STAT 714

LINEAR STATISTICAL MODELS

Fall, 2010

Lecture Notes

Joshua M. Tebbs

Department of Statistics

The University of South Carolina

Contents

1 Examples of the General Linear Model

Complementary reading from Monahan: Chapter 1.

INTRODUCTION: **Linear models** are models that are linear in their parameters. The general form of a linear model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observed responses, $\mathbf{X}$ is an $n \times p$ (design) matrix of fixed constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed but unknown parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of (unobserved) random errors. The model is called a linear model because the mean of the response vector $\mathbf{Y}$ is linear in the unknown parameter $\boldsymbol{\beta}$.

SCOPE: Several models commonly used in statistics are examples of the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. These include, but are not limited to, linear regression models and analysis of variance (ANOVA) models. Regression models generally refer to those for which $\mathbf{X}$ is full rank, while ANOVA models refer to those for which $\mathbf{X}$ consists of zeros and ones.

GENERAL CLASSES OF LINEAR MODELS:

- **Model I:** *Least squares model*: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. This model makes no assumptions on $\boldsymbol{\epsilon}$. The parameter space is $\boldsymbol{\Theta} = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{R}^p\}$.

- **Model II:** *Gauss Markov model*: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The parameter space is $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}, \sigma^2) : (\boldsymbol{\beta}, \sigma^2) \in \mathcal{R}^p \times \mathcal{R}^+\}$.

- **Model III:** *Aitken model*: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, $\mathbf{V}$ known. The parameter space is $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}, \sigma^2) : (\boldsymbol{\beta}, \sigma^2) \in \mathcal{R}^p \times \mathcal{R}^+\}$.

- **Model IV:** *General linear mixed model*: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\boldsymbol{\theta})$. The parameter space is $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}, \boldsymbol{\theta}) : (\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathcal{R}^p \times \boldsymbol{\Omega}\}$, where $\boldsymbol{\Omega}$ is the set of all values of $\boldsymbol{\theta}$ for which $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite.

GAUSS MARKOV MODEL: Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. This model is treated extensively in Chapter 4. We now highlight special cases of this model.

Example 1.1. *One-sample problem.* Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance $\sigma^2 > 0$. If $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are iid with mean $E(\epsilon_i) = 0$ and common variance σ^2, we can write the GM model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times 1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\beta}_{1\times 1} = \mu, \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. □

Example 1.2. *Simple linear regression.* Consider the model where a response variable Y is linearly related to an independent variable x via

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where the ϵ_i are uncorrelated random variables with mean 0 and common variance $\sigma^2 > 0$. If $x_1, x_2, ..., x_n$ are fixed constants, measured without error, then this is a GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. □

Example 1.3. *Multiple linear regression.* Suppose that a response variable Y is linearly related to several independent variables, say, $x_1, x_2, ..., x_k$ via

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where ϵ_i are uncorrelated random variables with mean 0 and common variance $\sigma^2 > 0$. If the independent variables are fixed constants, measured without error, then this model is a special GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times p} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

and $p = k + 1$. Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. □

Example 1.4. *One-way ANOVA*. Consider an experiment that is performed to compare $a \geq 2$ treatments. For the ith treatment level, suppose that n_i experimental units are selected at random and assigned to the ith treatment. Consider the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where the random errors ϵ_{ij} are uncorrelated random variables with zero mean and common variance $\sigma^2 > 0$. If the a treatment effects $\alpha_1, \alpha_2, ..., \alpha_a$ are best regarded as fixed constants, then this model is a special case of the GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. To see this, note that with $n = \sum_{i=1}^{a} n_i$,

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{an_a} \end{pmatrix}, \quad \mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{1}_{n_a} \end{pmatrix}, \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix},$$

where $p = a + 1$ and $\boldsymbol{\epsilon}_{n\times 1} = (\epsilon_{11}, \epsilon_{12}, ..., \epsilon_{an_a})'$, and where $\mathbf{1}_{n_i}$ is an $n_i \times 1$ vector of ones and $\mathbf{0}_{n_i}$ is an $n_i \times 1$ vector of zeros. Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

NOTE: In Example 1.4, note that the first column of $\mathbf{X}$ is the sum of the last a columns; i.e., there is a linear dependence in the columns of $\mathbf{X}$. From results in linear algebra, we know that $\mathbf{X}$ is not of full column rank. In fact, the rank of $\mathbf{X}$ is $r = a$, one less than the number of columns $p = a + 1$. This is a common characteristic of ANOVA models; namely, their $\mathbf{X}$ matrices are not of full column rank. On the other hand, (linear) regression models are models of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is of full column rank; see Examples 1.2 and 1.3. □

Example 1.5. *Two-way nested ANOVA*. Consider an experiment with two factors, where one factor, say, Factor B, is **nested** within Factor A. In other words, every level of B appears with exactly one level of Factor A. A statistical model for this situation is

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b_i$, and $k = 1, 2, ..., n_{ij}$. In this model, μ denotes the overall mean, α_i represents the effect due to the ith level of A, and β_{ij} represents the effect of the jth level of B, nested within the ith level of A. If all parameters are fixed, and the random errors ϵ_{ijk} are uncorrelated random variables with zero mean and constant unknown variance $\sigma^2 > 0$, then this is a special GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For example, with $a = 3$, $b = 2$, and $n_{ij} = 2$, we have

$$\mathbf{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{311} \\ Y_{312} \\ Y_{321} \\ Y_{322} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \\ \beta_{31} \\ \beta_{32} \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, ..., \epsilon_{322})'$. Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The $\mathbf{X}$ matrix is not of full column rank. The rank of $\mathbf{X}$ is $r = 6$ and there are $p = 10$ columns. □

Example 1.6. *Two-way crossed ANOVA with interaction.* Consider an experiment with two factors (A and B), where Factor A has a levels and Factor B has b levels. In general, we say that factors A and B are **crossed** if every level of A occurs in combination with every level of B. Consider the two-factor (crossed) ANOVA model given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$, where the random errors ϵ_{ij} are uncorrelated random variables with zero mean and constant unknown variance $\sigma^2 > 0$. If all the parameters are fixed, this is a special GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For example, with $a = 3$, $b = 2$, and $n_{ij} = 3$,

$$\mathbf{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{321} \\ Y_{322} \\ Y_{323} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{31} \\ \gamma_{32} \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, ..., \epsilon_{323})'$. Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The $\mathbf{X}$ matrix is not of full column rank. The rank of $\mathbf{X}$ is $r = 6$ and there are $p = 12$ columns. □

Example 1.7. *Two-way crossed ANOVA without interaction.* Consider an experiment with two factors (A and B), where Factor A has a levels and Factor B has b levels. The two-way crossed model without interaction is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$, where the random errors ϵ_{ij} are uncorrelated random variables with zero mean and common variance $\sigma^2 > 0$. Note that no-interaction model is a special case of the interaction model in Example 1.6 when $H_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{32} = 0$ is true. That is, the no-interaction model is a **reduced** version of the interaction model. With $a = 3$, $b = 2$, and $n_{ij} = 3$ as before, we have

$$\mathbf{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{321} \\ Y_{322} \\ Y_{323} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, ..., \epsilon_{323})'$. Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The $\mathbf{X}$ matrix is not of full column rank. The rank of $\mathbf{X}$ is $r = 4$ and there are $p = 6$ columns. Also note that the design matrix for the no-interaction model is the same as the design matrix for the interaction model, except that the last 6 columns are removed. □

Example 1.8. *Analysis of covariance.* Consider an experiment to compare $a \geq 2$ treatments after adjusting for the effects of a covariate x. A model for the analysis of covariance is given by

$$Y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \epsilon_{ij},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., n_i$, where the random errors ϵ_{ij} are uncorrelated random variables with zero mean and common variance $\sigma^2 > 0$. In this model, μ represents the overall mean, α_i represents the (fixed) effect of receiving the ith treatment (disregarding the covariates), and β_i denotes the slope of the line that relates Y to x for the ith treatment. Note that this model allows the treatment slopes to be different. The x_{ij}'s are assumed to be fixed values measured without error.

