BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 4: Ordinary least squares (rank-deficient case)

Version: 09/12/2023

Table of Contents

Linearly dependent columns

Motivation:

- Previously, we saw that $\widehat{\mathbf{y}}$—the projection of $\mathbf{y}$ onto the linear subspace spanned by the columns of $\mathbf{X}$, is unique.
- We also saw that the least squares estimate solves the normal equations, given by:

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} \quad = \quad \mathbf{X}^T\mathbf{y}.$$

- If $\mathbf{X}$ is of full rank, the solution is unique: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Motivation:

- You seek to predict lung function (via, for instance, forced expiratory volume) based on age. Leaving no stone unturned, you measure age in years and decades. The design matrix is defined by:
  - $\mathbf{x}_0 = \mathbf{1}$.
  - $\mathbf{x}_1$, age in decades.
  - $\mathbf{x}_2$, age in years.
- If a solution to the normal equations is given by:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix},$$

the normal equations have infinitely many solutions of the form:

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} + c \begin{bmatrix} 0 \\ -10 \\ 1 \end{bmatrix}.$$

Lemma 4.1: Non-uniqueness $\Rightarrow$ infinitely many solutions

Suppose $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are two different least squares estimators. Then, there are infinitely many least squares estimators of $\boldsymbol{\beta}$.

Lemma 4.1: Outline of proof

- The idea is to show that $\widetilde{\boldsymbol{\beta}} = \alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2$ (which is different than each of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$) also solves the normal equations.

Lemma 4.2: Two solutions to the normal equations

Suppose $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are two different least squares estimators. Then, $||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_1||^2 = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_2||^2$.

Lemma 4.2: Outline of proof

- First show that $\mathbf{y}^T\mathbf{X} = (\mathbf{X}\widehat{\boldsymbol{\beta}}_1)^T\mathbf{X}$.
- Expand $||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_1||^2$, invoke equivalence above and continue manipulating to make it look like $||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_2||^2$.
- Why does this matter?
  - Solutions to the normal equations do not simply correspond to local minima; they achieve the same value; no one is more "valid" than the other.

Linearly dependent columns

Motivation:

- If $\mathbf{X}$ is rank-deficient, meaning that $\text{rank}(\mathbf{X}) = R < K$, then $\widehat{\boldsymbol{\beta}}$ is not unique.
- More specifically, $\boldsymbol{\beta}$ is not *identifiable* (an inherent property of the model rather than a circumstantial property of the estimator).
- Solutions to this problem include:
  1. Using a generalized inverse, $(\mathbf{X}^T\mathbf{X})^-$.
  2. Reducing the model to one of full rank.
  3. Imposing identifiability constraints.

Table of Contents

Brief aside: ANOVA

Ideas:

- We will give ANOVA a more in-depth look later in the semester, though since ANOVA problems frequently are the source of rank-deficient design matrices, it makes sense to say a little about it now.
- ANOVA is an abbreviation for **An**alysis **O**f **Va**riance. We will see later why it is so termed.
- Historically, ANOVA has been the approach of choice in the social sciences, laboratory research, etc., allowing one to compare means across discrete subgroups.
- Because it is simply a specific case of the broader linear regression framework (think about categorical predictors), it doesn't get extraordinary emphasis in BIOS 6312 (or in this course).

Brief aside: ANOVA

Ideas:

- With regression, we're often thinking about the design matrix as a first step; information about it is often very directly embedded in the models we write down:
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- With ANOVA, the design matrix is usually an afterthought. Instead, the initial step is to write the outcome in terms of group-specific means. For instance, if there are $i = 1, \ldots, n$ observations in each of two subgroups ($j = 0, 1$), an ANOVA model may be written as:
  - $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$
  - Here, μ represents an overall mean, and α_j the *additional* amount to add on to the overall mean based on membership to group j.
  - This model has two groups and three parameters, and is therefore clearly over-specified. ANOVA doesn't care. Our goal in *this* set of notes is not to care so much about why we would ever do this, but instead to focus on ways to handle the rank-deficiency that arises in this situation.

