathbf{I}.$$

Liniowy model statystyczny w postaci scentrowanej

Liniowy model statystyczny $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$, i=1,2,...,n, (k=p-1=rank $\mathbf{X}$) zapiszmy w postaci wektorowo-macierzowej

1. $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
3. $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

gdzie $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$, $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$.

Model ten przedstawia losowy wektor $\mathbf{Y}$, którego wartość oczekiwana $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ jest liniową kombinacją kolumn nielosowej macierzy $\mathbf{X}$, zwanej macierzą planu eksperymentu. Nieznane są parametry β_0,...,β_k (współrzędne wektora $\boldsymbol{\beta}$). Losowy wektor zakłóceń $\boldsymbol{\varepsilon}$ ma zerową wartość oczekiwaną a jego poszczególne składowe są nieskorelowane i mają wspólną (nieznaną) wariancję σ^2. Powyższy model można zapisać w postaci

$$Y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \varepsilon_i, \quad i=1,2,...,n,$$

gdzie $\alpha = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_k\bar{x}_k$ i $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, j=1,…,k.

Przyjmijmy następujące oznaczenia

$$\mathbf{j} = \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{J} = \mathbf{j}\mathbf{j}^T = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}, \quad \mathbf{X}_1 = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \mathbf{X} = [\mathbf{j}, \mathbf{X}_1].$$

Macierz $\frac{1}{n}\mathbf{J} = \frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$ jest macierzą symetryczną i idempotentną reprezentującą rzut ortogonalny na 1-wymiarową przestrzeń $\text{span}\{\mathbf{1}\} = \text{span}\{\mathbf{j}\}$. Natomiast macierz $\mathbf{I} - \frac{1}{n}\mathbf{J}$ symetryczna i idempotentna zwana macierzą centrującą reprezentuje rzut ortogonalny na $\text{span}\{\mathbf{1}\}^{\perp} = \text{span}\{\mathbf{j}\}^{\perp}$.

Rzeczywiście $(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{X}_1 = (\mathbf{I} - \frac{1}{n}\mathbf{J})\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix} = \mathbf{X}_c$ prowadzi do scentrowanej formy macierz zmiennych objaśniających.

Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ w scentrowanej formie może być zapisany w postaci

$$\mathbf{Y} = [\mathbf{j}\ \mathbf{X}_c]\begin{bmatrix}\alpha\\ \boldsymbol{\beta}_1\end{bmatrix} + \boldsymbol{\varepsilon}\,.$$

Układ równań normalnych dla tego modelu jest postaci

$$[\mathbf{j}\ \mathbf{X}_c]^T[\mathbf{j}\ \mathbf{X}_c]\begin{bmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}_1\end{bmatrix} = [\mathbf{j}\ \mathbf{X}_c]^T\mathbf{Y} \Leftrightarrow \begin{bmatrix}\mathbf{j}^T\\ \mathbf{X}_c^T\end{bmatrix}[\mathbf{j}\ \mathbf{X}_c]\begin{bmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}_1\end{bmatrix} = \begin{bmatrix}\mathbf{j}^T\\ \mathbf{X}_c^T\end{bmatrix}\mathbf{Y} \Leftrightarrow$$

$$\Leftrightarrow \begin{bmatrix}\mathbf{j}^T\mathbf{j} & \mathbf{j}^T\mathbf{X}_c\\ \mathbf{X}_c^T\mathbf{j}^T & \mathbf{X}_c^T\mathbf{X}_c\end{bmatrix}\begin{bmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}_1\end{bmatrix} = \begin{bmatrix}n\overline{Y}\\ \mathbf{X}_c^T\mathbf{Y}\end{bmatrix} \Leftrightarrow \begin{bmatrix}n & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_c^T\mathbf{X}_c\end{bmatrix}\begin{bmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}_1\end{bmatrix} = \begin{bmatrix}n\overline{Y}\\ \mathbf{X}_c^T\mathbf{Y}\end{bmatrix}.$$

Stąd $\begin{bmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}_1\end{bmatrix} = \begin{bmatrix}n & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_c^T\mathbf{X}_c\end{bmatrix}^{-1}\begin{bmatrix}n\overline{Y}\\ \mathbf{X}_c^T\mathbf{Y}\end{bmatrix} = \begin{bmatrix}\frac{1}{n} & \mathbf{0}\\ \mathbf{0} & (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\end{bmatrix}\begin{bmatrix}n\overline{Y}\\ \mathbf{X}_c^T\mathbf{Y}\end{bmatrix} = \begin{bmatrix}\overline{Y}\\ (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{Y}\end{bmatrix}$

Wobec powyższego

$$\hat{\mathbf{Y}} = [\mathbf{j}\ \mathbf{X}_c]\begin{bmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}_1\end{bmatrix} = [\mathbf{j}\ \mathbf{X}_c]\begin{bmatrix}\overline{Y}\\ (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{Y}\end{bmatrix} = \mathbf{j}\overline{Y} + \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{Y} = (\tfrac{1}{n}\mathbf{J} + \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T)\mathbf{Y}$$

Porównując dwa wzory

$$\hat{\mathbf{Y}} = (\tfrac{1}{n}\mathbf{J} + \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T)\mathbf{Y} \text{ i } \hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

otrzymujemy drugą postać macierzy daszkowej $\mathbf{H} = \frac{1}{n}\mathbf{J} + \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T$.