NOTE: The analysis of covariance (ANCOVA) model is a special GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For example, with $a = 3$ and $n_1 = n_2 = n_3 = 3$, we have

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{12} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{13} & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{21} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{22} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{23} & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{31} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{32} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{33} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}.$$

Note that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The $\mathbf{X}$ matrix is not of full column rank. If there are no linear dependencies among the last 3 columns, the rank of $\mathbf{X}$ is $r = 6$ and there are $p = 7$ columns.

REDUCED MODEL: Consider the ANCOVA model in Example 1.8 which allows for unequal slopes. If $\beta_1 = \beta_2 = \cdots = \beta_a$; that is, all slopes are equal, then the ANCOVA model reduces to

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}.$$

That is, the common-slopes ANCOVA model is a **reduced** version of the model that allows for different slopes. Assuming the same error structure, this reduced ANCOVA model is also a special GM model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. With $a = 3$ and $n_1 = n_2 = n_3 = 3$, as before, we have

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} \\ 1 & 1 & 0 & 0 & x_{12} \\ 1 & 1 & 0 & 0 & x_{13} \\ 1 & 0 & 1 & 0 & x_{21} \\ 1 & 0 & 1 & 0 & x_{22} \\ 1 & 0 & 1 & 0 & x_{23} \\ 1 & 0 & 0 & 1 & x_{31} \\ 1 & 0 & 0 & 1 & x_{32} \\ 1 & 0 & 0 & 1 & x_{33} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}.$$

As long as at least one of the x_{ij}'s is different, the rank of $\mathbf{X}$ is $r = 4$ and there are $p = 5$ columns. □

GOAL: We now provide examples of linear models of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ that are not GM models.

TERMINOLOGY: A factor of classification is said to be **random** if it has an infinitely large number of levels and the levels included in the experiment can be viewed as a random sample from the population of possible levels.

Example 1.9. *One-way random effects ANOVA*. Consider the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where the treatment effects $\alpha_1, \alpha_2, ..., \alpha_a$ are best regarded as random; e.g., the a levels of the factor of interest are drawn from a large population of possible levels, and the random errors ϵ_{ij} are uncorrelated random variables with zero mean and common variance $\sigma^2 > 0$. For concreteness, let $a = 4$ and $n_{ij} = 3$. The model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ looks like

$$\begin{aligned}\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{41} \\ Y_{42} \\ Y_{43} \end{pmatrix} &= \mathbf{1}_{12}\mu + \underbrace{\begin{pmatrix} \mathbf{1}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{1}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{1}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{1}_3 \end{pmatrix}}_{= \mathbf{Z}_1} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}}_{= \boldsymbol{\epsilon}_1} + \underbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \\ \epsilon_{43} \end{pmatrix}}_{= \boldsymbol{\epsilon}_2} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2,\end{aligned}$$

where we identify $\mathbf{X} = \mathbf{1}_{12}$, $\boldsymbol{\beta} = \mu$, and $\boldsymbol{\epsilon} = \mathbf{Z}_1\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2$. This is not a GM model because

$$\text{cov}(\boldsymbol{\epsilon}) = \text{cov}(\mathbf{Z}_1\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2) = \mathbf{Z}_1\text{cov}(\boldsymbol{\epsilon}_1)\mathbf{Z}_1' + \text{cov}(\boldsymbol{\epsilon}_2) = \mathbf{Z}_1\text{cov}(\boldsymbol{\epsilon}_1)\mathbf{Z}_1' + \sigma^2\mathbf{I},$$

provided that the α_i's and the errors ϵ_{ij} are uncorrelated. Note that $\text{cov}(\boldsymbol{\epsilon}) \neq \sigma^2\mathbf{I}$. □

Example 1.10. *Two-factor mixed model.* Consider an experiment with two factors (A and B), where Factor A is fixed and has a levels and Factor B is random with b levels. A statistical model for this situation is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$. The α_i's are best regarded as fixed and the β_j's are best regarded as random. This model assumes no interaction.

APPLICATION: In a randomized block experiment, b blocks may have been selected randomly from a large collection of available blocks. If the goal is to make a statement about the large population of blocks (and not those b blocks in the experiment), then blocks are considered as random. The treatment effects $\alpha_1, \alpha_2, ..., \alpha_a$ are regarded as fixed constants if the a treatments are the only ones of interest.

NOTE: For concreteness, suppose that $a = 2$, $b = 4$, and $n_{ij} = 1$. We can write the model above as

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0}_4 \\ \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{1}_4 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}}_{= \mathbf{X}\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \mathbf{I}_4 \\ \mathbf{I}_4 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}}_{= \mathbf{Z}_1\boldsymbol{\epsilon}_1} + \underbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \end{pmatrix}}_{= \boldsymbol{\epsilon}_2}.$$

NOTE: If the α_i's are best regarded as **random** as well, then we have

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{pmatrix} = \mathbf{1}_8\mu + \underbrace{\begin{pmatrix} \mathbf{1}_4 & \mathbf{0}_4 \\ \mathbf{0}_4 & \mathbf{1}_4 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}}_{= \mathbf{Z}_1\boldsymbol{\epsilon}_1} + \underbrace{\begin{pmatrix} \mathbf{I}_4 \\ \mathbf{I}_4 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}}_{= \mathbf{Z}_2\boldsymbol{\epsilon}_2} + \underbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \end{pmatrix}}_{= \boldsymbol{\epsilon}_3}.$$

This model is also known as a **random effects** or **variance component** model. □

GENERAL FORM: A **linear mixed model** can be expressed generally as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\epsilon}_1 + \mathbf{Z}_2\boldsymbol{\epsilon}_2 + \cdots + \mathbf{Z}_k\boldsymbol{\epsilon}_k,$$

where $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k$ are known matrices (typically $\mathbf{Z}_k = \mathbf{I}_k$) and $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_k$ are uncorrelated random vectors with uncorrelated components.

Example 1.11. *Time series models.* When measurements are taken over time, the GM model may not be appropriate because observations are likely correlated. A linear model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, $\mathbf{V}$ known, may be more appropriate. The general form of $\mathbf{V}$ is chosen to model the correlation of the observed responses. For example, consider the statistical model

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t,$$

for $t = 1, 2, ..., n$, where $\epsilon_t = \rho\epsilon_{t-1} + a_t$, $a_t \sim \text{iid}\,\mathcal{N}(0, \sigma^2)$, and $|\rho| < 1$ (this is a stationarity condition). This is called a simple **linear trend model** where the error process $\{\epsilon_t : t = 1, 2, ..., n\}$ follows an autoregressive model of order 1, AR(1). It is easy to show that $E(\epsilon_t) = 0$, for all t, and that $\text{cov}(\epsilon_t, \epsilon_s) = \sigma^2\rho^{|t-s|}$, for all t and s. Therefore, if $n = 5$,

$$\mathbf{V} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}. \ \square$$

Example 1.12. *Random coefficient models.* Suppose that t measurements are taken (over time) on n individuals and consider the model

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \epsilon_{ij},$$

for $i = 1, 2, ..., n$ and $j = 1, 2, ..., t$; that is, the different $p \times 1$ regression parameters $\boldsymbol{\beta}_i$ are "subject-specific." If the individuals are considered to be a random sample, then we can treat $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_n$ as iid random vectors with mean $\boldsymbol{\beta}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, say. We can write this model as

$$\begin{aligned} Y_{ij} &= \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \epsilon_{ij} \\ &= \underbrace{\mathbf{x}'_{ij}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\mathbf{x}'_{ij}(\boldsymbol{\beta}_i - \boldsymbol{\beta}) + \epsilon_{ij}}_{\text{random}}. \end{aligned}$$

If the $\boldsymbol{\beta}_i$'s are independent of the ϵ_{ij}'s, note that

$$\text{var}(Y_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\Sigma}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{x}_{ij} + \sigma^2 \neq \sigma^2. \ \square$$

Example 1.13. *Measurement error models.* Consider the statistical model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma_\epsilon^2)$. The X_i's are not observed exactly; instead, they are measured with non-negligible error so that

$$W_i = X_i + U_i,$$

where $U_i \sim$ iid $\mathcal{N}(0, \sigma_U^2)$. Here,

$$\begin{aligned} \text{Observed data:} &\quad (Y_i, W_i) \\ \text{Not observed:} &\quad (X_i, \epsilon_i, U_i) \\ \text{Unknown parameters:} &\quad (\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_U^2). \end{aligned}$$