Table of Contents

Generalized inverses for the normal equations

One possible solution to the problem:

- If $\mathbf{X}$ is rank-deficient with $\text{rank}(\mathbf{X}) = R < K$, then $\mathbf{X}^T\mathbf{X}$ is also rank-deficient with $\text{rank}(\mathbf{X}^T\mathbf{X}) = R$, and hence $\mathbf{X}^T\mathbf{X}$ is not invertible.
- Recall: every matrix—even one that is singular—has a g-inverse.
- Let $(\mathbf{X}^T\mathbf{X})^-$ denote a g-inverse for $\mathbf{X}^T\mathbf{X}$, so that

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

- The normal equations can be written to reflect this:

$$\begin{aligned} \mathbf{X}^T\mathbf{y} &= \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}. \end{aligned}$$

- Therefore, $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$ solves the normal equations.

Generalized inverses for the normal equations

Theorem 4.1: Determining the orthogonal projection

The orthogonal projection, $\widehat{\mathbf{y}}$, of $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$ is given by $\mathbf{Py}$, where

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

for any generalized inverse, $(\mathbf{X}^T\mathbf{X})^-$.

- To prove this, we need a number of lemmas that you've already shown!

Lemma 4.3

If $(\mathbf{X}^T\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^T\mathbf{X}$, then $[(\mathbf{X}^T\mathbf{X})^-]^T$ is also a g-inverse of $\mathbf{X}^T\mathbf{X}$.

Generalized inverses for the normal equations

Lemma 4.4

If $(\mathbf{X}^T\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^T\mathbf{X}$, then $(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^-]^T$ is a symmetric, reflexive g-inverse of $\mathbf{X}^T\mathbf{X}$.

Generalized inverses for the normal equations

Lemma 4.5

If $(\mathbf{X}^T\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^T\mathbf{X}$, then $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^T$ and $(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is a g-inverse of $\mathbf{X}$.

Generalized inverses for the normal equations

Lemma 4.6

If $\mathbf{G}$ and $\widetilde{\mathbf{G}}$ are both g-inverses of $\mathbf{X}^T\mathbf{X}$, then $\mathbf{X}\mathbf{G}\mathbf{X}^T = \mathbf{X}\widetilde{\mathbf{G}}\mathbf{X}^T$.

Generalized inverses for the normal equations

Lemma 4.7

If $(\mathbf{X}^T\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^T\mathbf{X}$, then $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is symmetric.

Finding a g-inverse:

- Partition $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix},$$

where $\mathbf{X}_1$ consists of $R = \text{rank}(\mathbf{X})$ linearly independent columns of $\mathbf{X}$. Then, the following is a g-inverse of $\mathbf{X}^T\mathbf{X}$:

$$(\mathbf{X}^T\mathbf{X})^- = \begin{bmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

- Try to show this on your own! :)

Generalized inverses for the normal equations

Reminder: This is the statement of Theorem 4.1

- The orthogonal projection, $\widehat{\mathbf{y}}$, of $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$ is given by $\mathbf{Py}$, where

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

  for any generalized inverse, $(\mathbf{X}^T\mathbf{X})^-$.

Theorem 4.1: Proof

- First, let's prove that $\mathbf{P}$ is symmetric and idempotent.
  - Symmetry follows immediately from Lemma 4.7.
  - Idempotence follows from Lemma 4.5:

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

- It's therefore clear that $\mathbf{P}$ is an orthogonal projection matrix, as desired. Which one is it, though??

Theorem 4.1: Proof

- Now, since $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ should produce the same value regardless of the choice of $(\mathbf{X}^T\mathbf{X})^-$ (this is Lemma 4.6), let's choose the one proposed a few slides ago to see what that result is:

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}(\mathbf{X}_1^T\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{X}_1^T \\ \mathbf{X}_2^T\end{bmatrix} = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T.$$

- Therefore, $\mathbf{P}$ projects onto $\mathcal{C}(\mathbf{X}_1)$, the linearly independent columns of $\mathbf{X}$—which is to say that $\mathbf{P}$ projects onto $\mathcal{C}(\mathbf{X})$.

