STAT 8004 Lecture 3

Thursday, February 5, 2015 5:30 PM

Review:

Gauss-Markov Model

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

$E(\epsilon)=0$, $var(\epsilon) = \sigma^2 I$

Conventionally, we assume $\epsilon \sim N(0, \sigma^2 I)$

for statistical inferences $\begin{cases} \text{testing} \\ \text{confidence region} \end{cases}$

(next 2 weeks)

Take X as fixed, $E(Y) = X\beta$

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

objective, knowing $E(Y) = X\beta$, i.e. $E(Y) \in C(X)$

want to find $\hat{Y}$ $\left(\widehat{E(Y)}\right)$

$$\hat{Y} = \arg\min_{\hat{Y}\in C(X)} \underbrace{\|Y - \hat{Y}\|_2^2}\ \leftarrow \text{OLS}$$

- This task is "always" doable.

$$\boxed{\hat{Y} = X\underbrace{(X^T X)^-}X^T Y = PY}$$

- unique.

$\hat{Y} = X\hat{\beta}$: $\hat{\beta}$ may be ambiguous

$a^T\hat{Y}$: unique

$\underbrace{a^T}X\hat{\beta}$: unique

- Definition of estimable function. $c^T\beta$ is estimable iff $a^T X = c^T$ $\left(c \in C(X^T)\right)$

- Testablity $C = \begin{pmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_\ell^T \end{pmatrix}$

H_0: $C\beta = 0$ is testable.

if $\begin{cases} 1) \ c_1^T\beta, c_2^T\beta \cdots c_\ell^T\beta \text{ are estimable} \\ 2) \ \text{rank}(C) = \ell \end{cases}$

- Results: $Y = X\beta + \epsilon$

$Y = X\beta + \epsilon$

$\hat{Y} = PY = X(X^TX)^- X^T Y$

1) $E(\hat{Y}) = X(X^TX)^-X^T(E(Y))$

$= \underbrace{X(X^TX)^-X^T}_{P}(X\beta)$

$= X\beta$ (unbiasness)

2) $Var(\hat{Y}) = var(PY)$

$= P\sigma^2 I P^T = \sigma^2 PP = \sigma^2 P$

$= \sigma^2 X(X^TX)^-X^T$

3) $E(Y-\hat{Y}) = E(e) = 0$

$Var(Y-\hat{Y}) = var((I-P)Y) = \sigma^2(I-P)$

$- \quad \hat{Y}, \quad \underbrace{C = \begin{pmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_\ell^T \end{pmatrix} \triangleq AX}$

$\underline{\underline{C\beta}}$ Q: $\widehat{C\beta}$

$\widehat{C\beta} = C(X^TX)^-X^TY \quad (= APY)$

$E(\widehat{C\beta}) = C(X^TX)^-X^TX\beta$

$= C\beta$ (unbiased)

$var(\widehat{C\beta}) = var(APY)$

$= AP\sigma^2 I P^T A^T$

$= \sigma^2 APA^T$

$= \sigma^2 \underbrace{AX}_{C}(X^TX)^-\underbrace{X^TA^T}_{C^T}$

$= \sigma^2 C(X^TX)^-C^T$

if X is of full rank, and $C = I$

then $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

You can show that

$$E\left((Y-\hat{Y})^T(Y-\hat{Y})\right) = \underline{\underline{\sigma}}^2 (n-\text{rank}(X))$$

$$\hat{\sigma}^2 = \frac{(Y-\hat{Y})^T(Y-\hat{Y})}{\underline{\underline{n-\text{rank}(X)}}} = \frac{e^Te}{n-\text{rank}(X)}$$

$$= \frac{SSE}{n-\text{rank}(X)} = MSE$$

— unbiased estimator

- $\widehat{c\beta}$, $\widehat{\sigma^2}$ _____ × _____

- $\widehat{c\beta}$ is $\boxed{\text{the best}}$ unbiased estimator for $c\beta$

- (Gauss-Markov Theorem)

In the linear model, $Y = X\beta + \epsilon$, $E(\epsilon)=0$, $\text{var}(\epsilon)=\sigma^2 I$