As a frame of reference, suppose that Y is a continuous measurement of lung function in small children and that X denotes the long-term exposure to NO_2. It is unlikely that X can be measured exactly; instead, the surrogate W, the amount of NO_2 recorded at a clinic visit, is more likely to be observed. Note that the model above can be rewritten as

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(W_i - U_i) + \epsilon_i \\ &= \beta_0 + \beta_1 W_i + \underbrace{(\epsilon_i - \beta_1 U_i)}_{= \epsilon_i^*}. \end{aligned}$$

Because the W_i's are not fixed in advance, we would at least need $E(\epsilon_i^*|W_i) = 0$ for this to be a GM linear model. However, note that

$$\begin{aligned} E(\epsilon_i^*|W_i) &= E(\epsilon_i - \beta_1 U_i | X_i + U_i) \\ &= E(\epsilon_i | X_i + U_i) - \beta_1 E(U_i | X_i + U_i). \end{aligned}$$

The first term is zero if ϵ_i is independent of both X_i and U_i. The second term generally is not zero (unless $\beta_1 = 0$, of course) because U_i and $X_i + U_i$ are correlated. Therefore, this can not be a GM model. □

2 The Linear Least Squares Problem

Complementary reading from Monahan: Chapter 2 (except Section 2.4).

INTRODUCTION: Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observed responses, $\mathbf{X}$ is an $n \times p$ matrix of fixed constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed but unknown parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors. If $E(\boldsymbol{\epsilon}) = \mathbf{0}$, then

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

Since $\boldsymbol{\beta}$ is unknown, all we really know is that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$. To estimate $E(\mathbf{Y})$, it seems natural to take the vector in $\mathcal{C}(\mathbf{X})$ that is closest to $\mathbf{Y}$.

2.1 Least squares estimation

DEFINITION: An estimate $\widehat{\boldsymbol{\beta}}$ is a **least squares estimate** of $\boldsymbol{\beta}$ if $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is the vector in $\mathcal{C}(\mathbf{X})$ that is closest to $\mathbf{Y}$. In other words, $\widehat{\boldsymbol{\beta}}$ is a least squares estimate of $\boldsymbol{\beta}$ if

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathcal{R}^p} \ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

LEAST SQUARES: Let $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ and define the **error sum of squares**

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

the squared distance from $\mathbf{Y}$ to $\mathbf{X}\boldsymbol{\beta}$. The point where $Q(\boldsymbol{\beta})$ is minimized satisfies

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad \text{or, in other words,} \quad \begin{pmatrix} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

This minimization problem can be tackled either algebraically or geometrically.

Result 2.1. Let $\mathbf{a}$ and $\mathbf{b}$ be $p \times 1$ vectors and $\mathbf{A}$ be a $p \times p$ matrix of constants. Then

$$\frac{\partial \mathbf{a}'\mathbf{b}}{\partial \mathbf{b}} = \mathbf{a} \quad \text{and} \quad \frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = (\mathbf{A} + \mathbf{A}')\mathbf{b}.$$

Proof. See Monahan, pp 14. □

NOTE: In Result 2.1, note that

$$\frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = 2\mathbf{A}\mathbf{b}$$

if $\mathbf{A}$ is symmetric.

NORMAL EQUATIONS: Simple calculations show that

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Using Result 2.1, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta},$$

because $\mathbf{X}'\mathbf{X}$ is symmetric. Setting this expression equal to $\mathbf{0}$ and rearranging gives

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

These are the **normal equations**. If $\mathbf{X}'\mathbf{X}$ is nonsingular, then the unique least squares estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

When $\mathbf{X}'\mathbf{X}$ is singular, which can happen in ANOVA models (see Chapter 1), there can be multiple solutions to the normal equations. Having already proved algebraically that the normal equations are consistent, we know that the general form of the least squares solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} + [\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}]\mathbf{z},$$

for $\mathbf{z} \in \mathcal{R}^p$, where $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$.

2.2 Geometric considerations

CONSISTENCY: Recall that a linear system $\mathbf{Ax} = \mathbf{c}$ is **consistent** if there exists an $\mathbf{x}^*$ such that $\mathbf{Ax}^* = \mathbf{c}$; that is, if $\mathbf{c} \in \mathcal{C}(\mathbf{A})$. Applying this definition to

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y},$$

the normal equations are consistent if $\mathbf{X}'\mathbf{Y} \in \mathcal{C}(\mathbf{X}'\mathbf{X})$. Clearly, $\mathbf{X}'\mathbf{Y} \in \mathcal{C}(\mathbf{X}')$. Thus, we'll be able to establish consistency (geometrically) if we can show that $\mathcal{C}(\mathbf{X}'\mathbf{X}) = \mathcal{C}(\mathbf{X}')$.

Result 2.2. $\mathcal{N}(\mathbf{X}'\mathbf{X}) = \mathcal{N}(\mathbf{X})$.
Proof. Suppose that $\mathbf{w} \in \mathcal{N}(\mathbf{X})$. Then $\mathbf{Xw} = \mathbf{0}$ and $\mathbf{X}'\mathbf{Xw} = \mathbf{0}$ so that $\mathbf{w} \in \mathcal{N}(\mathbf{X}'\mathbf{X})$. Suppose that $\mathbf{w} \in \mathcal{N}(\mathbf{X}'\mathbf{X})$. Then $\mathbf{X}'\mathbf{Xw} = \mathbf{0}$ and $\mathbf{w}'\mathbf{X}'\mathbf{Xw} = 0$. Thus, $||\mathbf{Xw}||^2 = 0$ which implies that $\mathbf{Xw} = \mathbf{0}$; i.e., $\mathbf{w} \in \mathcal{N}(\mathbf{X})$. □

Result 2.3. Suppose that $\mathcal{S}_1$ and $\mathcal{T}_1$ are orthogonal complements, as well as $\mathcal{S}_2$ and $\mathcal{T}_2$. If $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\mathcal{T}_2 \subseteq \mathcal{T}_1$.
Proof. See Monahan, pp 244. □

CONSISTENCY: We use the previous two results to show that $\mathcal{C}(\mathbf{X}'\mathbf{X}) = \mathcal{C}(\mathbf{X}')$. Take $\mathcal{S}_1 = \mathcal{N}(\mathbf{X}'\mathbf{X})$, $\mathcal{T}_1 = \mathcal{C}(\mathbf{X}'\mathbf{X})$, $\mathcal{S}_2 = \mathcal{N}(\mathbf{X})$, and $\mathcal{T}_2 = \mathcal{C}(\mathbf{X}')$. We know that $\mathcal{S}_1$ and $\mathcal{T}_1$ ($\mathcal{S}_2$ and $\mathcal{T}_2$) are orthogonal complements. Because $\mathcal{N}(\mathbf{X}'\mathbf{X}) \subseteq \mathcal{N}(\mathbf{X})$, the last result guarantees $\mathcal{C}(\mathbf{X}') \subseteq \mathcal{C}(\mathbf{X}'\mathbf{X})$. But, $\mathcal{C}(\mathbf{X}'\mathbf{X}) \subseteq \mathcal{C}(\mathbf{X}')$ trivially, so we're done. Note also

$$\mathcal{C}(\mathbf{X}'\mathbf{X}) = \mathcal{C}(\mathbf{X}') \Longrightarrow r(\mathbf{X}'\mathbf{X}) = r(\mathbf{X}') = r(\mathbf{X}). \ \square$$

NOTE: We now state a result that characterizes all solutions to the normal equations.

