2.3 Gauss-Markov theorem

REMIND: Data : $(Y_i, X_i^T)^T$, $i = 1, \dots, n$ not necessarily iid

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} = (X^0, \dots, X^{k-1})$$

It is assumed $Y|X \sim (X\beta, \sigma^2 I_n)$

$\operatorname{rank}(X_{n \times k}) = r \leq k < n$

$E(Y|X) = X\beta$

$\operatorname{var}(Y|X) = \sigma^2 I_n$

‖ This is assumed ≡ MODEL (might be wrong)

- if $r = k$, $\hat{\beta} = (X^TX)^{-1}X^TY$ is unbiased estimator of β
- $\hat{Y} = X\hat{\beta} = HY$ is projection of Y into $\mathcal{M}(X)$

$H = X(X^TX)^{-1}X^T$ (if $r = k$)

$= QQ^T$ $(r \leq k)$,

$= X(X^TX)^-X^T$, $Q = (q_1, \dots, q_r)$ = orthonormal basis of $\mathcal{M}(X)$

- Statistical properties of $\hat{Y}$?

→ Gauss-Markov theorem

C.F. Gauss : 1821-23

A.A. Markov : 1912

- more "elegant" version of the proof

Theorem 2.4 Gauss-Markov

Assume a linear model $Y|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = r \le k < n$. The vector of fitted values $\hat{Y}$ is, conditionally given X, the best linear unbiased estimator (BLUE) of a vector parameter $\mu := E(Y|X)$ $(= X\beta$ for some $\beta \in \mathbb{R}^k)$. Further,

$$\operatorname{var}(\hat{Y}|X) = \sigma^2 H.$$

Proof:

Linearity: $\hat{Y} = HY \equiv$ linear function of Y

Unbiasedness: $E(\hat{Y}|X) = E(HY|X) =$ (HY: fun. of X)

$= HE(Y|X) \overset{\text{MODEL}}{=} HX\beta = X\beta$

(H: projection of cols. of X into $\mathcal{M}(X)$)

Optimality: Let $\tilde{Y} = a + BY$ be some other LUE of $\mu = X\beta$, B = fun. of X

That is

$\forall \beta \in \mathbb{R}^k \quad E(\tilde{Y}|X) = X\beta$

$\forall \beta \in \mathbb{R}^k \quad E(a + BY|X) = X\beta$

$\forall \beta \in \mathbb{R}^k \quad a + BE(Y|X) = X\beta$

$\forall \beta \in \mathbb{R}^k \quad a + BX\beta = X\beta$

- Take $\beta = 0 \Rightarrow a = 0$ (to maintain LUE property)
- That is, $\forall \beta \in \mathbb{R}^k \quad BX\beta = X\beta$
- Take $\beta = (0, .., 1, .., 0)^T$: $(BX)^j = X^j$ (j-th place)

$\Rightarrow BX = X$ (to maintain LUE property)

That is, if $\tilde{\tilde{y}} = a + BX$ is LUE of $X\beta$

$\Longleftrightarrow a = 0, \quad BX = X$ (trivial)

$BX = X$ implies : $BX = X \quad / \cdot (X^TX)^- X^T$

$$\underbrace{BX(X^TX)^-X^T}_{H} = \underbrace{X(X^TX)^-X^T}_{H}$$

$$BH = H \quad (= H^T)$$

$$H^T B^T = H^T \quad (= H)$$

$$H B^T = H \quad (= H^T)$$

Now: • $\operatorname{var}(\tilde{\tilde{y}}|X) = \operatorname{var}(Hy|X) = H \underbrace{\operatorname{var}(y|X)}_{\sigma^2 I_n} H^T =$

$$= \sigma^2 HH^T = \sigma^2 H$$

• $\operatorname{var}(\tilde{y}|X) = \operatorname{var}(By|X) = B \underbrace{\operatorname{var}(y|X)}_{\sigma^2 I_n} B^T =$

$$= \sigma^2 BB^T = \sigma^2 (H + \underline{B-H})(H + \underline{B-H})^T =$$

$$= \sigma^2 \left(\underbrace{HH^T}_{H} + \underbrace{H(B-H)^T}_{0} + \underbrace{(B-H)H^T}_{0} + (B-H)(B-H)^T \right)$$

$$= \sigma^2 H + \sigma^2 (B-H)(B-H)^T$$

from previous page:

$$\operatorname{var}(\hat{\vec{y}}/X) = \sigma^2 H$$

$$\operatorname{var}(\tilde{\vec{y}}/X) = \sigma^2 H + \sigma^2 (B-H)(B-H)^T$$

$$\Rightarrow \operatorname{var}(\tilde{\vec{y}}/X) - \operatorname{var}(\hat{\vec{y}}/X) = \sigma^2 (B-H)(B-H)^T \geq 0$$

positive semidefinite matrix

KLJ

Gauss-Markov in short:

$$\hat{\vec{y}}/X \sim (X\beta, \sigma^2 H)$$

+ optimality

REMARK: Gauss-Markov theorem is the key tool to show easily that any linear combination of LSE $\hat{\beta}$ (if $r=k$) is BLUE of respective linear combination of β.

WILL NOW BE ASSUMED:

$\operatorname{rank}(X_{n\times k}) = k$

$\ell := (\ell_0, \dots, \ell_{k-1})^T \in \mathbb{R}^k$, $\ell \neq 0$

$L := \begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_m^T \end{pmatrix}$, $\ell_j \in \mathbb{R}^k$, $\ell_j \neq 0$, $j = 1, \dots, m$

$m \leq k$, rows in L linearly independent

- Interest in estimation of
  - (i) $\theta := \ell^T \beta$
  - (ii) $\theta := L\beta = \begin{pmatrix} \ell_1^T \beta \\ \vdots \\ \ell_m^T \beta \end{pmatrix}$

Theorem 2.5 Gauss-Markov for linear combinations

Assume a linear model $Y|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = k < n$. Then

(i) $\hat{\theta} := \ell^T \hat{\beta}$ is BLUE of $\theta = \ell^T \beta$ with

$\operatorname{var}(\hat{\theta}|X) = \sigma^2 \ell^T (X^T X)^{-1} \ell > 0$.

(ii) $\hat{\theta} := L\hat{\beta}$ is BLUE of $\theta = L\beta$ with

$\operatorname{var}(\hat{\theta}|X) = \sigma^2 L (X^T X)^{-1} L$ which is a positive definite matrix.

Proof: KEY STEP (the rest is more or less trivial):

$\mathcal{M}(X^T)$ = linear span of ROWS of X, each row is $\in \mathbb{R}^k$, number of rows is $n > k$, $\text{rank}(X_{n\times k}) = k$

$= \mathbb{R}^k$

(i) That is, $\forall \ell \in \mathbb{R}^k$, $\ell \in \mathcal{M}(X^T)$.

Hence $\forall \ell \in \mathbb{R}^k \; \exists a \in \mathbb{R}^n \quad \ell = X^T a$

(linear comb. of n rows from X)

Hence $\ell^T \hat{\beta} = a^T X \hat{\beta}$ for some $a \in \mathbb{R}^n$

$= a^T \hat{Y}$ — " —

$(= a^T H Y)$

That is, $\ell^T \hat{\beta}$ is a linear functn of $\hat{Y}$

this is BLUE of $X\beta$.

$\Rightarrow \ell^T \hat{\beta}$ is BLUE of $a^T X \beta = \ell^T \beta$

REMARK: Unbiasedness now holds also unconditionally:

$E(\ell^T \hat{\beta}) = E(E(\ell^T \hat{\beta} | X)) = E(\ell^T \beta) =$

$= \ell^T \beta$ for any $\beta \in \mathbb{R}^k$

• $var(\ell^T\hat{\beta}|X) = var(a^T\hat{y}|X) = a^T var(\hat{y}|X)a =$

$= a^T(\sigma^2 H)a = \sigma^2 \underbrace{a^T X}_{\ell^T} (X^TX)^{-1} \underbrace{X^T a}_{\ell} =$

$= \sigma^2 \ell^T (X^TX)^{-1}\ell$

OR CALCULATED DIRECTLY:

$var(\ell^T\hat{\beta}|X) = \ell^T var(\hat{\beta}|X)\ell =$

$\overset{\text{Lemma 2.2}}{=} \ell^T(\sigma^2(X^TX)^{-1})\ell = \sigma^2\ell^T(X^TX)^{-1}\ell$

$var(\ell^T\hat{\beta}|X) > 0$ since $\ell \neq 0$ and

$(X^TX)^{-1}$ is positive definite matrix

(ii) The BLUE property of $\hat{\theta} = L\hat{\beta}$ as an estimator of $\theta = L\beta$ follows from (i).