Then for $c \in C(X)$, $\widehat{c^T\beta}$ (OLS estimator) is the best linear unbiased estimator (BLUE) of $c^T\beta$ in the sense that $\text{var}(c^T\beta)$ is the smallest among all linear unbiased estimators for $c^T\beta$.

$$c^T\beta = c^T(X^TX)^- \underbrace{X^T}Y \leftarrow \text{a linear combination of } Y.$$

Proof: let d^TY be a unbiased estimator of $c^T\beta$

$$E(d^TY) = d^TE(Y) = \underbrace{d^TX}\beta = \underbrace{c^T}\beta$$

$$\Rightarrow \underline{d^TX = c^T} \qquad (*)$$

- $\boxed{\text{var}(d^TY) \geq \text{var}(\widehat{c^T\beta})}$ we need to show

$$\text{var}(d^TY) = \text{var}(\widehat{c^T\beta} + d^TY - \widehat{c^T\beta})$$

$$= \text{var}(\widehat{c^T\beta}) + \underbrace{\boxed{\text{var}(d^TY - \widehat{c^T\beta})}}_{\geq 0}$$

$$+ 2\,\text{cov}(\widehat{c^T\beta}, d^TY - \widehat{c^T\beta})$$

$$\geq \text{var}(\widehat{c^T\beta}) + 2\underbrace{\text{cov}(\widehat{c^T\beta}, d^TY - \widehat{c^T\beta})}$$

$$\text{cov}(\widehat{c^T\beta}, d^TY - \widehat{c^T\beta})$$

$$= \text{cov}(a^TPY, (d^T - a^TP)Y)$$

$$= a^T P \boxed{\sigma^2 I} (d - Pa)$$

$$= \sigma^2 \left(a^T P d - a^T \overset{= P}{PP} a \right)$$

$$= \sigma^2 \left(\underbrace{a^T X}_{= C^T} (X^T X)^{-} \underbrace{X^T d}_{= C} - \underbrace{a^T X}_{C^T} (X^T X)^{-} \underbrace{X^T a}_{= C} \right)$$

$$= 0$$

—— x ——

- $\widehat{C^T\beta}$ is the optimal unbiased (BLUE)

- Remark: The optimality does not hold more broadly.

$\boxed{GM}$	Aiken's model
$E(\epsilon) = 0$	$E(\epsilon) = 0$
$var(\epsilon) = \sigma^2 I$	$var(\epsilon) = \sigma^2 V$
	V: a general variance covariance matrix

Example: one-way model, 3 groups, 2 obs per group

$$y_{ij} = \mu_{ij} + \epsilon_{ij} \quad \left(\begin{matrix} i = 1, 2, 3 \\ j = 1, 2 \end{matrix} \right)$$

Now assume that the second observation is more accurate, whose variance is $1/100$ of the first one

$$Y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & & \\ 1 & & \\ & 1 & \\ & 1 & \\ & & 1 \\ & & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \epsilon$$

$$var(\epsilon) = diag(1, 0.01, 1, 0.01, 1, 0.01)\,\eta^2$$

$$\overset{OLS}{\Rightarrow} \quad \hat{Y} = \begin{pmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{3\cdot} \end{pmatrix} \qquad \boxed{\bar{y}_{i\cdot} = \frac{y_{i1} + y_{i2}}{2}}$$

$$\hat{\beta} = \begin{pmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{3\cdot} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix}$$

$$\boxed{var(\hat{\mu}_1)} = var\left(\frac{y_{11} + y_{12}}{2} \right) = \frac{1}{4} var(y_{11} + y_{12})$$

$= \frac{1}{4}(1+0.01)\eta^2$

$\boxed{var(\hat{\mu}_1) = var(y_{12})}$ $\quad (\tilde{\mu}_1 = y_{12})$

$= 0.01\,\eta^2$

$var(\tilde{\mu}_1) < var(\hat{\mu}_1)$

— y —

- What to do in an Aiken model?

- Answer: Generalized (weighted) least square.

$E(\epsilon) = 0 \quad , \quad var(\epsilon) = \sigma^2 V$

Fact: $\quad V = V^{1/2} V^{1/2} \quad$ (square root matrix $V^{1/2}$)

let $U = V^{-1/2} Y$

Now $E(U) = V^{-1/2} E(Y) = \underbrace{\boxed{V^{-1/2} X}}_{= W} \beta$

$var(U) = V^{-1/2}\, var(Y)\, V^{-1/2}$

$= V^{-1/2}\, V\, V^{-1/2}\, \sigma^2$

$= \sigma^2 I$

GM assumptions hold for

$\underline{U = W\beta + \epsilon^*}$

Now, OLS for U makes perfect sense.