Generalized inverses for the normal equations

Example: One-way ANOVA with two groups

- As previously mentioned, this sort of situation arises frequently in the "ANOVA" formulation of a regression model.
- Consider the ANOVA model: $Y_{ij} = \mu + \gamma_j + \epsilon_{ij}$, with N_1 subjects in group $j = 1$ and N_2 subjects in group $j = 2$:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Clearly, this model is over-specified, with $\mathbf{X}$ only having two linearly independent columns.

Generalized inverses for the normal equations

Example: One-way ANOVA with two groups

- Under this model, we can find a solution to the normal equations by finding a generalized inverse of $\mathbf{X}^T\mathbf{X}$.
- Partition $\mathbf{X}$ so that $\mathbf{X}_1$ comprises its first two columns. In this case:

$$(\mathbf{X}^T\mathbf{X})^- = \begin{bmatrix} \frac{1}{N_2} & -\frac{1}{N_2} & 0 \\ -\frac{1}{N_2} & \frac{N}{N_1 N_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- With this in mind, note that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{1}{N_2} & -\frac{1}{N_2} & 0 \\ -\frac{1}{N_2} & \frac{N}{N_1 N_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{N_1} Y_{1i} + \sum_{i=1}^{N_2} Y_{2i} \\ \sum_{i=1}^{N_1} Y_{1i} \\ \sum_{i=1}^{N_2} Y_{2i} \end{bmatrix} = \begin{bmatrix} \overline{Y}_2 \\ \overline{Y}_1 - \overline{Y}_2 \\ 0 \end{bmatrix}$$

Generalized inverses for the normal equations

Example: One-way ANOVA with two groups

- It seems like a strange solution, though perhaps not if we account for the fact that it is not a unique solution to the normal equations.
- Nevertheless, we hope that our computation of $\widehat{\mathbf{y}}$ will give us an answer that *does* make sense, given that we're supposed to get the same answer irrespective of our selection of $(\mathbf{X}^T\mathbf{X})^-$.
- Indeed, here's the punchline:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \overline{Y}_2 \\ \overline{Y}_1 - \overline{Y}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_2 \end{bmatrix}.$$

Table of Contents

Reducing the model to full rank

Another possible solution to the problem:

- Suppose $\text{rank}(\mathbf{X}) = R < K$. Then, we can factor as $\mathbf{X} = \mathbf{K}_{N\times R}\mathbf{L}_{R\times K}$, where $\text{rank}(\mathbf{K}) = \text{rank}(\mathbf{L}) = R$.
- Re-parameterizing the model: $\text{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta} = \mathbf{K}\mathbf{L}\boldsymbol{\beta} = \mathbf{K}\boldsymbol{\alpha}$.
  - This model is of full rank, because $\mathbf{K}$ is of full rank!
- The (unique) least squares estimate of $\boldsymbol{\alpha}$ is $\widehat{\boldsymbol{\alpha}} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{y}$.
- Therefore, $\widehat{\mathbf{y}} = \mathbf{K}\widehat{\boldsymbol{\alpha}} = \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{y}$, from which it follows that $\mathbf{P} = \mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T$.
- Intuitive factorization: $\mathbf{X}_1$ denotes R linearly independent columns of $\mathbf{X}$, with $\mathbf{X}_2$ denoting the remaining $K - R$ columns, then:

$$\mathbf{X} \;=\; \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_1\mathbf{M}\end{bmatrix} = \mathbf{X}_1\begin{bmatrix}\mathbf{I} & \mathbf{M}\end{bmatrix}$$

- Therefore, $\mathbf{K} = \mathbf{X}_1$ and $\mathbf{L} = \begin{bmatrix}\mathbf{I} & \mathbf{M}\end{bmatrix}$, with $\mathbf{P} = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$.

More details:

- Leveraging the special factorization we identified, we can determine $\mathbf{M}$ as follows. Since $\mathbf{X}_2 = \mathbf{X}_1\mathbf{M}$, it follows that:

$$\begin{aligned} \mathbf{X}_1^T\mathbf{X}_2 &= \mathbf{X}_1^T\mathbf{X}_1\mathbf{M} \\ \Rightarrow (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2 &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_1\mathbf{M} \\ \Rightarrow (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2 &= \mathbf{M}, \end{aligned}$$

  where the second line follows from the fact that $\mathbf{X}_1$ is of full rank.
- Therefore, $\mathbf{L} = \begin{bmatrix} \mathbf{I} & (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2 \end{bmatrix}$, with $\boldsymbol{\alpha} = \mathbf{L}\boldsymbol{\beta}$.

