Predictive Distributions & Properties of MLES

Merlise Clyde

STA721 Linear Models

Duke University

September 15, 2016

Outline

Topics

- Predictive Distributions
- OLS/MLES Unbiased Estimation
- Gauss-Markov Theorem

Readings: Christensen Chapter 2, Chapter 6.3, (Appendix A, and Appendix B as needed)

Prediction

- Predict Y_* at $\mathbf{x}_*^T$ (could be new point or existing point) $\mathbf{Y}_* = \mathbf{x}_*^T + \epsilon_*$
- $\mathrm{E}[Y_* \mid \mathbf{x}_*] = \mathbf{x}_*^T \boldsymbol{\beta} = \mu_*$ minimizes squared error loss for predicting Y_* at $\mathbf{X}_*^T$

$$\begin{aligned}
\mathrm{E}[Y_* - f(\mathbf{x}_*)]^2 &= \mathrm{E}[Y_* - \mu_* + \mu_* - f(x_*)]^2 \\
&= E[Y_* - \mu_*]^2 + E[\mu_* - f(x_*)]^2 + \\
&\quad 2\mathrm{E}[(Y_* - \mu_*)(\mu_* - f(x_*))] \\
&\geq E[Y_* - \mu_*]^2
\end{aligned}$$

Crossproduct term is 0:

$$\mathrm{E}[\mathrm{E}[(Y_* - \mu_x)(\mu_* - f(x_*)) \mid \mathbf{x}_*]] = \mathrm{E}[0 \cdot (\mu_* - f(x_*))]$$

- $E[Y_* \mid \mathbf{x}_*]$ is the "best" predictor of Y_*
- MLE of μ_* is $\mathbf{x}_*^T \hat{\boldsymbol{\beta}} = \hat{Y}_*$ (is this unique?)
- OLS Best Linear predictor of $\mathbf{Y}_*$
- Under joint Normality of $\mathbf{Y}, \mathbf{X}$ Best Predictor

Predictive Distribution

Look at

$$Y_* - \hat{Y}_* = \mathbf{x}^{*T}\boldsymbol{\beta} - \mathbf{x}_*^T\hat{\boldsymbol{\beta}} + \epsilon_*$$

$$\text{var}(Y - \hat{Y}) = \text{var}(\mathbf{x}_*^T\boldsymbol{\beta} - \mathbf{x}_*^T\hat{\boldsymbol{\beta}}) + \text{var}(\epsilon_*)$$

Two Sources of variation:

- Variation of estimator around true regression
- Variation of error around true regression

Distribution

Distribution of

$$\frac{Y_* - \mathbf{x}_*^T \hat{\beta}}{\sqrt{\mathrm{MSE}(1 + \mathbf{x}_* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_*^T)}} \sim t(n - p - 1, 0, 1)$$

$(1 - \alpha)100$ % Prediction Interval

$$\mathbf{x}_*^T \hat{\beta} \pm t_{\alpha/2} \sqrt{\mathrm{MSE}(1 + \mathbf{x}_* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_*^T)}$$

Models & MLEs

- $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ with $\boldsymbol{\mu} \in C(\mathbf{X}) \Leftrightarrow \boldsymbol{\mu} = \mathbf{X}\beta$
- Maximum Likelihood Estimator (MLE) of $\boldsymbol{\mu}$ is $\mathbf{P_X Y}$
- $\mathbf{P_X}$ is the orthogonal projection operator on the column space of $\mathbf{X}$; e.g. $\mathbf{X}$ full rank $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- If $\mathbf{X}^T\mathbf{X}$ is not invertible use a generalized inverse

A generalize inverse of $\mathbf{A}$: $\mathbf{A}^-$ satisfies $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$

Lemma (B.43)

If $\mathbf{G}$ *and* $\mathbf{H}$ *are generalized inverses of* $(\mathbf{X}^T\mathbf{X})$ *then*

1. $\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{X} = \mathbf{X}$
2. $\mathbf{X}\mathbf{G}\mathbf{X}^T = \mathbf{X}\mathbf{H}\mathbf{X}^T$

$\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is the orthogonal projection operator onto $C(\mathbf{X})$ (does not depend on choice of generalized inverse!) [See proof in Theorem B.44]

Generalize Inverses

A generalize inverse of $\mathbf{A}$: $\mathbf{A}^-$ satisfies $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$
Special Case: Moore-Penrose Generalized Inverse