Example (continued)

$$\begin{cases} V = diag(1, 0.01, 1, 0.01, 1, 0.01) \\ V^{-1} = diag(1, 100, 1, 100, 1, 100) \\ V^{-1/2} = diag(1, 10, 1, 10, 1, 10) \end{cases}$$

$$U = V^{-1/2} Y = \begin{pmatrix} Y_{11} \\ 10 Y_{12} \\ Y_{21} \\ 10 Y_{22} \\ Y_{31} \end{pmatrix}$$

$$\begin{pmatrix} \vdots \\ y_{21} \\ 10y_{22} \\ y_{31} \\ 10y_{32} \end{pmatrix}$$

$$W = V^{-\frac{1}{2}} X = \begin{pmatrix} 1 & 0 & \\ 10 & & \\ & 1 & 0 \\ 0 & 10 & \\ & 0 & 1 \\ & & 10 \end{pmatrix}$$

This model for $U = W\beta + \epsilon^*$

$$\Rightarrow \hat{\beta} = \begin{pmatrix} \frac{y_{11}+100y_{12}}{101} \\ \frac{y_{21}+100y_{22}}{101} \\ \frac{y_{31}+100y_{32}}{100} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix}$$

$$\hat{U} = \begin{pmatrix} \hat{\mu}_1 \\ 10\hat{\mu}_1 \\ \vdots \\ \vdots \end{pmatrix}$$

To get $\hat{Y} = V^{\frac{1}{2}} \hat{U} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_3 \end{pmatrix}$.

Note. $\hat{\beta}$ minimizes

$$(U - W\beta)^T (U - W\beta)$$

$$= (V^{-\frac{1}{2}}Y - V^{-\frac{1}{2}}X\beta)^T (V^{-\frac{1}{2}}Y - V^{-\frac{1}{2}}X\beta)$$

$$= (Y - X\beta)^T V^{-\frac{1}{2}} V^{-\frac{1}{2}} (Y - X\beta)$$

$$= \underbrace{(Y - X\beta)^T V^{-1} (Y - X\beta)}$$

Remark: Things will change substantially if we step out "Linear".

$$(\widehat{c^T\beta}(U) = c^T (W^T W)^- W^T U$$

$$= c^T (W^T W)^- W^T V^{-\frac{1}{2}} Y .$$

$$= c^T (W^T W)^- W^T V^{-\frac{1}{2}} y$$

Example : $y_{ij} = \mu + \epsilon_{ij}$ for 2 groups
3 obs per group

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 \end{pmatrix} \mu + \epsilon$$

$$var(\epsilon) = diag(\sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_2^2)$$

Now, if $\sigma_1^2 / \sigma_2^2 = r$ is known

$$Var(\epsilon) = \sigma_2^2 diag(r, r, r, 1, 1, 1)$$

Then with the Aiken's model.

$$\hat{\mu} = \frac{\bar{y}_1 + r\bar{y}_2}{1+r} = \underbrace{\frac{\frac{1}{\sigma_1^2}\bar{y}_1 + \frac{1}{\sigma_2^2}\bar{y}_2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}}$$

But, if σ_1^2 & σ_2^2 are unknown

$$\hat{\mu} = \frac{\frac{1}{S_1^2}\bar{y}_1 + \frac{1}{S_2^2}\bar{y}_2}{\frac{1}{S_1^2} + \frac{1}{S_2^2}}$$

is a reasonable estimator.

However, there is nothing can be said about its optimality.

. In fact, it is not even a <u>linear</u> estimator.

— x —

- $E(\epsilon) = 0$, $var(\epsilon) = \sigma^2 V$.

$$\epsilon \sim f(0, \sigma^2 V)$$

— x —

$$\epsilon \sim f(0, \sigma^2 V)$$ —x—

- $\boxed{\text{OLS, GLS}}$

- Reparametrization.