Zauważmy, że $\mathbf{S} = \frac{1}{n-1}\left[\sum_{s=1}^{n}(X_{is} - \overline{X}_i)(X_{js} - \overline{X}_j)\right] = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$

Wyznaczmy *i*-ty element diagonalny macierzy daszkowej

$$h_i = H_{ii} = \frac{1}{n} + \begin{bmatrix}x_{i1} - \overline{x}_1 & \cdots & x_{ik} - \overline{x}_k\end{bmatrix}(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\begin{bmatrix}x_{i1} - \overline{x}_1\\ \vdots\\ x_{ik} - \overline{x}_k\end{bmatrix} =$$

$$\frac{1}{n} + \begin{bmatrix}x_{i1} - \overline{x}_1 & \cdots & x_{ik} - \overline{x}_k\end{bmatrix}[(n-1)\mathbf{S}]^{-1}\begin{bmatrix}x_{i1} - \overline{x}_1\\ \vdots\\ x_{ik} - \overline{x}_k\end{bmatrix} = \frac{1}{n} + \frac{1}{n-1}d_M^2(\mathbf{x}_i, \overline{\mathbf{x}})$$

gdzie $\mathbf{x}_i = \begin{bmatrix}x_{i1} & \cdots & x_{ik}\end{bmatrix}$, jest *i*- tym wierszem macierzy $\mathbf{X}_1 = \begin{bmatrix}x_{11} & \cdots & x_{1k}\\ \vdots & \ddots & \vdots\\ x_{n1} & \cdots & x_{nk}\end{bmatrix}$ a $\overline{\mathbf{x}} = \begin{bmatrix}\overline{x}_1 & \cdots & \overline{x}_k\end{bmatrix}$.

Uzyskany związek $h_i = H_{ii} = \frac{1}{n} + \frac{1}{n-1}d_M^2(\mathbf{x}_i, \overline{\mathbf{x}})$ tłumaczy zastosowanie tzw. dźwigni (czyli elementów h_i) do identyfikacji obserwacji wpływowych.

Klasyczny model liniowy ma zastosowanie wówczas, gdy wartości zmiennych objaśnianych są ustalone w powtarzalnych próbach. W praktyce częściej mamy do czynienia z losowymi zmiennymi objaśniającymi. Sytuacja ta prowadzi do poważnych komplikacji zagadnienia estymacji i estymatory MNK tracą swoje cenne własności. Jest jednak pewien przypadek szczególny, który nie prowadzi do zbytnich komplikacji. Modyfikujemy klasyczny model liniowy zastępując założenie o nielosowości zmiennych objaśniających założeniem (Goldberger):

X jest macierzą losową o rozkładzie niezależnym od **ε**.

Założenie to wzięte wraz z pozostałymi (nie zmodyfikowanymi) założeniami modelu liniowego prowadzi do zależności

$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$

$E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$

$V(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}$

Wykorzystując fundamentalną własność iterowania wartości oczekiwanych możemy stwierdzić, że:

Jeżeli w klasycznym modelu regresji liniowej założenie o nielosowości zmiennych objaśniających zastąpimy założeniem, że X ma rozkład niezależny od ε, to estymatory rozważanych parametrów uzyskane MNK są nieobciążone.

Przy losowości zmiennych objaśniających estymator MNK $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ nie jest liniowy, gdyż jest on teraz stochastyczną funkcją wektora **Y** i jako taki nie może być najlepszym liniowym estymatorem. Uwagę tę można pominąć, jeżeli zmienne objaśniające są zmiennymi losowymi o rozkładzie niezależnym od składników losowych, ponieważ wówczas uzyskane estymatory mają własności estymatorów uzyskanych MNK. Można pokazać, że estymator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ jest identyczny z estymatorem największej wiarygodności przy założeniu, że składniki losowe mają rozkłady normalne, a rozkład **X** nie zależy od parametrów **β** i σ^2. Rozważając heurystycznie można stwierdzić, że stochastyczna niezależność składników losowych i zmiennych objaśniających zapewnia, że estymatory uzyskane MNK mają własności estymatorów uzyskanych klasyczną MNK pod warunkiem, że macierz **X** przyjmuje dowolną wartość, co jest równoważne bezwarunkowemu posiadaniu klasycznych własności dla każdej macierzy **X**. Podobnie zachowują swą ważność klasyczna estymacja przedziałowa i weryfikacja hipotez: na przykład przedział, który pokrywa prawdziwą wartość parametru z prawdopodobieństwem 1-α dla każdej ustalonej wartości **X**, będzie pokrywać prawdziwą wartość parametru z prawdopodobieństwem 1-α dla wszystkich możliwych wartości **X**. Osłabienie założenia niezależności zmiennych objaśniających od składnika losowego prowadzi do modeli które należy analizować oddzielnie, istotnie wykorzystując charakter tych odstępstw. Opisy różnych klas modeli uzyskiwanych poprzez osłabianie założenia niezależności można znaleźć w monografii:

Fuller W.,A.; Measurement Error Models, Wiley, 1987