Chapter 2 Estimation

Hakbae Lee

The Department of Statistics and Data Science, Yonsei University

Let's consider a linear model

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{\epsilon}_{n\times 1} = \left(\mathbf{x}_i^T\beta\right) + \epsilon$$

where $\mathbf{x}_i^T$ is the i-th row vector of $\mathbf{X}$, $i = 1, \ldots, n$

$$\mathrm{E}(\epsilon) = 0 \quad \text{and} \quad \mathrm{Cov}(\epsilon) = \sigma^2 I$$

or

$$\mathrm{E}(\epsilon) = 0 \quad \text{and} \quad \mathrm{Cov}(\epsilon) = \sigma^2\Sigma$$

Identifiability and Estimability

- A general linear model is a parameterization

$$E(\mathbf{Y}) = f(\mathbf{X}) = E(\mathbf{X}\beta + \epsilon) = \mathbf{X}\beta + E(\epsilon) = \mathbf{X}\beta$$

- **Definition 2.1.1** The parameter β is identifiable if for any β_1 and β_2, $f(\beta_1) = f(\beta_2)$ implies $\beta_1 = \beta_2$. If β is identifiable, we say that the parameterization f(β) is identifiable. Moreover, a vector-valued function $g(\beta)$ is identifiable if $f(\beta_1) = f(\beta_2)$ implies $g(\beta_1) = g(\beta_2)$.
- For regression models for which $r(X) = p$, the parameters are identifiable: $\mathbf{X}^T\mathbf{X}$ is nonsingular, so if $\mathbf{X}\beta_1 = \mathbf{X}\beta_2$, then

$$\beta_1 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta_1 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta_2 = \beta_2$$

- **Theorem 2.1.2** A function $g(\beta)$ is identifiable if and only if $g(\beta)$ is a function of $f(\beta)$.

Identifiability and Estimability

- **Definition 2.1.3** A vector-valued linear function of β, $\Lambda^T\beta$ is estimable if $\Lambda^T\beta = P^T\mathbf{X}\beta$ for some matrix P; In other words, $\Lambda^T\beta$ is estimable if

$$\Lambda = \mathbf{X}^T P \in \mathcal{C}(\mathbf{X}^T)$$

- Clearly, if $\Lambda^T\beta$ is estimable, it is identifiable and therefore it is a reasonable thing to estimate.
- For estimable functions $\Lambda^T\beta = P^T\mathbf{X}\beta$, although P need not be unique, its perpendicular projection (columnwise) onto $\mathcal{C}(\mathbf{X})$ is unique:
  Let P_1 and P_2 be matrices with $\Lambda^T = P_1^T\mathbf{X} = P_2^T\mathbf{X}$, then

$$MP_1 = \mathbf{X}(\mathbf{X}^T X)^- X^T P_1 = X(X^T X)^- \Lambda = \mathbf{X}(\mathbf{X}^T X)^- X^T P_2 = MP_2.$$

- Example 2.1.4 and 2.1.5

Identifiability and Estimability

- **Definition 2.1.7** An estimate $f(\mathbf{Y})$ of $g(\beta)$ is unbiased if $\mathrm{E}[f(\mathbf{Y})] = g(\beta)$ for any β.

- **Definition 2.1.8** $f(\mathbf{Y})$ is a linear estimate of $\Lambda^T\beta$ if $f(\mathbf{Y}) = a_0 + a^T\mathbf{Y}$ for some scalar a_0 and vector a

- **Proposition 2.1.9** A linear estimate $a_0 + a^T\mathbf{Y}$ is unbiased for $\Lambda^T\beta$ if and only if $a_0 = 0$ and $a^T\mathbf{X} = \Lambda^T$; say,

$$\Lambda = \mathbf{X}^T a \in \mathcal{C}(\mathbf{X}^T)$$

- **Corollary 2.1.10** $\Lambda^T\beta$ is estimable if and only if there exists ρ such that $\mathrm{E}(\rho^T\mathbf{Y}) = \Lambda^T\beta$ for any β .