Finally:

$var(L\hat{\beta}|X) = L\, var(\hat{\beta}|X)\, L^T =$ ($var(\hat{\beta}|X) = \sigma^2(X^TX)^{-1}$)

$= \sigma^2 L(X^TX)^{-1}L^T$

IS THIS MATRIX POSITIVE DEFINITE?

one more observation:

$$\mathcal{M}(\underbrace{X^TX}_{k\times k \text{ regular matrix}}) = \mathbb{R}^k = \mathcal{M}(X^T)$$

Hence, for each row from L: $l_j \in \mathcal{M}(X^TX)$,

i.e. $\mathcal{M}(\underset{k\times m}{L^T}) \subset \mathcal{M}(X^TX)$

$\Rightarrow$ There exists $B_{k\times m}$ with linearly indep. columns (since L has lin. indep. rows) such that

$$L^T = X^T\underbrace{XB}_{=:A_{n\times m}}$$

(A also lin. indep. columns)

$\Rightarrow$

$$L(X^TX)^{-1}L^T = B^TX^TX(X^TX)^{-1}X^TXB =$$

$$= B^TX^TXB = \underbrace{A^TA}_{m\times m \text{ matrix}}$$

A has linear indep. columns, hence A^TA is regular and positive definite matrix. ∎

In particular (take $L = I_k$):

$\hat{\beta} = (X^TX)^{-1}X^TY$ is BLUE of β.

2.4 Residuals, properties

REMIND: Data $(Y_i, X_{i\cdot}^T)^T$, $i = 1, \dots, n$ not neces. iid

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} = (X^0, \dots, X^{k-1})$$

It is assumed $Y | X \sim (X\beta, \sigma^2 I_n)$

$\operatorname{rank}(X_{n\times k}) = r \leq k < n$

MODEL: $\mathbb{E}(Y|X) = X\beta$ $\longrightarrow$ its BLUE: $\hat{Y} = H \cdot Y \in \mathcal{M}(X)$

$\operatorname{var}(Y|X) = \sigma^2 I_n$ $\quad = QQ^T Y$

$\quad = X(X^TX)^- X^T Y$

how to estimate this?

Residuals $U = Y - \hat{Y} = MY \in \mathcal{M}(X)^\perp$

$= NN^T Y$

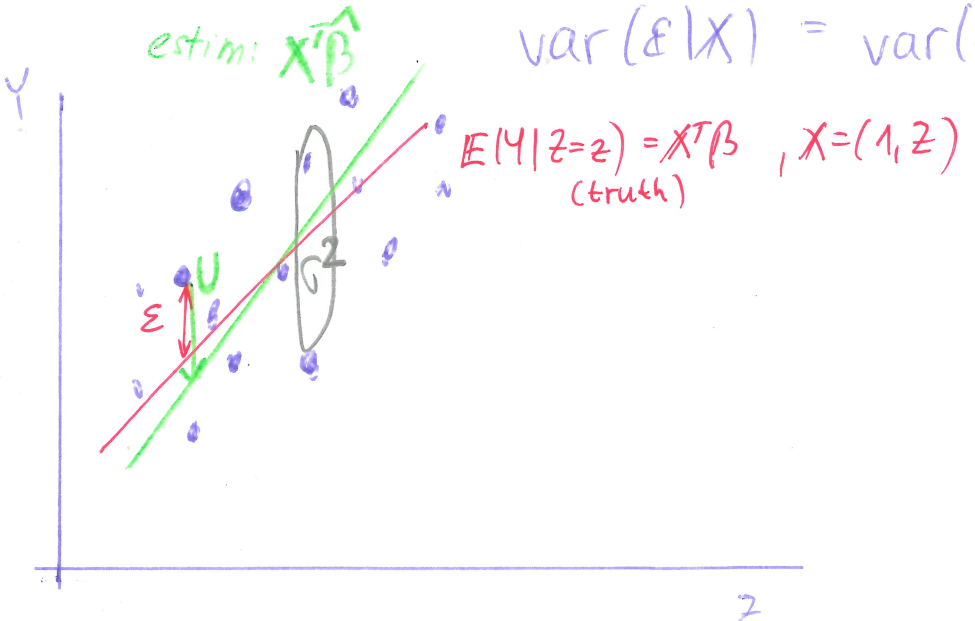
$= (I_n - H)Y$

We also have:

Error terms $\varepsilon = Y - X\beta = Y - \mathbb{E}(Y|X)$

already know: $\mathbb{E}(\varepsilon|X) = \mathbb{E}(\varepsilon) = 0$

$\operatorname{var}(\varepsilon|X) = \operatorname{var}(\varepsilon) = \sigma^2 I_n$



Def 2.5 **Residual sum of squares**

Consider a linear model $Y/X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = r \leq k < n$. The quantity

$$SS_e = \|U\|^2 \left(= \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2\right)$$

will be called the **residual sum of squares**.

