

# Predictive Distributions & Properties of MLES

Merlise Clyde

STA721 Linear Models

Duke University

September 11, 2019

# Outline

## Topics

- ▶ Predictive Distributions
- ▶ OLS/MLES Unbiased Estimation
- ▶ Gauss-Markov Theorem (if time)

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)  
 $\mathbf{Y}_* = \mathbf{x}_*^T \boldsymbol{\beta} + \epsilon_*$
- ▶  $E[Y_* | \mathbf{x}_*] = \mathbf{x}_*^T \boldsymbol{\beta} = \mu_*$  minimizes squared error loss for predicting  $Y_*$  at  $\mathbf{X}_*^T$

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

Crossproduct term is 0:

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

- ▶ equality if  $f(\mathbf{x}) = E[Y_* | \mathbf{x}_*]$ , the “best” predictor of  $Y_*$
- ▶ MLE of  $\mu_*$  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 (reducible error)
- ▶ Variation of error around true regression (irreducible error)

# Distribution

Distribution of pivotal quantity

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

Number of columns (rank) of  $\mathbf{X}$  is  $p$

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

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

# Models & MLEs

- ▶  $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$  with  $\boldsymbol{\mu} \in C(\mathbf{X}) \Leftrightarrow \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$
- ▶ Maximum Likelihood Estimator (MLE) of  $\boldsymbol{\mu}$  is  $\mathbf{P}_\mathbf{X}\mathbf{Y}$
- ▶  $\mathbf{P}_\mathbf{X}$  is the orthogonal projection operator on the column space of  $\mathbf{X}$ ; e.g.  $\mathbf{X}$  full rank  $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- ▶ 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.  $\mathbf{XGX}^T\mathbf{X} = \mathbf{XHX}^T\mathbf{X} = \mathbf{X}$
2.  $\mathbf{XGX}^T = \mathbf{XHX}^T$

$\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T$  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  $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$

▶  $\mathbf{A}_{MP}^- = \mathbf{U}\mathbf{\Lambda}^-\mathbf{U}^T$

▶  $\mathbf{\Lambda}^-$  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 $\beta$

$$\begin{aligned}\mathbf{P}_\mathbf{X}\mathbf{Y} &= \mathbf{X}\hat{\beta} \\ \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} &= \mathbf{X}\hat{\beta}\end{aligned}$$

- ▶ MLE of  $\beta$  iff  $\mathbf{P}_\mathbf{X}\mathbf{Y} = \mathbf{X}\hat{\beta}$
- ▶ 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

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

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

## Identifiability

$\boldsymbol{\beta}$  and  $\sigma^2$  are identifiable if distribution of  $\mathbf{Y}$ ,

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}_1, \sigma_1^2) = f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}_2, \sigma_2^2) \text{ implies that } (\boldsymbol{\beta}_1, \sigma_1^2)^T = (\boldsymbol{\beta}_2, \sigma_2^2)^T$$

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

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

- ▶ Let  $\boldsymbol{\beta}_1 = (\mu, \tau_1, \dots, \tau_J)^T$
- ▶ Let  $\boldsymbol{\beta}_2 = (\mu - 42, \tau_1 + 42, \dots, \tau_J + 42)^T$
- ▶ 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  $\mathbf{L}\beta$  is *estimable* if  $\mathbf{L}\beta = \mathbf{A}\mathbf{X}\beta$  for some matrix  $\mathbf{A}$

Equivalently

## Definition

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

# Estimability

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

## Theorem

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

## Proof.

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

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

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



# Estimability of Individual $\beta_j$

## Proposition

For

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

$\beta_j$  is not identifiable if and only if there exists  $\alpha_j$  such that

$$\mathbf{X}_j = \sum_{i \neq j} \mathbf{X}_i \alpha_i$$

One-way Anova Model:

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

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

Are any parameters  $\mu$  or  $\tau_j$  identifiable?

# Gauss-Markov Theorem

## Theorem

*Under the assumptions:*

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

*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.*