Estimation: Least Squares

- Estimating $E(\mathbf{Y})$ is to take a vector in $\mathcal{C}(\mathbf{X})$ closest to $\mathbf{Y}$;

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{X}\beta \in \mathcal{C}(\mathbf{X}) \\ \hat{\beta} &= \text{least squares estimate (LSE) of } \beta \\ &= \min_{\beta}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta) = \min_{\beta}||\mathbf{Y}-\mathbf{X}\beta||^2 \end{aligned}$$

- LSE of $\Lambda^T\beta = \Lambda^T\hat{\beta}$ for any least squares estimate $\hat{\beta}$.

Theorem 2.2.1 $\hat{\beta}$ is a LSE of β if and only if $\mathbf{X}\hat{\beta} = M\mathbf{Y}$, where M is the perpendicular projection operator onto $\mathcal{C}(\mathbf{X})$.

Corollary 2.2.2

$$\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y} = \text{ LSE of } \beta$$

Estimation: Least Squares

Corollary 2.2.3 The unique LSE of $\rho^T \mathbf{X}\beta = \rho^T M\mathbf{Y}$.

Note: The unique LSE of $\Lambda^T\beta = \Lambda^T\hat{\beta} = P^T M\mathbf{Y}$.

Theorem 2.2.4 The LSE of $\lambda^T\beta$ is unique only if $\lambda^T\beta$ is estimable: $\lambda = \mathbf{X}^T\rho$ if $\lambda^T\hat{\beta}_1 = \lambda^T\hat{\beta}_2$ so that $\mathbf{X}\hat{\beta}_1 = \mathbf{X}\hat{\beta}_2 = M\mathbf{Y}$

Note: When β is not identifiable, we need side conditions imposed on the parameters to estimate nonidentifiable parameters.

Note: With $r = r(\mathbf{X}) < p$(overparameterized model), we need $p - r$ individual side conditions to identify and estimate the parameters

Proposition 2.2.5 If $\lambda = \mathbf{X}^T\rho$, then $\mathrm{E}(\rho^T M\mathbf{Y}) = \lambda^T\beta$.

Estimation: Least Squares

Let's decompose **Y** as

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} = M\mathbf{Y} + (I - M)\mathbf{Y}$$

where

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} = M\mathbf{Y} = \text{fitted values of } \mathbf{Y} \in \mathcal{C}(X) \\ \mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\beta} = (I - M)\mathbf{Y} = \text{residuals} \in \mathcal{C}(X)^{\perp} \end{aligned}$$

Theorem 2.2.6 Let $r(\mathbf{X}) = r$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I$. Then

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T(I - M)\mathbf{Y}}{n - r} = \text{an unbiased estimate of } \sigma^2 = MSE$$

where

$$n - r = \mathrm{rank}(I - M) = \text{degrees of freedom for error}$$

Definition 2.3.1 $a^T\mathbf{Y}$ is a best linear unbiased estimate(BLUE) of $\lambda^T\beta$ if $a^T\mathbf{Y}$ is unbiased, i.e., $\mathrm{E}(a^T\mathbf{Y}) = \lambda^T\beta$ and if for any other linear unbiased estimate $b^T\mathbf{Y}$, $\mathrm{Var}(a^T\mathbf{Y}) \leq \mathrm{Var}(b^T\mathbf{Y})$.

Gauss-Markov Theorem 2.3.2
Consider $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I$. Let $\lambda^T\beta$ be estimable. Then

$$\text{LSE of } \lambda^T\beta = \text{ BLUE of } \lambda^T\beta.$$

Corollary 2.3.3 Let $\sigma^2 > 0$. Then there exists a unique BLUE for any estimable function $\lambda^T\beta$.

