Identifiability, Gauss Markov & Predictive Distributions

Merlise Clyde

STA721 Linear Models

Duke University

September 15, 2019

Outline

Topics

- Gauss-Markov Theorem
- Estimability and Prediction

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

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 $g(\boldsymbol{\beta})$ is identifiable if and only if $g(\boldsymbol{\beta})$ is a function of $\mu(\boldsymbol{\beta})$

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

Definition

A scalar function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is *estimable* if $\boldsymbol{\lambda}^T\boldsymbol{\beta} = \mathbf{a}^T\mathbf{X}\boldsymbol{\beta}$ for some vector $\mathbf{a} \in \mathbb{R}^n$

Equivalently

Definition

A function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is *estimable* if it has an unbiased linear estimator, i.e. there exists an $\mathbf{a}$ such that $\mathrm{E}(\mathbf{a}^T\mathbf{Y}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$

Estimability

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 $\boldsymbol{\beta}$ $\square$

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 $\psi = \boldsymbol{\lambda}^T \boldsymbol{\beta}$ 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. $\mathbf{a}^* = \mathbf{P_X a}$.

Unique Unbiased Estimator

Proof

- Since ψ is estimable, there exists an $\mathbf{a} \in \mathbb{R}^n$ for which $\mathsf{E}[\mathbf{a}^T\mathbf{Y}] = \boldsymbol{\lambda}^T\boldsymbol{\beta} = \psi$ with $\boldsymbol{\lambda}^T = \mathbf{a}^T\mathbf{X}$
- 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 $\boldsymbol{\beta}$

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

- 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 (G-M)

- 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

duke.eps

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 $\boldsymbol{\lambda}^T = \mathbf{a}^{*T}\mathbf{X}$ or $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$ □

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 (MVUE); not just linear estimators (requires Completeness and Rao-Blackwell Theorem - next semester)

Prediction

- For predicting at new $\mathbf{x}_*$ is there always a unique unbiased estimator of $E[\mathbf{Y} \mid \mathbf{x}_*]$?
- If one does exist, how do we know that if we are given $\boldsymbol{\lambda}$?

Existence

- $\mathbf{x}_*^T\boldsymbol{\beta}$ has a unique unbiased estimator if $\mathbf{x}_* \equiv \boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$
- Clearly if $\mathbf{x}_* = \mathbf{x}_i$ (ith row of observed data) then it is estimable with $\mathbf{a}$ equal to the vector with a 1 in the ith position even if $\mathbf{X}$ is not full rank!
- What about out of sample prediction?

Example

```
x1 = -4:4
x2 = c(-2, 1, -1, 2, 0, 2, -1, 1, -2)
x3 = 3*x1  -2*x2
x4 = x2 - x1 + 4
Y = 1+x1+x2+x3+x4 + c(-.5,.5,.5,-.5,0,.5,-.5,-.5,.5)
dev.set = data.frame(Y, x1, x2, x3, x4)
lm1234 = lm(Y ~ x1 + x2 + x3 + x4, data=dev.set)
round(coefficients(lm1234), 4)

## (Intercept)          x1          x2          x3          x4
##           5           3           0          NA          NA

lm3412 = lm(Y ~ x3 + x4 + x1 + x2, data = dev.set)
round(coefficients(lm3412), 4)

## (Intercept)          x3          x4          x1          x2
##         -19           3           6          NA          NA
```

In Sample Predictions

```
cbind(dev.set, predict(lm1234), predict(lm3412))

##      Y x1 x2  x3 x4 predict(lm1234) predict(lm3412)
## 1 -7.5 -4 -2  -8  6              -7              -7
## 2 -3.5 -3  1 -11  8              -4              -4
## 3 -0.5 -2 -1  -4  5              -1              -1
## 4  1.5 -1  2  -7  7               2               2
## 5  5.0  0  0   0  4               5               5
## 6  8.5  1  2  -1  5               8               8
## 7 10.5  2 -1   8  1              11              11
## 8 13.5  3  1   7  2              14              14
## 9 17.5  4 -2  16 -2              17              17
```

Both models agree for estimating the mean at the observed **X** points!

Out of Sample

```
out = data.frame(test.set,
      Y1234=predict(lm1234, new=test.set),
      Y3412=predict(lm3412, new=test.set))
out
```

```
##   x1 x2 x3 x4 Y1234 Y3412
## 1  3  1  7  2    14    14
## 2  6  2 14  4    23    47
## 3  6  2 14  0    23    23
## 4  0  0  0  4     5     5
## 5  0  0  0  0     5   -19
## 6  1  2  3  4     8    14
```

Agreement for cases 1, 3, and 4 only! Can we determine that without finding the predictions and comparing?

Determining Estimable $\boldsymbol{\lambda}$

- Estimable means that $\boldsymbol{\lambda}^T = \mathbf{a}^T\mathbf{X}$ for $\mathbf{a} \in C(\mathbf{X})$
- Transpose: $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$ for $\mathbf{a} \in C(\mathbf{X})$
- $\boldsymbol{\lambda} \in C(\mathbf{X}^T)$ $(\boldsymbol{\lambda} \in R(\mathbf{X}))$
- $\boldsymbol{\lambda} \perp C(\mathbf{X}^T)^\perp$
- $C(\mathbf{X}^T)^\perp$ is the null space of $\mathbf{X}$

$$\mathbf{v} \perp C(\mathbf{X}^T) : \mathbf{X}\mathbf{v} = 0 \Leftrightarrow \mathbf{v} \in N(\mathbf{X})$$

- $\boldsymbol{\lambda} \perp N(\mathbf{X})$
- if $\mathbf{P}$ is a projection onto $C(\mathbf{X}^T)$ then $\mathbf{I} - \mathbf{P}$ is a projection onto $N(\mathbf{X})$ and therefore $(\mathbf{I} - \mathbf{P})\boldsymbol{\lambda} = \mathbf{0}$ if $\boldsymbol{\lambda}$ is estimable

Take $\mathbf{P}_{\mathbf{X}^T} = (\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^-$ as a projection onto $C(\mathbf{X}^T)$ and show $(\mathbf{I} - \mathbf{P}_{\mathbf{X}^T})\boldsymbol{\lambda} = \mathbf{0}_p$

Example

```r
library("estimability" )
cbind(epredict(lm1234, test.set), epredict(lm3412, test.se
```

```
##   [,1] [,2]
## 1   14   14
## 2   NA   NA
## 3   23   23
## 4    5    5
## 5   NA   NA
## 6   NA   NA
```

Rows 2, 5, and 6 are not estimable! No linear unbiased estimator

Summary

- When BLUEs exist, under normality they are MVUE (ditto for prediction - BLUP)
- BLUE/BLUP do not always exist for estimation/prediction if $\mathbf{X}$ is not full rank
- may occur with redundancies for modest $p < n$ and of course $p > n$
- Eliminate redundancies by removing variables (variable selection)
- Consider alternative estimators (Bayes and related)

Other Estimators

What about some estimator $g(\mathbf{Y})$ that is not unbiased?

- 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!

Bayes

- Next Class Bayes Theorem & Conjugate Normal-Gamma Prior/Posterior distributions
- Read Chapter 2 in Christensen or Wakefield 5.7
- Review Multivariate Normal and Gamma distributions