Result 2.4. $Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is minimized at $\widehat{\boldsymbol{\beta}}$ if and only if $\widehat{\boldsymbol{\beta}}$ is a solution to the normal equations.
Proof. ($\Longleftarrow$) Suppose that $\widehat{\boldsymbol{\beta}}$ is a solution to the normal equations. Then,

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}), \end{aligned}$$

since the cross product term $2(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})=0$; verify this using the fact that $\widehat{\boldsymbol{\beta}}$ solves the normal equations. Thus, we have shown that $Q(\boldsymbol{\beta})=Q(\widehat{\boldsymbol{\beta}})+\mathbf{z}'\mathbf{z}$, where $\mathbf{z}=\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}$. Therefore, $Q(\boldsymbol{\beta})\geq Q(\widehat{\boldsymbol{\beta}})$ for all $\boldsymbol{\beta}$ and, hence, $\widehat{\boldsymbol{\beta}}$ minimizes $Q(\boldsymbol{\beta})$. ($\Longrightarrow$) Now, suppose that $\widetilde{\boldsymbol{\beta}}$ minimizes $Q(\boldsymbol{\beta})$. We already know that $Q(\widetilde{\boldsymbol{\beta}})\geq Q(\widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$, by assumption, but also $Q(\widetilde{\boldsymbol{\beta}})\leq Q(\widehat{\boldsymbol{\beta}})$ because $\widetilde{\boldsymbol{\beta}}$ minimizes $Q(\boldsymbol{\beta})$. Thus, $Q(\widetilde{\boldsymbol{\beta}})=Q(\widehat{\boldsymbol{\beta}})$. But because $Q(\widetilde{\boldsymbol{\beta}})=Q(\widehat{\boldsymbol{\beta}})+\mathbf{z}'\mathbf{z}$, where $\mathbf{z}=\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\widetilde{\boldsymbol{\beta}}$, it must be true that $\mathbf{z}=\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\widetilde{\boldsymbol{\beta}}=\mathbf{0}$; that is, $\mathbf{X}\widehat{\boldsymbol{\beta}}=\mathbf{X}\widetilde{\boldsymbol{\beta}}$. Thus,

$$\mathbf{X}'\mathbf{X}\widetilde{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{Y},$$

since $\widehat{\boldsymbol{\beta}}$ is a solution to the normal equations. This shows that $\widetilde{\boldsymbol{\beta}}$ is also solution to the normal equations. □

INVARIANCE: In proving the last result, we have discovered a very important fact; namely, if $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ both solve the normal equations, then $\mathbf{X}\widehat{\boldsymbol{\beta}}=\mathbf{X}\widetilde{\boldsymbol{\beta}}$. In other words, $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is **invariant** to the choice of $\widehat{\boldsymbol{\beta}}$.

NOTE: The following result ties least squares estimation to the notion of a perpendicular projection matrix. It also produces a general formula for the matrix.

Result 2.5. An estimate $\widehat{\boldsymbol{\beta}}$ is a least squares estimate if and only if $\mathbf{X}\widehat{\boldsymbol{\beta}}=\mathbf{M}\mathbf{Y}$, where $\mathbf{M}$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$.

Proof. We will show that

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=(\mathbf{Y}-\mathbf{M}\mathbf{Y})'(\mathbf{Y}-\mathbf{M}\mathbf{Y})+(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}).$$

Both terms on the right hand side are nonnegative, and the first term does not involve $\boldsymbol{\beta}$. Thus, $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$ is minimized by minimizing $(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$, the squared distance between $\mathbf{M}\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta}$. This distance is zero if and only if $\mathbf{M}\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}$, which proves the result. Now to show the above equation:

$$\begin{aligned}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) &= (\mathbf{Y}-\mathbf{M}\mathbf{Y}+\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{M}\mathbf{Y}+\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\\ &= (\mathbf{Y}-\mathbf{M}\mathbf{Y})'(\mathbf{Y}-\mathbf{M}\mathbf{Y})+\underbrace{(\mathbf{Y}-\mathbf{M}\mathbf{Y})'(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}_{(*)}\\ &\quad+\underbrace{(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{M}\mathbf{Y})}_{(**)}+(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{M}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

It suffices to show that (∗) and (∗∗) are zero. To show that (∗) is zero, note that

$$(\mathbf{Y}-\mathbf{MY})'(\mathbf{MY}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'(\mathbf{I}-\mathbf{M})(\mathbf{MY}-\mathbf{X}\boldsymbol{\beta}) = [(\mathbf{I}-\mathbf{M})\mathbf{Y}]'(\mathbf{MY}-\mathbf{X}\boldsymbol{\beta}) = 0,$$

because $(\mathbf{I}-\mathbf{M})\mathbf{Y} \in \mathcal{N}(\mathbf{X}')$ and $\mathbf{MY}-\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$. Similarly, $(**) = 0$ as well. □

Result 2.6. The perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$ is given by

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

Proof. We know that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ is a solution to the normal equations, so it is a least squares estimate. But, by Result 2.5, we know $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{MY}$. Because perpendicular projection matrices are unique, $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ as claimed. □

NOTATION: Monahan uses $\mathbf{P}_\mathbf{X}$ to denote the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. We will henceforth do the same; that is,

$$\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'.$$

PROPERTIES: Let $\mathbf{P}_\mathbf{X}$ denote the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. Then

(a) $\mathbf{P}_\mathbf{X}$ is idempotent

(b) $\mathbf{P}_\mathbf{X}$ projects onto $\mathcal{C}(\mathbf{X})$

(c) $\mathbf{P}_\mathbf{X}$ is invariant to the choice of $(\mathbf{X}'\mathbf{X})^{-}$

(d) $\mathbf{P}_\mathbf{X}$ is symmetric

(e) $\mathbf{P}_\mathbf{X}$ is unique.

We have already proven (a), (b), (d), and (e); see Matrix Algebra Review 5. Part (c) must be true; otherwise, part (e) would not hold. However, we can prove (c) more rigorously.

Result 2.7. If $(\mathbf{X}'\mathbf{X})^{-}_1$ and $(\mathbf{X}'\mathbf{X})^{-}_2$ are generalized inverses of $\mathbf{X}'\mathbf{X}$, then

1. $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}_1\mathbf{X}'\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}_2\mathbf{X}'\mathbf{X} = \mathbf{X}$

2. $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}_1\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}_2\mathbf{X}'$.

Proof. For $\mathbf{v} \in \mathcal{R}^n$, let $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1 \in \mathcal{C}(\mathbf{X})$ and $\mathbf{v}_2 \perp \mathcal{C}(\mathbf{X})$. Since $\mathbf{v}_1 \in \mathcal{C}(\mathbf{X})$, we know that $\mathbf{v}_1 = \mathbf{X}\mathbf{d}$, for some vector $\mathbf{d}$. Then,

$$\mathbf{v}'\mathbf{X}(\mathbf{X}'\mathbf{X})_1^-\mathbf{X}'\mathbf{X} = \mathbf{v}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})_1^-\mathbf{X}'\mathbf{X} = \mathbf{d}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})_1^-\mathbf{X}'\mathbf{X} = \mathbf{d}'\mathbf{X}'\mathbf{X} = \mathbf{v}'\mathbf{X},$$

since $\mathbf{v}_2 \perp \mathcal{C}(\mathbf{X})$. Since $\mathbf{v}$ and $(\mathbf{X}'\mathbf{X})_1^-$ were arbitrary, we have shown the first part. To show the second part, note that

$$\mathbf{X}(\mathbf{X}'\mathbf{X})_1^-\mathbf{X}'\mathbf{v} = \mathbf{X}(\mathbf{X}'\mathbf{X})_1^-\mathbf{X}'\mathbf{X}\mathbf{d} = \mathbf{X}(\mathbf{X}'\mathbf{X})_2^-\mathbf{X}'\mathbf{X}\mathbf{d} = \mathbf{X}(\mathbf{X}'\mathbf{X})_2^-\mathbf{X}'\mathbf{v}.$$

Since $\mathbf{v}$ is arbitrary, the second part follows as well. □

Result 2.8. Suppose $\mathbf{X}$ is $n \times p$ with rank $r \leq p$, and let $\mathbf{P_X}$ be the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. Then $r(\mathbf{P_X}) = r(\mathbf{X}) = r$ and $r(\mathbf{I} - \mathbf{P_X}) = n - r$.
Proof. Note that $\mathbf{P_X}$ is $n \times n$. We know that $\mathcal{C}(\mathbf{P_X}) = \mathcal{C}(\mathbf{X})$, so the first part is obvious. To show the second part, recall that $\mathbf{I} - \mathbf{P_X}$ is the perpendicular projection matrix onto $\mathcal{N}(\mathbf{X}')$, so it is idempotent. Thus,

$$r(\mathbf{I} - \mathbf{P_X}) = tr(\mathbf{I} - \mathbf{P_X}) = tr(\mathbf{I}) - tr(\mathbf{P_X}) = n - r(\mathbf{P_X}) = n - r,$$

because the trace operator is linear and because $\mathbf{P_X}$ is idempotent as well. □