Estimation: Maximum Likelihood

Assume that $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 I_n)$. Then the maximum likelihood estimates (MLEs) of β and σ^2 are obtained by maximizing the log of the likelihood so that

$$\begin{aligned}
(\hat{\beta}, \hat{\sigma}^2) &= \text{MLE of } (\beta, \sigma^2) \\
&= \max_{(\beta,\sigma^2)} \left\{ \frac{-n}{2}\log(2\pi) - \frac{1}{2}\log[(\sigma^2)^n] - \frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^2} \right\} \\
\hat{\beta} &= \text{LSE of } \beta \\
\hat{\sigma}^2 &= \frac{Y^T(I-M)\mathbf{Y}}{n}
\end{aligned}$$

Estimation: Minimum Variance Unbiased

Assume that $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$.

Definition 2.5.1 A vector-valued sufficient statistic $T(\mathbf{Y})$ is said to be complete if $\mathrm{E}[h(T(\mathbf{Y}))] = 0$ implies that $\Pr[h(T(\mathbf{Y})) = 0] = 1$ for all β and σ^2.

Theorem 2.5.2 If $T(\mathbf{Y})$ is a complete sufficient statistic, then $f(T(\mathbf{Y}))$ is a minimum variance unbiased estimate (MVUE) of $\mathrm{E}[f(T(\mathbf{Y}))]$.

Estimation: Minimum Variance Unbiased

Theorem 2.5.3 Let $\theta = (\theta_1, ..., \theta_s)^T$ and let $\mathbf{Y}$ be a random vector with p.d.f

$$f(\mathbf{Y}) = c(\theta) \exp\left[\sum_{i=1}^{s} \theta_i T_i(\mathbf{Y})\right] h(\mathbf{Y})$$

then $T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}), ..., T_s(\mathbf{Y}))^T$ is a complete sufficient statistic provided that neither θ nor $T(\mathbf{Y})$ satisfy any linear constraints.

Theorem 2.5.4 MSE is a minimum variance unbiased estimate(MVUE) of σ^2 and $\rho^T M\mathbf{Y}$ is MVUE of $\rho^T \mathbf{X}\beta$ whenever $\epsilon \sim N(0, I)$.

Sampling Distributions of Estimates

Assume that $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$. Then $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 I_n)$.

$$\Lambda^T \hat{\beta} = P^T M \mathbf{Y} \sim N(\Lambda^T \beta, \sigma^2 P^T M P) = N(\Lambda^T \beta, \sigma^2 \Lambda^T (\mathbf{X}^T\mathbf{X})^- \Lambda)$$

since $M = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$

$$\hat{\mathbf{Y}} = M\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 M)$$

If $\mathbf{X}$ is of full rank, then

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \sim N(\beta, \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1})$$

Exercise: Show that

$$\mathbf{Y}^T (I - M)\mathbf{Y}/\sigma^2 \sim \chi^2(r(I - M), \beta^T \mathbf{X}^T (I - M)\mathbf{X}\beta/2\sigma^2).$$

Do Exercise 2.1.

Generalized Least Squares(GLS)

Assume that for some known positive definite Σ,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \mathrm{E}(\epsilon) = 0, \quad \mathrm{Cov}(\epsilon) = \sigma^2\Sigma. \qquad (1)$$

By SD, $\Sigma = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$, (1) can be rewritten as

$$\begin{aligned} \Sigma^{\frac{-1}{2}}\mathbf{Y} &= \Sigma^{\frac{-1}{2}}\mathbf{X}\beta + \Sigma^{\frac{-1}{2}}\epsilon, \quad \mathrm{E}(\Sigma^{\frac{-1}{2}}\epsilon) = 0, \ \mathrm{Cov}(\Sigma^{\frac{-1}{2}}\epsilon) = \sigma^2\mathbf{I}. \qquad (2) \\ \mathbf{Y}_* &= \mathbf{X}_*\beta + \epsilon_*, \quad \mathrm{E}(\epsilon_*) = 0, \quad \mathrm{Cov}(\epsilon_*) = \sigma^2\mathbf{I} \end{aligned}$$