Reducing the model to full rank

Example: One-way ANOVA with two groups

- To revisit our ANOVA example:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Let's see if we can use this new method to obtain the same fitted values.

Reducing the model to full rank

Example: One-way ANOVA with two groups

- Let $\mathbf{X}_1$ denote the first two columns of $\mathbf{X}$ (which are linearly independent).
- With a little matrix algebra, we should find $\mathbf{L}$ to be given by:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & N_2/N_1 \end{bmatrix}$$

- Therefore, it follows that:

$$\boldsymbol{\alpha} = \begin{bmatrix} \mu \\ \gamma_1 + \frac{N_2}{N_1}\gamma_2 \end{bmatrix}$$

Reducing the model to full rank

Example: One-way ANOVA with two groups

- Because we previously made use of the first two columns, it should come as no surprise that:

$$\widehat{\boldsymbol{\alpha}} = \begin{bmatrix} \overline{Y}_2 \\ \overline{Y}_1 - \overline{Y}_2 \end{bmatrix}, \text{ and } \widehat{\mathbf{y}} = \begin{bmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_2 \end{bmatrix}.$$

Table of Contents

Imposing identifiability constraints

Yet another possible solution to the problem:

- Suppose again $\text{rank}(\mathbf{X}) = R < K$. Then, one approach we can take is to impose a total of $K - R$ constraints on $\boldsymbol{\beta}$ taking the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, such that $\boldsymbol{\beta}$ is uniquely determined (identifiable).
  - $\mathbf{C}$ has dimensions $(K - R) \times K$.
- If we choose $\mathbf{C}$ properly, then for $\widehat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$, there is a unique $\widehat{\boldsymbol{\beta}}$ satisfying:

$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}} \text{ and } \mathbf{C}\widehat{\boldsymbol{\beta}} = \mathbf{0}.$$

- How do we choose $\mathbf{C}$ properly? In other words, under what conditions on $\mathbf{X}$ and $\mathbf{C}$ would it be the case that there is a unique solution to:

$$\begin{bmatrix} \widehat{\mathbf{y}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{C} \end{bmatrix} \widehat{\boldsymbol{\beta}} =: \mathbf{D}\widehat{\boldsymbol{\beta}}.$$

for any $\widehat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$?

Lemma 4.8: Conditions for uniqueness

There is a unique solution to:

$$\begin{bmatrix} \widehat{\mathbf{y}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{C} \end{bmatrix} \widehat{\boldsymbol{\beta}} =: \mathbf{D}\widehat{\boldsymbol{\beta}}$$

for any $\widehat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$ if and only if $\text{rank}(\mathbf{D}) = K$ and the rows of $\mathbf{C}$ are linearly independent of the rows of $\mathbf{X}$.

- That is to say that we require $\text{rank}(\mathbf{D}) = K$ and $\text{rank}(\mathbf{C}) = K - R$.

Solving the normal equations:

- To solve:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{C} \end{bmatrix} \boldsymbol{\beta} =: \mathbf{D}\boldsymbol{\beta}$$

  note that the normal equations are given by:

$$\mathbf{D}^T\mathbf{D}\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}.$$

- Because $\mathbf{D}$ is of full rank, so too is $\mathbf{D}^T\mathbf{D}$, and hence we obtain a unique solution:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{y}.$$

- Indeed, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{X}^T$.

Example: One-way ANOVA with two groups

- To revisit our ANOVA example:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Let's see if we can use this new method to obtain the same fitted values.

Imposing identifiability constraints

Example: One-way ANOVA with two groups

- For simplicity, let's assume that $N_1 = N_2 = N$ for this particular example (though we don't need to).
- I claim that this model requires only *one* constraint. Why?
- I claim that the constraint $\gamma_1 + \gamma_2 = 0$ is one that satisfies the criteria for this method.
  - To see this, let's check the corresponding constraint matrix, $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

  - Clearly, no linear combination of the rows of $\mathbf{X}$ will ever produce the row of $\mathbf{C}$.
- Question: could I have suggested the constraint $\gamma_1 = \gamma_2$?