SUMMARY: Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$; in what follows, the $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ assumption is not needed. We have shown that a least squares estimate of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}.$$

This solution is not unique (unless $\mathbf{X}'\mathbf{X}$ is nonsingular). However,

$$\mathbf{P_X}\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} \equiv \widehat{\mathbf{Y}}$$

is unique. We call $\widehat{\mathbf{Y}}$ the vector of **fitted values**. Geometrically, $\widehat{\mathbf{Y}}$ is the point in $\mathcal{C}(\mathbf{X})$ that is closest to $\mathbf{Y}$. Now, recall that $\mathbf{I} - \mathbf{P_X}$ is the perpendicular projection matrix onto $\mathcal{N}(\mathbf{X}')$. Note that

$$(\mathbf{I} - \mathbf{P_X})\mathbf{Y} = \mathbf{Y} - \mathbf{P_X}\mathbf{Y} = \mathbf{Y} - \widehat{\mathbf{Y}} \equiv \widehat{\mathbf{e}}.$$

We call $\widehat{\mathbf{e}}$ the vector of **residuals**. Note that $\widehat{\mathbf{e}} \in \mathcal{N}(\mathbf{X}')$. Because $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}')$ are orthogonal complements, we know that $\mathbf{Y}$ can be uniquely decomposed as

$$\mathbf{Y} = \widehat{\mathbf{Y}} + \widehat{\mathbf{e}}.$$

We also know that $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{e}}$ are orthogonal vectors. Finally, note that

$$\begin{aligned}\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{I}\mathbf{Y} &= \mathbf{Y}'(\mathbf{P_X} + \mathbf{I} - \mathbf{P_X})\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{P_X}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{P_X}\mathbf{P_X}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})(\mathbf{I} - \mathbf{P_X})\mathbf{Y} \\ &= \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \widehat{\mathbf{e}}'\widehat{\mathbf{e}},\end{aligned}$$

since $\mathbf{P_X}$ and $\mathbf{I}-\mathbf{P_X}$ are both symmetric and idempotent; i.e., they are both perpendicular projection matrices (but onto orthogonal spaces). This orthogonal decomposition of $\mathbf{Y}'\mathbf{Y}$ is often given in a tabular display called an **analysis of variance** (**ANOVA**) table.

ANOVA TABLE: Suppose that $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$ with rank $r \leq p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\epsilon}$ is $n \times 1$. An ANOVA table looks like

Source	df	SS
Model	r	$\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \mathbf{Y}'\mathbf{P_X}\mathbf{Y}$
Residual	$n-r$	$\widehat{\mathbf{e}}'\widehat{\mathbf{e}} = \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$
Total	n	$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{I}\mathbf{Y}$

It is interesting to note that the sum of squares column, abbreviated "SS," catalogues 3 quadratic forms, $\mathbf{Y}'\mathbf{P_X}\mathbf{Y}$, $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X}\mathbf{Y})$, and $\mathbf{Y}'\mathbf{I}\mathbf{Y}$. The degrees of freedom column, abbreviated "df," catalogues the ranks of the associated quadratic form matrices; i.e.,

$$\begin{aligned} r(\mathbf{P_X}) &= r \\ r(\mathbf{I} - \mathbf{P_X}) &= n - r \\ r(\mathbf{I}) &= n. \end{aligned}$$

The quantity $\mathbf{Y}'\mathbf{P_X}\mathbf{Y}$ is called the (uncorrected) **model** sum of squares, $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$ is called the **residual** sum of squares, and $\mathbf{Y}'\mathbf{Y}$ is called the (uncorrected) **total** sum of squares.

NOTE: The following "visualization" analogy is taken liberally from Christensen (2002).

VISUALIZATION: One can think about the geometry of least squares estimation in three dimensions (i.e., when $n = 3$). Consider your kitchen table and take one corner of the table to be the origin. Take $\mathcal{C}(\mathbf{X})$ as the two dimensional subspace determined by the surface of the table, and let $\mathbf{Y}$ be any vector originating at the origin; i.e., any point in $\mathcal{R}^3$. The linear model says that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, which just says that $E(\mathbf{Y})$ is somewhere on the table. The least squares estimate $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P_X Y}$ is the perpendicular projection of $\mathbf{Y}$ onto the surface of the table. The residual vector $\widehat{\mathbf{e}} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y}$ is the vector starting at the origin, perpendicular to the surface of the table, that reaches the same height as $\mathbf{Y}$. Another way to think of the residual vector is to first connect $\mathbf{Y}$ and $\mathbf{P_X Y}$ with a line segment (that is perpendicular to the surface of the table). Then, shift the line segment along the surface (keeping it perpendicular) until the line segment has one end at the origin. The residual vector $\widehat{\mathbf{e}}$ is the perpendicular projection of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{I} - \mathbf{P_X}) = \mathcal{N}(\mathbf{X}')$; that is, the projection onto the orthogonal complement of the table surface. The orthogonal complement $\mathcal{C}(\mathbf{I} - \mathbf{P_X})$ is the one-dimensional space in the vertical direction that goes through the origin. Once you have these vectors in place, sums of squares arise from Pythagorean's Theorem.

A SIMPLE PPM: Suppose $Y_1, Y_2, ..., Y_n$ are iid with mean $E(Y_i) = \mu$. In terms of the general linear model, we can write $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \mu, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$ is given by

$$\mathbf{P_1} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-}\mathbf{1}' = n^{-1}\mathbf{1}\mathbf{1}' = n^{-1}\mathbf{J},$$

where $\mathbf{J}$ is the $n \times n$ matrix of ones. Note that

$$\mathbf{P_1 Y} = n^{-1}\mathbf{J}\mathbf{Y} = \overline{Y}\mathbf{1},$$

where $\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i$. The perpendicular projection matrix $\mathbf{P_1}$ projects $\mathbf{Y}$ onto the space

$$\mathcal{C}(\mathbf{P_1}) = \{\mathbf{z} \in \mathcal{R}^n : \mathbf{z} = (a, a, ..., a)';\ a \in \mathcal{R}\}.$$

Note that $r(\mathbf{P_1}) = 1$. Note also that

$$(\mathbf{I} - \mathbf{P_1})\mathbf{Y} = \mathbf{Y} - \mathbf{P_1}\mathbf{Y} = \mathbf{Y} - \overline{Y}\mathbf{1} = \begin{pmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{pmatrix},$$

the vector which contains the deviations from the mean. The perpendicular projection matrix $\mathbf{I} - \mathbf{P_1}$ projects $\mathbf{Y}$ onto

$$\mathcal{C}(\mathbf{I} - \mathbf{P_1}) = \left\{\mathbf{z} \in \mathcal{R}^n : \mathbf{z} = (a_1, a_2, ..., a_n)';\ a_i \in \mathcal{R},\ \sum_{i=1}^{n} a_i = 0\right\}.$$

Note that $r(\mathbf{I} - \mathbf{P_1}) = n - 1$.