$$\begin{aligned} \hat{\beta}_{GLS} &= \text{generalized squares estimate (GLSE) of } \beta \\ &= \min_{\beta}(\mathbf{Y}_* - \mathbf{X}_*\beta)^T(\mathbf{Y}_* - \mathbf{X}_*\beta) = \min_{\beta}||\mathbf{Y}_* - \mathbf{X}_*\beta||^2 \\ &= \min_{\beta}(\mathbf{Y} - \mathbf{X}\beta)^T\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\beta) \end{aligned}$$

Generalized Least Squares(GLS)

Theorem 2.7.1

(a) $\lambda^T\beta$ estimable in model (1) if and only if $\lambda^T\beta$ is estimable in model (2).

(b) $\hat{\beta}$ is a GLSE of β if and only if

$$\mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-}\mathbf{X}^T\Sigma^{-1}\mathbf{Y} = \mathbf{X}\hat{\beta}: \text{ Normal Equation of GLS}$$

For any estimable function there exists a unique GLSE.

(c) GLSE estimate of estimable $\lambda^T\beta$ = BLUE of $\lambda^T\beta$.

(d) Let $\epsilon \sim N(0, \Sigma^2\Sigma)$. Then, GLSE of estimable $\lambda^T\beta$ = MVUE.

(e) Let $\epsilon \sim N(0, \Sigma^2\Sigma)$. Then, $\hat{\beta}_{GLS}$=$\hat{\beta}_{MLE}$.

Generalized Least Squares(GLS)

Normal Equation of GLS can be rewritten as

$$A\mathbf{Y} = \mathbf{X}\hat{\beta}$$

where $A = \mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-}\mathbf{X}^T\Sigma^{-1}$

Proposition 2.7.2. A is a projection operator onto $\mathcal{C}(\mathbf{X})$.

Proposition 2.7.3.

$$\text{Cov}(\mathbf{X}\hat{\beta}_{GLS}) = \sigma^2\mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-}\mathbf{X}^T$$

Corollary 2.7.4. Let $\lambda^T\beta$ be estimable. Then

$$\text{Var}(\lambda^T\hat{\beta}_{GLS}) = \sigma^2\lambda^T(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-}\lambda$$

Generalized Least Squares(GLS)

Note: $(\mathbf{I} - A)Y$ = residual vector of GLSE.

$$\begin{aligned}
\text{SSE}_{GLS} &= (\mathbf{Y}_* - \hat{\mathbf{Y}}_*)^T(\mathbf{Y}_* - \hat{\mathbf{Y}}_*) \\
&\vdots \qquad \text{check!} \\
&= \mathbf{Y}^T(\mathbf{I} - A)^T\Sigma^{-1}(\mathbf{I} - A)\mathbf{Y} \\
\text{MSE}_{GLS} &= \hat{\sigma}^2 = \frac{\text{SSE}_{GLS}}{n - r(\mathbf{X})}
\end{aligned}$$

$$\frac{\lambda^T\hat{\beta}_{GLS} - \lambda^T\beta_{GLS}}{\hat{\sigma}^2\lambda^T(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^-\lambda} \sim t(n - r(\mathbf{X}))$$

Generalized Least Squares(GLS)

Proposition 2.7.5. Let Σ be nonsingular and $\mathcal{C}(\Sigma \mathbf{X}) \subset \mathcal{C}(\mathbf{X})$. Then least squares estimates are BLUEs.

NOTE: For diagonal Σ, GLS is referred to as weighted least squares (WLS).

Exercise 2.5 Show that A is the perpendicular projection operator onto $\mathcal{C}(X)$ when the inner product between two vectors x and y is defined as $(x, y)_\Sigma \equiv x^T \Sigma^{-1} y$.