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & & \\ 1 & & \\ & 1 & \\ & & 1 \\ & & 1 \\ & & 1 \end{pmatrix}}_{X} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \epsilon$$

$$= \underbrace{\begin{pmatrix} 1 & 1 & & \\ 1 & 1 & & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & 1 \\ 1 & & & 1 \end{pmatrix}}_{W} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \epsilon$$

$$\boxed{C(X) = C(W)}$$

- $Y = X\beta + \epsilon$ & $Y = W\underline{\gamma} + \epsilon$

are fundamentally the same if $C(X) = C(W)$.

- Since the column spaces are same,

$\hat{Y}$, $\widehat{c^T\beta}$, $Y - \hat{Y}$, etc. are all the same.

$$\underline{a^T\hat{Y}} = \boxed{\underbrace{a^T X}_{c^T}\hat{\beta}} = \boxed{\underbrace{a^T W}_{d^T}\hat{\gamma}}$$

Now, what is the connection between

$\boxed{c^T\beta \ \& \ d^T\gamma}$? —x—

- If $C(X) = C(W)$, then it must be true

that $W = XF$ for some F

$$c^T\beta = \underbrace{a^T X}\beta = a^T W \gamma$$

$$= a^T (XF)\gamma$$
$$= \underbrace{(a^T X}F)\gamma$$
$$= \underbrace{(c^T F)}\gamma = \underbrace{d^T}\gamma$$

$$d^T = c^T \underline{F}$$

example.
$$\underbrace{\begin{pmatrix} 1 & & \\ 1 & & \\ & 1 & \\ & & 1 \\ & & 1 \\ & & 1 \end{pmatrix}}_{X} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

$$\underbrace{\begin{pmatrix} 1 & 1 & & \\ 1 & 1 & & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & 1 \\ 1 & & & 1 \end{pmatrix}}_{W} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

$$F = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

$$W = XF$$

Now. for $\underbrace{(1, 0, 0)}_{c^T} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \underline{\mu_1}$

$$d^T = c^T F = (1, 1, 0, 0)$$
$$d^T \gamma = (1, 1, 0, 0) \begin{pmatrix} \mu \\ \alpha_1 \\ \hline \alpha_2 \\ \alpha_3 \end{pmatrix}$$
$$= \boxed{\mu + \alpha_1}$$

Remark: How to choose between X & W ?
(full rank) (effect)

1. for computational / formula convenience one may prefer a full rank parametrization.

one may prefer a full rank parametrization.

$$\boxed{\begin{array}{l}\text{2. for scientific interpretations between parameters. Sometimes, one may use}\\ \text{other parametrizations,}\\ \text{"e.g." two way model } (\alpha_i,\ \beta_j,\ \delta_{ij})\end{array}}$$

- Get a full ranked parametrization

example. 2×2 factorial design with 2 obs per group.

$$\begin{array}{cc} & A \\ B & \begin{array}{|c|c|}\hline & \\ \hline & \\ \hline\end{array}\end{array}$$

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

for k^{th} obs in the ij group.

$$\begin{pmatrix} Y_{111}\\ Y_{112}\\ Y_{121}\\ Y_{122}\\ Y_{211}\\ Y_{212}\\ Y_{221}\\ Y_{222}\end{pmatrix} = \underbrace{\begin{pmatrix} 1 & & & \\ 1 & & & \\ & 1 & & \\ & 1 & & \\ & & 1 & \\ & & 1 & \\ & & & 1 \\ & & & 1 \end{pmatrix}}_{\text{full rank.}} \begin{pmatrix}\mu_{11}\\ \mu_{12}\\ \mu_{21}\\ \mu_{22}\end{pmatrix} + \epsilon$$