REMARK: The matrix $\mathbf{P_1}$ plays an important role in linear models, and here is why. Most linear models, when written out in non-matrix notation, contain an **intercept term**. For example, in simple linear regression,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

or in ANOVA-type models like

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

the intercept terms are β_0 and μ, respectively. In the corresponding design matrices, the first column of $\mathbf{X}$ is $\mathbf{1}$. If we discard the "other" terms like $\beta_1 x_i$ and $\alpha_i + \beta_j + \gamma_{ij}$ in the models above, then we have a reduced model of the form $Y_i = \mu + \epsilon_i$; that is, a model that relates Y_i to its overall mean, or, in matrix notation $\mathbf{Y} = \mathbf{1}\mu + \boldsymbol{\epsilon}$. The perpendicular projection matrix onto $\mathcal{C}(\mathbf{1})$ is $\mathbf{P_1}$ and

$$\mathbf{Y}'\mathbf{P_1}\mathbf{Y} = \mathbf{Y}'\mathbf{P_1}\mathbf{P_1}\mathbf{Y} = (\mathbf{P_1}\mathbf{Y})'(\mathbf{P_1}\mathbf{Y}) = n\overline{Y}^2.$$

This is the model sum of squares for the model $Y_i = \mu + \epsilon_i$; that is, $\mathbf{Y}'\mathbf{P_1Y}$ is the sum of squares that arises from fitting the overall mean μ. Now, consider a general linear model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and suppose that the first column of $\mathbf{X}$ is $\mathbf{1}$. In general, we know that

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{IY} = \mathbf{Y}'\mathbf{P_X Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}.$$

Subtracting $\mathbf{Y}'\mathbf{P_1Y}$ from both sides, we get

$$\mathbf{Y}'(\mathbf{I} - \mathbf{P_1})\mathbf{Y} = \mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}.$$

The quantity $\mathbf{Y}'(\mathbf{I}-\mathbf{P_1})\mathbf{Y}$ is called the **corrected total** sum of squares and the quantity $\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}$ is called the **corrected model** sum of squares. The term "corrected" is understood to mean that we have removed the effects of "fitting the mean." This is important because this is the sum of squares breakdown that is commonly used; i.e.,

Source	df	SS
Model (Corrected)	$r - 1$	$\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}$
Residual	$n - r$	$\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$
Total (Corrected)	$n - 1$	$\mathbf{Y}'(\mathbf{I} - \mathbf{P_1})\mathbf{Y}$

In ANOVA models, the corrected model sum of squares $\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}$ is often broken down further into smaller components which correspond to different parts; e.g., orthogonal contrasts, main effects, interaction terms, etc. Finally, the degrees of freedom are simply the corresponding ranks of $\mathbf{P_X} - \mathbf{P_1}$, $\mathbf{I} - \mathbf{P_X}$, and $\mathbf{I} - \mathbf{P_1}$.

NOTE: In the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the residual vector from the least squares fit $\widehat{\mathbf{e}} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y} \in \mathcal{N}(\mathbf{X}')$, so $\widehat{\mathbf{e}}'\mathbf{X} = \mathbf{0}$; that is, the residuals in a least squares fit are orthogonal to the columns of $\mathbf{X}$, since the columns of $\mathbf{X}$ are in $\mathcal{C}(\mathbf{X})$. Note that if $\mathbf{1} \in \mathcal{C}(\mathbf{X})$, which is true of all linear models with an intercept term, then

$$\widehat{\mathbf{e}}'\mathbf{1} = \sum_{i=1}^{n} \widehat{e}_i = 0,$$

that is, the sum of the residuals from a least squares fit is zero. This is not necessarily true of models for which $\mathbf{1} \notin \mathcal{C}(\mathbf{X})$.

Result 2.9. If $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$, then $\mathbf{P_X} - \mathbf{P_W}$ is the perpendicular projection matrix onto $\mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}]$.

Proof. It suffices to show that (a) $\mathbf{P_X} - \mathbf{P_W}$ is symmetric and idempotent and that (b) $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) = \mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}]$. First note that $\mathbf{P_X P_W} = \mathbf{P_W}$ because the columns of $\mathbf{P_W}$ are in $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. By symmetry, $\mathbf{P_W P_X} = \mathbf{P_W}$. Now,

$$\begin{aligned}(\mathbf{P_X} - \mathbf{P_W})(\mathbf{P_X} - \mathbf{P_W}) &= \mathbf{P}^2_{\mathbf{X}} - \mathbf{P_X P_W} - \mathbf{P_W P_X} + \mathbf{P}^2_{\mathbf{W}} \\ &= \mathbf{P_X} - \mathbf{P_W} - \mathbf{P_W} + \mathbf{P_W} = \mathbf{P_X} - \mathbf{P_W}.\end{aligned}$$

Thus, $\mathbf{P_X} - \mathbf{P_W}$ is idempotent. Also, $(\mathbf{P_X} - \mathbf{P_W})' = \mathbf{P}'_{\mathbf{X}} - \mathbf{P}'_{\mathbf{W}} = \mathbf{P_X} - \mathbf{P_W}$, so $\mathbf{P_X} - \mathbf{P_W}$ is symmetric. Thus, $\mathbf{P_X} - \mathbf{P_W}$ is a perpendicular projection matrix onto $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W})$. Suppose that $\mathbf{v} \in \mathcal{C}(\mathbf{P_X} - \mathbf{P_W})$; i.e., $\mathbf{v} = (\mathbf{P_X} - \mathbf{P_W})\mathbf{d}$, for some $\mathbf{d}$. Write $\mathbf{d} = \mathbf{d}_1 + \mathbf{d}_2$, where $\mathbf{d}_1 \in \mathcal{C}(\mathbf{X})$ and $\mathbf{d}_2 \in \mathcal{N}(\mathbf{X}')$; that is, $\mathbf{d}_1 = \mathbf{Xa}$, for some $\mathbf{a}$, and $\mathbf{X}'\mathbf{d}_2 = \mathbf{0}$. Then,

$$\begin{aligned}\mathbf{v} &= (\mathbf{P_X} - \mathbf{P_W})(\mathbf{d}_1 + \mathbf{d}_2) \\ &= (\mathbf{P_X} - \mathbf{P_W})(\mathbf{Xa} + \mathbf{d}_2) \\ &= \mathbf{P_X Xa} + \mathbf{P_X d}_2 - \mathbf{P_W Xa} - \mathbf{P_W d}_2 \\ &= \mathbf{Xa} + \mathbf{0} - \mathbf{P_W Xa} - \mathbf{0} \\ &= (\mathbf{I} - \mathbf{P_W})\mathbf{Xa} \in \mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}].\end{aligned}$$

Thus, $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) \subseteq \mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}]$. Now, suppose that $\mathbf{w} \in \mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}]$. Then $\mathbf{w} = (\mathbf{I} - \mathbf{P_W})\mathbf{Xc}$, for some $\mathbf{c}$. Thus,

$$\mathbf{w} = \mathbf{Xc} - \mathbf{P_W Xc} = \mathbf{P_X Xc} - \mathbf{P_W Xc} = (\mathbf{P_X} - \mathbf{P_W})\mathbf{Xc} \in \mathcal{C}(\mathbf{P_X} - \mathbf{P_W}).$$

This shows that $\mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}] \subseteq \mathcal{C}(\mathbf{P_X} - \mathbf{P_W})$. □

TERMINOLOGY: Suppose that $\mathcal{V}$ is a vector space and that $\mathcal{S}$ is a subspace of $\mathcal{V}$; i.e., $\mathcal{S} \subset \mathcal{V}$. The subspace

$$\mathcal{S}^{\perp}_{\mathcal{V}} = \{\mathbf{z} \in \mathcal{V} : \mathbf{z} \perp \mathcal{S}\}$$

is called the **orthogonal complement** of $\mathcal{S}$ with respect to $\mathcal{V}$. If $\mathcal{V} = \mathcal{R}^n$, then $\mathcal{S}^{\perp}_{\mathcal{V}} = \mathcal{S}^{\perp}$ is simply referred to as the orthogonal complement of $\mathcal{S}$.

Result 2.10. If $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$, then $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) = \mathcal{C}[(\mathbf{I} - \mathbf{P_W})\mathbf{X}]$ is the orthogonal complement of $\mathcal{C}(\mathbf{P_W})$ with respect to $\mathcal{C}(\mathbf{P_X})$; that is,

$$\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) = \mathcal{C}(\mathbf{P_W})^{\perp}_{\mathcal{C}(\mathbf{P_X})}.$$

Proof. $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) \perp \mathcal{C}(\mathbf{P_W})$ because $(\mathbf{P_X} - \mathbf{P_W})\mathbf{P_W} = \mathbf{P_X}\mathbf{P_W} - \mathbf{P}^2_{\mathbf{W}} = \mathbf{P_W} - \mathbf{P_W} = \mathbf{0}$. Because $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) \subset \mathcal{C}(\mathbf{P_X})$, $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W})$ is contained in the orthogonal complement of $\mathcal{C}(\mathbf{P_W})$ with respect to $\mathcal{C}(\mathbf{P_X})$. Now suppose that $\mathbf{v} \in \mathcal{C}(\mathbf{P_X})$ and $\mathbf{v} \perp \mathcal{C}(\mathbf{P_W})$. Then,

$$\mathbf{v} = \mathbf{P_X}\mathbf{v} = (\mathbf{P_X} - \mathbf{P_W})\mathbf{v} + \mathbf{P_W}\mathbf{v} = (\mathbf{P_X} - \mathbf{P_W})\mathbf{v} \in \mathcal{C}(\mathbf{P_X} - \mathbf{P_W}),$$

showing that the orthogonal complement of $\mathcal{C}(\mathbf{P_W})$ with respect to $\mathcal{C}(\mathbf{P_X})$ is contained in $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W})$. □