- Decompose symmetric $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$
- $\mathbf{A}^-_{MP} = \mathbf{U}\boldsymbol{\Lambda}^-\mathbf{U}^T$
- $\boldsymbol{\Lambda}^-$ is diagonal with

$$\lambda_i^- = \begin{cases} 1/\lambda_i & \text{if } \lambda_i \neq 0 \\ 0 & \text{if } \lambda_i = 0 \end{cases}$$

- Symmetric $\mathbf{A}^-_{MP} = (\mathbf{A}^-_{MP})^T$
- Reflexive $\mathbf{A}^-_{MP}\mathbf{A}\mathbf{A}^-_{MP} = \mathbf{A}^-_{MP}$

If $\mathbf{P}$ is an orthogonal projection matrix, the generalized inverse of $\mathbf{P}$, $\mathbf{P}^- = \mathbf{P}$

MLE of β

$$\begin{aligned} \mathbf{P_X Y} &= \mathbf{X}\hat{\beta} \\ \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y} &= \mathbf{X}\hat{\beta} \end{aligned}$$

- MLE of β iff $\mathbf{P_X Y} = \mathbf{X}\hat{\beta}$
- If $\mathbf{X}^T\mathbf{X}$ is invertible, then

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

  and is unique
- But if $\mathbf{X}^T\mathbf{X}$ is not invertible,

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

  is one solution which depends on choice of generalized inverse

What can we estimate uniquely?

Identifiability

$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$

- Distribution of $\mathbf{Y}$ determined by $\boldsymbol{\mu}$ and σ^2
- $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \mu(\boldsymbol{\beta})$

Identifiability

$\boldsymbol{\beta}$ and σ^2 are identifiable if distribution of $\mathbf{Y}$, $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}_1, \sigma_1^2) = f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}_2, \sigma_2^2)$ implies that $(\boldsymbol{\beta}_1, \sigma_1^2)^T = (\boldsymbol{\beta}_2, \sigma_2^2)^T$

For linear models, equivalent definition is that $\boldsymbol{\beta}$ is identifiable if for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ $\mu(\boldsymbol{\beta}_1) = \mu(\boldsymbol{\beta}_2)$ implies that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. If $r(\mathbf{X}) = p$ then $\boldsymbol{\beta}$ is identifiable If $\mathbf{X}$ is not full rank, there exists

$\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$, but $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$ and hence $\boldsymbol{\beta}$ is not identifiable

Non-Identifiable

Recall the One-way ANOVA model

$$\mu_{ij} = \mu + \tau_j \qquad \boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{n_1 1}, \mu_{12}, \ldots, \mu_{n_2,2}, \ldots, \mu_{1J}, \ldots, \mu_{n_J J})^T$$

- Let $\boldsymbol{\beta}_1 = (\mu, \tau_1, \ldots, \tau_J)^T$
- Let $\boldsymbol{\beta}_2 = (\mu - 42, \tau_1 + 42, \ldots, \tau_J + 42)^T$
- Then $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ even though $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$
- $\boldsymbol{\beta}$ is not identifiable
- yet $\boldsymbol{\mu}$ is identifiable, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ (a linear combination of $\boldsymbol{\beta}$)

Identifiability and Estimability

Theorem

A function $g(\beta)$ is identifiable if and only if $g(\beta)$ is a function of $\mu(\beta)$

In linear models, historical focus on linear functions. Identifiable linear functions are called *estimable* functions

Definition

A vector valued function $\mathbf{\Lambda}\beta$ is *estimable* if $\mathbf{\Lambda}\beta = \mathbf{AX}\beta$ for some matrix $\mathbf{A}$

Equivalently

Definition

A vector valued function $\mathbf{\Lambda}\beta$ is *estimable* if it has an unbiased linear estimator, i.e. there exists an $\mathbf{A}$ such that $\mathsf{E}(\mathbf{AY}) = \mathbf{\Lambda}\beta$ for all β

Estimability

Work with scalar functions $\psi = \boldsymbol{\lambda}^T \boldsymbol{\beta}$

Theorem

The function $\psi = \boldsymbol{\lambda}^T \beta$ is estimable if and only if $\boldsymbol{\lambda}^T$ is a linear combination of the rows of $\mathbf{X}$. i.e. there exists $\mathbf{a}^T$ such that $\boldsymbol{\lambda}^T = \mathbf{a}^T \mathbf{X}$

Proof.