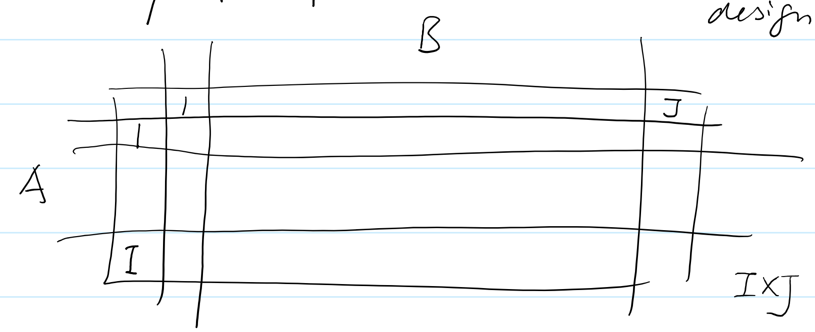
effect model.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$$

$$\boxed{\underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & & 1 & & & \\ 1 & 1 & 0 & 1 & & 1 & & & \\ 1 & 1 & 0 & 0 & & & 1 & & \\ 1 & 1 & 0 & 0 & & & 1 & & \\ 1 & 0 & 1 & 0 & & & & 1 & \\ 1 & 0 & 1 & 0 & & & & 1 & \\ 1 & 0 & 1 & 1 & & & & & 1 \\ 1 & 0 & 1 & 1 & & & & & 1 \end{pmatrix}}_{\text{none full rank design.}} \begin{pmatrix}\mu\\ \alpha_1\\ \alpha_2\\ \beta_1\\ \beta_2\\ \delta_{11}\\ \delta_{12}\\ \delta_{21}\\ \delta_{22}\end{pmatrix}}$$

- Using contraints / restriction to produce full rank design matrix

- Two ways for imposing constraints in $I \times J$ factorial design



$\alpha_1, \cdots \alpha_I$

$\beta_1, \cdots \beta_J$

$\delta_{ij}, \quad i = 1, \cdots I$
$\qquad\quad j = 1, \cdots J$

Constraint 1	2
Sum restrictions	Baseline restriction (SAS)
$\sum \alpha_i = 0$	$\alpha_I = 0$
$\sum \beta_j = 0$	$\beta_J = 0$
$\sum_i \delta_{ij} = \sum_j \delta_{ij} = 0$	$\delta_{Ij} = 0, j = 1 \cdots J$
	$\delta_{iJ} = 0, i = 1, \cdots I$

- Why do these help?

- example. 2×2 factorial.

a) Sum restrictions

δ_{11}	δ_{12}
δ_{21}	δ_{22}

$\alpha_1 + \alpha_2 = 0 \Rightarrow \alpha_2 = -\underline{\underline{\alpha_1}}$

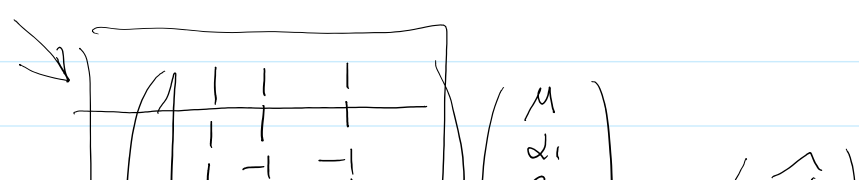
$\beta_1 + \beta_2 = 0 \Rightarrow \beta_2 = -\underline{\underline{\beta_1}}$

$\delta_{11} + \delta_{12} = 0 \Rightarrow \delta_{12} = -\delta_{11}$

$\delta_{21} + \delta_{22} = 0 \Rightarrow \delta_{22} = -\delta_{21} = -\underline{\underline{\delta_{11}}}$

$\delta_{12} + \delta_{22} = 0 \Rightarrow \delta_{22} = -\delta_{12}$
$\qquad = \delta_{11}$

(μ)



$$\begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \delta_{11} \end{pmatrix} \quad (\hat{Y})$$

of full rank.

$$E(Y_{111}) = \mu + \alpha_1 + \beta_1 + \delta_{11}$$

$$E(Y_{222}) = \mu - \alpha_1 - \beta_1 + \delta_{11}$$

b) baseline constraints, $\boxed{\alpha_2 = 0, \quad \beta_2 = 0 \quad \delta_{12} = 0 \quad \delta_{21} = 0 \quad \delta_{22} = 0}$

$$\begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \delta_{11} \end{pmatrix} \quad \boxed{(\hat{Y})}$$

Another design of full rank.

$E(Y_{111})$

$$E(Y_{222}) = \mu.$$

$C(X)$, $\hat{Y}$ $\rightarrow$ all others.

$H_0: \widehat{c\beta} = 0$?

$\widehat{c^T\beta}$ inferences.

next two weeks.