REMARK: The preceding two results are important for hypothesis testing in linear models. Consider the linear models

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and} \quad \mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. As we will learn later, the condition $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$ implies that $\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ is a **reduced model** when compared to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, sometimes called the **full model**. If $E(\boldsymbol{\epsilon}) = \mathbf{0}$, then, if the full model is correct,

$$E(\mathbf{P_X}\mathbf{Y}) = \mathbf{P_X}E(\mathbf{Y}) = \mathbf{P_X}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X}).$$

Similarly, if the reduced model is correct, $E(\mathbf{P_W}\mathbf{Y}) = \mathbf{W}\boldsymbol{\gamma} \in \mathcal{C}(\mathbf{W})$. Note that if the reduced model $\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ is correct, then the full model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is also correct since $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. Thus, if the reduced model is correct, $\mathbf{P_X}\mathbf{Y}$ and $\mathbf{P_W}\mathbf{Y}$ are attempting to estimate the same thing and their difference $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ should be small. On the other hand, if the reduced model is not correct, but the full model is, then $\mathbf{P_X}\mathbf{Y}$ and $\mathbf{P_W}\mathbf{Y}$ are estimating different things and one would expect $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ to be large. The question about whether or not to "accept" the reduced model as plausible thus hinges on deciding whether or not $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$, the (perpendicular) projection of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{P_X} - \mathbf{P_W}) = \mathcal{C}(\mathbf{P_W})^{\perp}_{\mathcal{C}(\mathbf{P_X})}$, is large or small.

2.3 Reparameterization

REMARK: For estimation in the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, we can only learn about $\boldsymbol{\beta}$ through $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$. Thus, the crucial item needed is $\mathbf{P_X}$, the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. For convenience, we call $\mathcal{C}(\mathbf{X})$ the **estimation space**. $\mathbf{P_X}$ is the perpendicular projection matrix onto the estimation space. We call $\mathcal{N}(\mathbf{X}')$ the **error space**. $\mathbf{I} - \mathbf{P_X}$ is the perpendicular projection matrix onto the error space.

IMPORTANT: Any two linear models with the same estimation space are really the same model; the models are said to be **reparameterizations** of each other. Any two such models will give the same predicted values, the same residuals, the same ANOVA table, etc. In particular, suppose that we have two linear models:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and} \quad \mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}.$$

If $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$, then $\mathbf{P_X}$ does not depend on which of $\mathbf{X}$ or $\mathbf{W}$ is used; it depends only on $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$. As we will find out, the least-squares estimate of $E(\mathbf{Y})$ is

$$\widehat{\mathbf{Y}} = \mathbf{P_X}\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{W}\widehat{\boldsymbol{\gamma}}.$$

IMPLICATION: The $\boldsymbol{\beta}$ parameters in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, are not really all that crucial. Because of this, it is standard to reparameterize linear models (i.e., change the parameters) to exploit computational advantages, as we will soon see. The essence of the model is that $E(\mathbf{Y}) \in \mathcal{C}(\mathbf{X})$. As long as we do not change $\mathcal{C}(\mathbf{X})$, the design matrix $\mathbf{X}$ and the corresponding model parameters can be altered in a manner suitable to our liking.

EXAMPLE: Recall the simple linear regression model from Chapter 1 given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$. Although not critical for this discussion, we will assume that $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are uncorrelated random variables with mean 0 and common variance $\sigma^2 > 0$. Recall that, in matrix notation,

$$\mathbf{Y}_{n\times 1}=\begin{pmatrix} Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix},\quad \mathbf{X}_{n\times 2}=\begin{pmatrix} 1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix},\quad \boldsymbol{\beta}_{2\times 1}=\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix},\quad \boldsymbol{\epsilon}_{n\times 1}=\begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \vdots\\ \epsilon_n\end{pmatrix}.$$

As long as $(x_1, x_2, ..., x_n)'$ is not a multiple of $\mathbf{1}_n$ and at least one $x_i \neq 0$, then $r(\mathbf{X}) = 2$ and $(\mathbf{X}'\mathbf{X})^{-1}$ exists. Straightforward calculations show that

$$\mathbf{X}'\mathbf{X}=\begin{pmatrix} n & \sum_i x_i\\ \sum_i x_i & \sum_i x_i^2\end{pmatrix},\quad (\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix} \frac{1}{n}+\frac{\overline{x}^2}{\sum_i(x_i-\overline{x})^2} & -\frac{\overline{x}}{\sum_i(x_i-\overline{x})^2}\\ -\frac{\overline{x}}{\sum_i(x_i-\overline{x})^2} & \frac{1}{\sum_i(x_i-\overline{x})^2}.\end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{Y}=\begin{pmatrix}\sum_i Y_i\\ \sum_i x_iY_i\end{pmatrix}.$$

Thus, the (unique) least squares estimator is given by

$$\widehat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{pmatrix}=\begin{pmatrix}\overline{Y}-\widehat{\beta}_1\overline{x}\\ \frac{\sum_i(x_i-\overline{x})(Y_i-\overline{Y})}{\sum_i(x_i-\overline{x})^2}\end{pmatrix}.$$

For the simple linear regression model, it can be shown (verify!) that the perpendicular projection matrix $\mathbf{P_X}$ is given by

$$\begin{aligned}\mathbf{P_X} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\\ &= \begin{pmatrix} \frac{1}{n}+\frac{(x_1-\overline{x})^2}{\sum_i(x_i-\overline{x})^2} & \frac{1}{n}+\frac{(x_1-\overline{x})(x_2-\overline{x})}{\sum_i(x_i-\overline{x})^2} & \cdots & \frac{1}{n}+\frac{(x_1-\overline{x})(x_n-\overline{x})}{\sum_i(x_i-\overline{x})^2}\\ \frac{1}{n}+\frac{(x_1-\overline{x})(x_2-\overline{x})}{\sum_i(x_i-\overline{x})^2} & \frac{1}{n}+\frac{(x_2-\overline{x})^2}{\sum_i(x_i-\overline{x})^2} & \cdots & \frac{1}{n}+\frac{(x_2-\overline{x})(x_n-\overline{x})}{\sum_i(x_i-\overline{x})^2}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{1}{n}+\frac{(x_1-\overline{x})(x_n-\overline{x})}{\sum_i(x_i-\overline{x})^2} & \frac{1}{n}+\frac{(x_2-\overline{x})(x_n-\overline{x})}{\sum_i(x_i-\overline{x})^2} & \cdots & \frac{1}{n}+\frac{(x_n-\overline{x})^2}{\sum_i(x_i-\overline{x})^2}\end{pmatrix}.\end{aligned}$$

A reparameterization of the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is

$$Y_i = \gamma_0 + \gamma_1(x_i - \overline{x}) + \epsilon_i$$

or $\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where

$$\mathbf{Y}_{n\times 1}=\begin{pmatrix} Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix},\quad \mathbf{W}_{n\times 2}=\begin{pmatrix} 1 & x_1-\overline{x}\\ 1 & x_2-\overline{x}\\ \vdots & \vdots\\ 1 & x_n-\overline{x}\end{pmatrix},\quad \boldsymbol{\gamma}_{2\times 1}=\begin{pmatrix}\gamma_0\\ \gamma_1\end{pmatrix},\quad \boldsymbol{\epsilon}_{n\times 1}=\begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \vdots\\ \epsilon_n\end{pmatrix}.$$

To see why this is a reparameterized model, note that if we define

$$\mathbf{U} = \begin{pmatrix} 1 & -\overline{x} \\ 0 & 1 \end{pmatrix},$$

then $\mathbf{W} = \mathbf{X}\mathbf{U}$ and $\mathbf{X} = \mathbf{W}\mathbf{U}^{-1}$ (verify!) so that $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$. Moreover, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma} = \mathbf{X}\mathbf{U}\boldsymbol{\gamma}$. Taking $\mathbf{P}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ leads to $\boldsymbol{\beta} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} = \mathbf{P}'\mathbf{X}\mathbf{U}\boldsymbol{\gamma} = \mathbf{U}\boldsymbol{\gamma}$; i.e.,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \gamma_0 - \gamma_1\overline{x} \\ \gamma_1 \end{pmatrix} = \mathbf{U}\boldsymbol{\gamma}.$$