The function $\psi = \boldsymbol{\lambda}^T \boldsymbol{\beta}$ is estimable if there exists an $\mathbf{a}^T$ such that $\mathsf{E}[\mathbf{a}^T \mathbf{Y}] = \boldsymbol{\lambda}^T \boldsymbol{\beta}$

$$\begin{aligned} \mathsf{E}[\mathbf{a}^T \mathbf{Y}] &= \mathbf{a}^T \mathsf{E}[\mathbf{Y}] \\ &= \mathbf{a}^T \mathbf{X} \boldsymbol{\beta} \\ &= \boldsymbol{\lambda}^T \boldsymbol{\beta} \end{aligned}$$

if and only if $\boldsymbol{\lambda}^T = \mathbf{a}^T \mathbf{X}$ for all $\boldsymbol{\beta}$ □

Estimability of Individual β_j

Proposition

For

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \sum_j \mathbf{X}_j \beta_j$$

β_j is not identifiable if and only if there exists α_j such that $\mathbf{X}_j = \sum_{i \neq j} \mathbf{X}_i \alpha_i$

One-way Anova Model:

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots & \\ \mathbf{1}_{n_J} & \mathbf{0}_{n_J} & \mathbf{0}_{n_J} & \dots & \mathbf{1}_{n_J} \end{bmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_J \end{pmatrix}$$

Are any parameters μ or τ_j identifiable?

Gauss-Markov Theorem

Theorem

Under the assumptions:

$$\begin{aligned} E[\mathbf{Y}] &= \boldsymbol{\mu} \\ Cov(\mathbf{Y}) &= \sigma^2 \mathbf{I}_n \end{aligned}$$

every estimable function $\psi = \boldsymbol{\lambda}^T \boldsymbol{\beta}$ *has a unique unbiased linear estimator* $\hat{\psi}$ *which has minimum variance in the class of all unbiased linear estimators.* $\hat{\psi} = \boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}$ *where* $\hat{\boldsymbol{\beta}}$ *is any set of ordinary least squares estimators.*

Unique Unbiased Estimator

Lemma

- If $\psi = \boldsymbol{\lambda}^T \beta$ is estimable, there exists a unique linear unbiased estimator of $\psi = \mathbf{a}^{*T}\mathbf{Y}$ with $a^* \in C(\mathbf{X})$.
- If $\mathbf{a}^T\mathbf{Y}$ is any unbiased linear estimator of ψ then a^* is the projection of $\mathbf{a}$ onto $C(\mathbf{X})$, i.e. $\mathbf{a}^* = \mathbf{P}_{\mathbf{X}}\mathbf{a}$.

Unique Unbiased Estimator

Proof

- Since ψ is estimable, there exists an $\mathbf{a} \in \mathbb{R}^n$ for which $\mathsf{E}[\mathbf{a}^T\mathbf{Y}] = \boldsymbol{\lambda}^T\boldsymbol{\beta} = \psi$
- Let $\mathbf{a} = \mathbf{a}^* + \mathbf{u}$ where $\mathbf{a}^* \in C(\mathbf{X})$ and $\mathbf{u} \in C(\mathbf{X})^\perp$
- Then

$$\begin{aligned} \psi = \mathsf{E}[\mathbf{a}^T\mathbf{Y}] &= \mathsf{E}[\mathbf{a}^{*T}\mathbf{Y}] + \mathsf{E}[\mathbf{u}^T\mathbf{Y}] \\ &= \mathsf{E}[\mathbf{a}^{*T}\mathbf{Y}] + 0 \end{aligned}$$

$$\mathsf{E}[\mathbf{u}^T\mathbf{Y}] = \mathbf{u}^T\mathbf{X}\boldsymbol{\beta}$$

since $\mathbf{u} \perp C(\mathbf{X})$ (i.e. $\mathbf{u} \in C(\mathbf{X})^\perp$) $\mathsf{E}[\mathbf{u}^T\mathbf{Y}] = 0$

- Thus $\mathbf{a}^{*T}\mathbf{Y}$ is also an unbiased linear estimator of ψ with $\mathbf{a}^* \in C(\mathbf{X})$

Uniqueness

Proof.