Example: One-way ANOVA with two groups

- We need to determine the matrix $\mathbf{D}^T\mathbf{D} = \mathbf{X}^T\mathbf{X} + \mathbf{C}^T\mathbf{C}$:

$$\mathbf{D}^T\mathbf{D} = \begin{bmatrix} 2N & N & N \\ N & N+1 & 1 \\ N & 1 & N+1 \end{bmatrix}$$

- With a little bit of work, it can be shown that:

$$(\mathbf{D}^T\mathbf{D})^{-1} = \begin{bmatrix} \frac{N+2}{4N} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{N+2}{4N} & \frac{N-2}{4N} \\ -\frac{1}{4} & \frac{N-2}{4N} & \frac{N+2}{4N} \end{bmatrix} = \begin{bmatrix} \frac{1}{4}+\frac{1}{2N} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4}+\frac{1}{2N} & \frac{1}{4}-\frac{1}{2N} \\ -\frac{1}{4} & \frac{1}{4}-\frac{1}{2N} & \frac{1}{4}+\frac{1}{2N} \end{bmatrix}$$

Example: One-way ANOVA with two groups

- With a bit more work, it can be shown that:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{2}\begin{bmatrix}\overline{Y}_1 + \overline{Y}_2 \\ \overline{Y}_1 - \overline{Y}_2 \\ \overline{Y}_2 - \overline{Y}_1\end{bmatrix}$$

- Finally, we obtain

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix}\overline{Y}_1 \\ \vdots \\ \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_2\end{bmatrix}.$$

Table of Contents

Estimable functions

Example: One-way ANOVA with two groups

- To revisit our ANOVA example:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Can we estimate μ?
- Can we estimate $\mu + \gamma_1$?
- Can we estimate $\mu - \gamma_2$?
- It's not yet clear what these questions even mean, though our prior approaches to solving the normal equations have given us some clues.

Estimable functions

Estimable functions: Definition

- We say that $\mathbf{a}^T\boldsymbol{\beta}$ is estimable if for all $\boldsymbol{\beta}$ there is some $\mathbf{c}$ such that $\mathrm{E}[\mathbf{c}^T\mathbf{y}] = \mathbf{a}^T\boldsymbol{\beta}$ (that is, if $\mathbf{a}^T\boldsymbol{\beta}$ has a linear unbiased estimate).

Estimable functions

Lemma 4.9: Which functions are estimable?

The quantity $\mathbf{a}^T\boldsymbol{\beta}$ is estimable if and only if $\mathbf{a} \in \mathcal{R}(\mathbf{X})$.

Lemma 4.9: Proof

- $\mathrm{E}[\mathbf{c}^T\mathbf{y}] = \mathbf{c}^T\mathbf{X}\boldsymbol{\beta}$ (this follows from the fact that the linear model is correctly specified). Now, $\mathbf{c}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$ if and only if $\mathbf{a} = \mathbf{X}^T\mathbf{c}$, which is to say that $\mathbf{a} \in \mathcal{R}(\mathbf{X})$.
  - Intuition: Each observation is an unbiased estimate of its expected value in the sense that $\mathrm{E}[Y_i] = \mathbf{x}_i^T\boldsymbol{\beta}$. Therefore, we should be able to estimate anything of the form $\sum_i \alpha_i \mathbf{x}_i^T\boldsymbol{\beta}$ in an unbiased way (these are linear combinations of the rows of $\mathbf{X}$).
- Note: If $\mathbf{X}$ is of full rank, then any quantity of the form $\mathbf{a}^T\boldsymbol{\beta}$ will be estimable. Further, $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$ is the unique BLUE of $\mathbf{a}^T\boldsymbol{\beta}$. However, if $\mathbf{X}$ is not of full rank, our statements about estimability and optimality need to be crafted more cautiously.