Lemma 2.6 **Alternative expressions of residuals and residual sum of squares**

Let $Y/X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = r \leq k < n$. The following then holds:

(i) $U = M\varepsilon$, where $\varepsilon = Y - X\beta$.

(ii) $SS_e = Y^T M Y = \varepsilon^T M \varepsilon$.

Proof (i) $U = MY = M(\underbrace{X\beta + \varepsilon}_{Y}) =$

$= \underbrace{MX}_{O_{n\times k}}\beta + M\varepsilon = M\varepsilon$.

($O_{n\times k}$ (projections of cols. of X into $\mathcal{M}(X)^\perp$)

(ii) $SS_e = \|U\|^2 = U^T U = (MY)^T MY =$

$= Y^T \underbrace{M^T M}_{M} Y = Y^T M Y$

$\overset{(i)}{=} \varepsilon^T M^T M \varepsilon = \varepsilon^T M \varepsilon$

Lemma 2.7 Moments of residuals and resid. sum of sq.

Let $Y|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = r \le k < n$.

Then (i) $E(U|X) = 0_n$, $\operatorname{var}(U|X) = \sigma^2 M$.

(ii) $E(SSe|X) = E(SSe) = (n-r)\sigma^2$.

Proof: (i) $E(U|X) = E(MY|X) =$

($M = I_n - X(X^TX)^-X^T$)

$= ME(Y|X) = \underbrace{MX}_{0_{n\times k}}\beta = 0$

(projections of cols. of X into $\mathcal{M}(X)^\perp$)

$\operatorname{var}(U|X) = \operatorname{var}(MY|X) = M \underbrace{\operatorname{var}(Y|X)}_{\sigma^2 I_n} M^T =$

$= \sigma^2 MM^T = \sigma^2 M$.

(ii) $E(SSe|X) \overset{\text{Lemma 2.6}}{=} E(\varepsilon^T M \varepsilon | X) =$

$= E(\operatorname{tr}(\varepsilon^T M \varepsilon)|X) = E(\operatorname{tr}(M\varepsilon\varepsilon^T)|X) =$

$= \operatorname{tr}\{E(M\varepsilon\varepsilon^T|X)\} = \operatorname{tr}\{M \underbrace{E(\varepsilon\varepsilon^T|X)}_{\operatorname{var}(\varepsilon|X) = \sigma^2 I_n}\} =$

$= \operatorname{tr}(M \cdot \sigma^2 I_n) = \sigma^2 \operatorname{tr}(M) = \sigma^2 \operatorname{tr}(NN^T) =$

(orth. basis of $\mathcal{M}(X)^\perp$ in cols. / vec-dim = $n-r$)

$= \sigma^2 \operatorname{tr}(N^TN) = \sigma^2 \operatorname{tr}(I_{n-r}) =$

$= \sigma^2(n-r)$

$E(SSe) = E(E(SSe|X)) = E(\sigma^2(n-r)) = \sigma^2(n-r)$

□

Lemma 2.7 (i) in short: $U|X \sim (0, \sigma^2 M)$

Consequence of Lemma 2.7 (ii):

Quantity $\frac{SSe}{n-r}$ is unbiased (both conditionally given X and unconditionally) estimator of σ^2.

Def. 2.6 **Residual mean square and residual degrees of freedom**

Consider a linear model $Y|X \sim (X\beta, \sigma^2 I_n)$, $\text{rank}(X_{n\times k}) = r \leq k < n$.

(i) The residual mean square of the model is the quantity $\frac{SSe}{n-r}$ and will be denoted as MSe, i.e. $MSe = \frac{SSe}{n-r}$.

(ii) The residual degrees of freedom of the model is the vector dimension of the residual space $\mathcal{M}(X)^\perp$ and will be denoted as ν_e, i.e. $\nu_e = n - r$.