Suppose that there is another $\mathbf{v} \in C(\mathbf{X})$ such that $\mathsf{E}[\mathbf{v}^T\mathbf{Y}] = \psi$. Then for all β

$$\begin{aligned} 0 &= \mathsf{E}[\mathbf{a}^{*T}\mathbf{Y}] - \mathsf{E}[\mathbf{v}^T\mathbf{Y}] \\ &= (\mathbf{a}^* - \mathbf{v})^T\mathbf{X}\beta \\ \text{So } (\mathbf{a}^* - \mathbf{v})^T\mathbf{X} &= 0 \quad \text{for all } \beta \end{aligned}$$

- Implies $(\mathbf{a}^* - \mathbf{v}) \in C(\mathbf{X})^\perp$
- but by assumption $(\mathbf{a}^* - \mathbf{v}) \in C(\mathbf{X})$ ($C(\mathbf{X})$ is a vector space)
- the only vector in BOTH is $\mathbf{0}$, so $\mathbf{a}^* = \mathbf{v}$

Therefore $\mathbf{a}^{*T}\mathbf{Y}$ is the unique linear unbiased estimator of ψ with $\mathbf{a}^* \in C(\mathbf{X})$. $\square$

Proof of Minimum Variance

- Let ${\mathbf{a}^*}^T\mathbf{Y}$ be the unique unbiased linear estimator of ψ with $\mathbf{a}^* \in C(\mathbf{X})$.
- Let $\mathbf{a}^T\mathbf{Y}$ be any unbiased estimate of ψ; $\mathbf{a} = \mathbf{a}^* + \mathbf{u}$ with $\mathbf{a}^* \in C(\mathbf{X})$ and $\mathbf{u} \in C(\mathbf{X})^\perp$

$$
\begin{aligned}
\mathrm{Var}(\mathbf{a}^T\mathbf{Y}) &= \mathbf{a}^T\mathrm{Cov}(\mathbf{Y})\mathbf{a} \\
&= \sigma^2\|\mathbf{a}\|^2 \\
&= \sigma^2(\|\mathbf{a}^*\|^2 + \|\mathbf{u}\|^2 + 2{\mathbf{a}^*}^T\mathbf{u}) \\
&= \sigma^2(\|\mathbf{a}^*\|^2 + \|\mathbf{u}\|^2) + 0 \\
&= \mathrm{Var}({\mathbf{a}^*}^T\mathbf{Y}) + \sigma^2\|\mathbf{u}\|^2 \\
&\geq \mathrm{Var}({\mathbf{a}^*}^T\mathbf{Y})
\end{aligned}
$$

with equality if and only if $\mathbf{a} = \mathbf{a}^*$

Hence ${\mathbf{a}^*}^T\mathbf{Y}$ is the unique linear unbiased estimator of ψ with minimum variance "BLUE" = Best Linear Unbiased Estimator

Continued

Proof.

Show that $\hat{\psi} = \mathbf{a}^{*T}\mathbf{Y} = \boldsymbol{\lambda}^T\hat{\beta}$

Since $\mathbf{a}^* \in C(\mathbf{X})$ we have $\mathbf{a}^* = \mathbf{P_X}\mathbf{a}^*$

$$\begin{aligned} \mathbf{a}^{*T}\mathbf{Y} &= \mathbf{a}^{*T}\mathbf{P}_X^T\mathbf{Y} \\ &= \mathbf{a}^{*T}\mathbf{P}_x\mathbf{Y} \\ &= \mathbf{a}^{*T}\mathbf{X}\hat{\beta} \\ &= \boldsymbol{\lambda}^T\hat{\beta} \end{aligned}$$

for $\boldsymbol{\lambda}^T = \mathbf{a}^{*T}\mathbf{X}$ $\square$

MVUE

- Gauss-Markov Theorem says that OLS has minimum variance in the class of all Linear Unbiased estimators
- Requires just first and second moments
- Additional assumption of normality, OLS = MLEs have minimum variance out of ALL unbiased estimators; not just linear estimators (requires Completeness and Rao-Blackwell Theorem - next semester)
- Mean Squared Error for estimator $g(\mathbf{Y})$ of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is

$$\mathsf{E}[g(\mathbf{Y}) - \boldsymbol{\lambda}^T\boldsymbol{\beta}]^2 = \mathsf{Var}(g(\mathbf{Y})) + \mathsf{Bias}^2(g(\mathbf{Y}))$$

  where $\mathsf{Bias} = \mathsf{E}[g(\mathbf{Y})] - \boldsymbol{\lambda}^T\boldsymbol{\beta}$
- Bias vs Variance tradeoff
- Can have smaller MSE if we allow some Bias!