Lemma 4.10: Uniqueness

If $\mathbf{a}^T\boldsymbol{\beta}$ is estimable, there is a unique $\mathbf{c}_* \in \mathcal{C}(\mathbf{X})$ such that $\mathbf{a} = \mathbf{X}^T\mathbf{c}_*$.

Lemma 4.10: Proof

- Suppose $\mathbf{a}^T\boldsymbol{\beta}$ is estimable. Then, by Lemma 4.9, $\mathbf{a} \in \mathcal{R}(\mathbf{X})$ so that $\mathbf{a} = \mathbf{X}^T\mathbf{c}$ for some $\mathbf{c}$. Any $\mathbf{c} \in \mathbb{R}^N$ can be uniquely decomposed as $\mathbf{c} = \mathbf{c}_* + \mathbf{c}_*^\perp$, with $\mathbf{c}_* \in \mathcal{C}(\mathbf{X})$ and $\mathbf{c}_*^\perp \in \mathcal{N}(\mathbf{X}^T)$.
  - Why? Recall: $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}^T)$ are orthogonal complements; $\mathcal{C}(\mathbf{X}) \oplus \mathcal{N}(\mathbf{X}^T) = \mathbb{R}^N$.
- Then,

$$\mathbf{a} \;=\; \mathbf{X}^T\mathbf{c} = \mathbf{X}^T(\mathbf{c}_* + \mathbf{c}_*^\perp) = \mathbf{X}^T\mathbf{c}_* + \mathbf{X}^T\mathbf{c}_*^\perp = \mathbf{X}^T\mathbf{c}_*$$

Lemma 4.11: Property of estimable functions

If $\mathbf{a}^T\boldsymbol{\beta}$ is estimable, then $\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{a}^T$ for any g-inverse $(\mathbf{X}^T\mathbf{X})^-$.

Lemma 4.11: Proof

- Suppose $\mathbf{a}^T\boldsymbol{\beta}$ is estimable. Then, by Lemma 4.10, $\mathbf{a} = \mathbf{X}^T\mathbf{c}_*$ with $\mathbf{c}_* \in \mathcal{C}(\mathbf{X})$. Then,

$$\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{c}_*^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{c}_*^T\mathbf{P}\mathbf{X} = \mathbf{c}_*^T\mathbf{X} = \mathbf{a}^T.$$

Lemma 4.12: Variance of estimable functions

If $\mathbf{a}^T\boldsymbol{\beta}$ is estimable, then $\text{Var}[\mathbf{a}^T\widehat{\boldsymbol{\beta}}] = \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{a}$.

Lemma 4.12: Proof

- Let $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$. Then,

$$\begin{aligned}
\text{Var}[\mathbf{a}^T\widehat{\boldsymbol{\beta}}] &= \text{Var}[\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}] \\
&= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{a} \\
&= \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{a} \\
&= \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{a}.
\end{aligned}$$

- Note: The "collapsing" we see in the final line of the proof follows from Lemma 4.11 (it does not necessarily imply that $(\mathbf{X}^T\mathbf{X})^-$ is a reflexive g-inverse, though it may be).
- Further, note that $\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{a} = \mathbf{c}_*^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{c}_* = \mathbf{c}_*^T\mathbf{P}\mathbf{c}_*$ is unique (meaning, the same for any g-inverse $(\mathbf{X}^T\mathbf{X})^-$).

Variance of estimable functions: Estimation

- Once again, we can't count on knowing the error variance, σ^2. How do we estimate it? We again rely on a previous useful result about the expectation of quadratic forms:

$$\begin{aligned} \mathrm{E}[\mathrm{RSS}] &= \mathrm{E}[\mathbf{y}^T(\mathbf{I}-\mathbf{P})\mathbf{y}] \\ &= \mathrm{trace}((\mathbf{I}-\mathbf{P})\mathrm{Cov}[\mathbf{y}]) + (\mathrm{E}[\mathbf{y}])^T(\mathbf{I}-\mathbf{P})\mathrm{E}[\mathbf{y}] \\ &= \mathrm{trace}((\mathbf{I}-\mathbf{P})\sigma^2\mathbf{I}) + (\mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2\mathrm{trace}(\mathbf{I}-\mathbf{P}) + 0 \\ &= (N-R)\sigma^2. \end{aligned}$$

- From this, we deduce that the following estimator is unbiased for σ^2:

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{N-R} = \frac{1}{N-R}\sum_{i=1}^{N}\widehat{\epsilon}_i^2.$$

Estimable functions

Example: One-way ANOVA with two groups

- To revisit our ANOVA example:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Can we estimate μ?
  - Put another way, is it the case that $\mathbf{a} = (1, 0, 0)^T \in \mathcal{R}(\mathbf{X})$?
  - No!

Estimable functions

Example: One-way ANOVA with two groups

- To revisit our ANOVA example:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Can we estimate $\mu + \gamma_1$?
  - Put another way, is it the case that $\mathbf{a} = (1, 1, 0)^T \in \mathcal{R}(\mathbf{X})$?
  - Yes!

Estimable functions

Example: One-way ANOVA with two groups

- To revisit our ANOVA example:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}$$

- Can we estimate $\mu - \gamma_2$?
  - Put another way, is it the case that $\mathbf{a} = (1, 0, -1)^T \in \mathcal{R}(\mathbf{X})$?
  - No!

Table of Contents

Revisiting the Gauss-Markov theorem

Reminder: The full-rank case

- When $\mathbf{X}$ was of full rank, we had a statement about the optimality of OLS under homoscedasticity ($\mathbf{a}^T\widehat{\boldsymbol{\beta}}$ was the unique BLUE of $\mathbf{a}^T\boldsymbol{\beta}$).
- If $\mathbf{X}$ is not of full rank, then the solution to the normal equations is not unique, so instead we make statements about $\widehat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, the projection of $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$.

Revisiting the Gauss-Markov theorem

Theorem 4.2: The Gauss-Markov theorem (version 2)

- Suppose that each of the following conditions is satisfied:
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
  - $\mathbf{X}$ may be rank-deficient ($\mathrm{rank}(\mathbf{X}) = R < K$).
  - $\mathrm{Cov}[\mathbf{y}] = \sigma^2\mathbf{I}$.
- Then, $\mathbf{a}^T\widehat{\mathbf{y}}$ is the unique estimate of $\mathbf{a}^T\mathbf{X}\boldsymbol{\beta}$ that achieves minimum variance among all unbiased linear estimators of $\mathbf{a}^T\mathbf{X}\boldsymbol{\beta}$.

- The proof is similar to that of the full-rank case.

Revisiting the Gauss-Markov theorem

Theorem 4.3: The Gauss-Markov theorem (version 3)

- Suppose that each of the following conditions is satisfied:
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
  - $\mathbf{X}$ may be rank-deficient ($\mathrm{rank}(\mathbf{X}) = R < K$).
  - $\mathrm{Cov}[\mathbf{y}] = \sigma^2\mathbf{I}$.
- If $\mathbf{a}^T\boldsymbol{\beta}$ is estimable, then $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$ is unique (i.e., the same for all $\widehat{\boldsymbol{\beta}}$ solving the normal equations), and is the BLUE of $\mathbf{a}^T\boldsymbol{\beta}$.

Theorem 4.3: Proof

- Proof of uniqueness: By Lemma 4.10, if $\mathbf{a}^T\boldsymbol{\beta}$ is estimable, then $\mathbf{a} = \mathbf{X}^T\mathbf{c}_*$ for a unique $\mathbf{c}_* \in \mathcal{C}(\mathbf{X})$. Therefore, $\mathbf{a}^T\widehat{\boldsymbol{\beta}} = \mathbf{c}_*^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{c}_*^T\widehat{\mathbf{y}}$, which is unique because $\widehat{\mathbf{y}}$ is unique.
- Proof of **BLUE**-ness: We know that $\mathbf{a}^T\widehat{\boldsymbol{\beta}} = \mathbf{c}_*^T\widehat{\mathbf{y}}$, which is the BLUE for $\mathbf{c}_*^T\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}$ by Theorem 4.2 (Gauss-Markov theorem, version 2).

Summary

So far:

- Ordinary least squares in the case of a rank-deficient design matrix.

Summary

Up next:

- Weighted least squares.