To find the least-squares estimator for $\boldsymbol{\gamma}$ in the reparameterized model, observe that

$$\mathbf{W}'\mathbf{W} = \begin{pmatrix} n & 0 \\ 0 & \sum_i (x_i - \overline{x})^2 \end{pmatrix} \quad \text{and} \quad (\mathbf{W}'\mathbf{W})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_i (x_i-\overline{x})^2} \end{pmatrix}.$$

Note that $(\mathbf{W}'\mathbf{W})^{-1}$ is diagonal; this is one of the benefits to working with this parameterization. The least squares estimator of $\boldsymbol{\gamma}$ is given by

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} = \begin{pmatrix} \widehat{\gamma}_0 \\ \widehat{\gamma}_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \frac{\sum_i (x_i-\overline{x})(Y_i-\overline{Y})}{\sum_i (x_i-\overline{x})^2} \end{pmatrix},$$

which is different than $\widehat{\boldsymbol{\beta}}$. However, it can be shown directly (verify!) that the perpendicular projection matrix onto $\mathcal{C}(\mathbf{W})$ is

$$\begin{aligned} \mathbf{P}_{\mathbf{W}} &= \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' \\ &= \begin{pmatrix} \frac{1}{n} + \frac{(x_1-\overline{x})^2}{\sum_i (x_i-\overline{x})^2} & \frac{1}{n} + \frac{(x_1-\overline{x})(x_2-\overline{x})}{\sum_i (x_i-\overline{x})^2} & \cdots & \frac{1}{n} + \frac{(x_1-\overline{x})(x_n-\overline{x})}{\sum_i (x_i-\overline{x})^2} \\ \frac{1}{n} + \frac{(x_1-\overline{x})(x_2-\overline{x})}{\sum_i (x_i-\overline{x})^2} & \frac{1}{n} + \frac{(x_2-\overline{x})^2}{\sum_i (x_i-\overline{x})^2} & \cdots & \frac{1}{n} + \frac{(x_2-\overline{x})(x_n-\overline{x})}{\sum_i (x_i-\overline{x})^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} + \frac{(x_1-\overline{x})(x_n-\overline{x})}{\sum_i (x_i-\overline{x})^2} & \frac{1}{n} + \frac{(x_2-\overline{x})(x_n-\overline{x})}{\sum_i (x_i-\overline{x})^2} & \cdots & \frac{1}{n} + \frac{(x_n-\overline{x})^2}{\sum_i (x_i-\overline{x})^2} \end{pmatrix}. \end{aligned}$$

which is the same as $\mathbf{P}_{\mathbf{X}}$. Thus, the fitted values will be the same; i.e., $\widehat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{W}\widehat{\boldsymbol{\gamma}} = \mathbf{P}_{\mathbf{W}}\mathbf{Y}$, and the analysis will be the same under both parameterizations.

Exercise: Show that the one way fixed effects ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, and the cell means model $Y_{ij} = \mu_i + \epsilon_{ij}$ are reparameterizations of each other. Does one parameterization confer advantages over the other?

3 Estimability and Least Squares Estimators

Complementary reading from Monahan: Chapter 3 (except Section 3.9).

3.1 Introduction

REMARK: Estimability is one of the most important concepts in linear models. Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$. In our discussion that follows, the assumption $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ is not needed. Suppose that $\mathbf{X}$ is $n \times p$ with rank $r \leq p$. If $r = p$ (as in regression models), then estimability concerns vanish as $\boldsymbol{\beta}$ is estimated uniquely by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If $r < p$, (a common characteristic of ANOVA models), then $\boldsymbol{\beta}$ can not be estimated uniquely. However, even if $\boldsymbol{\beta}$ is not estimable, certain functions of $\boldsymbol{\beta}$ may be estimable.

3.2 Estimability

DEFINITIONS:

1. An estimator $t(\mathbf{Y})$ is said to be **unbiased** for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ iff $E\{t(\mathbf{Y})\} = \boldsymbol{\lambda}'\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$.

2. An estimator $t(\mathbf{Y})$ is said to be a **linear** estimator in $\mathbf{Y}$ iff $t(\mathbf{Y}) = c + \mathbf{a}'\mathbf{Y}$, for $c \in \mathcal{R}$ and $\mathbf{a} = (a_1, a_2, ..., a_n)'$, $a_i \in \mathcal{R}$.

3. A function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is said to be (linearly) **estimable** iff there exists a linear unbiased estimator for it. Otherwise, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is **nonestimable**.

Result 3.1. Under the model assumptions $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, a linear function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff there exists a vector $\mathbf{a}$ such that $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$; that is, $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$.
Proof. ($\Longleftarrow$) Suppose that there exists a vector $\mathbf{a}$ such that $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$. Then, $E(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$. Therefore, $\mathbf{a}'\mathbf{Y}$ is a linear unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ and hence

$\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable. ($\Longrightarrow$) Suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable. Then, there exists an estimator $c+\mathbf{a}'\mathbf{Y}$ that is unbiased for it; that is, $E(c+\mathbf{a}'\mathbf{Y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$. Note that $E(c+\mathbf{a}'\mathbf{Y}) = c + \mathbf{a}'\mathbf{X}\boldsymbol{\beta}$, so $\boldsymbol{\lambda}'\boldsymbol{\beta} = c + \mathbf{a}'\mathbf{X}\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$. Taking $\boldsymbol{\beta} = \mathbf{0}$ shows that $c = 0$. Successively taking $\boldsymbol{\beta}$ to be the standard unit vectors convinces us that $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$; i.e., $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$. □

Example 3.1. Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where $E(\epsilon_{ij}) = 0$. Take $a = 3$ and $n_i = 2$ so that

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

Note that $r(\mathbf{X}) = 3$, so $\mathbf{X}$ is not of full rank; i.e., $\boldsymbol{\beta}$ is not uniquely estimable. Consider the following parametric functions $\boldsymbol{\lambda}'\boldsymbol{\beta}$:

Parameter	$\boldsymbol{\lambda}'$	$\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$?	Estimable?
$\boldsymbol{\lambda}_1'\boldsymbol{\beta} = \mu$	$\boldsymbol{\lambda}_1' = (1, 0, 0, 0)$	no	no
$\boldsymbol{\lambda}_2'\boldsymbol{\beta} = \alpha_1$	$\boldsymbol{\lambda}_2' = (0, 1, 0, 0)$	no	no
$\boldsymbol{\lambda}_3'\boldsymbol{\beta} = \mu + \alpha_1$	$\boldsymbol{\lambda}_3' = (1, 1, 0, 0)$	yes	yes
$\boldsymbol{\lambda}_4'\boldsymbol{\beta} = \alpha_1 - \alpha_2$	$\boldsymbol{\lambda}_4' = (0, 1, -1, 0)$	yes	yes
$\boldsymbol{\lambda}_5'\boldsymbol{\beta} = \alpha_1 - (\alpha_2 + \alpha_3)/2$	$\boldsymbol{\lambda}_5' = (0, 1, -1/2, -1/2)$	yes	yes

Because $\boldsymbol{\lambda}_3'\boldsymbol{\beta} = \mu + \alpha_1$, $\boldsymbol{\lambda}_4'\boldsymbol{\beta} = \alpha_1 - \alpha_2$, and $\boldsymbol{\lambda}_5'\boldsymbol{\beta} = \alpha_1 - (\alpha_2 + \alpha_3)/2$ are (linearly) estimable, there must exist linear unbiased estimators for them. Note that

$$\begin{aligned} E(\overline{Y}_{1+}) &= E\left(\frac{Y_{11} + Y_{12}}{2}\right) \\ &= \frac{1}{2}(\mu + \alpha_1) + \frac{1}{2}(\mu + \alpha_1) = \mu + \alpha_1 = \boldsymbol{\lambda}_3'\boldsymbol{\beta} \end{aligned}$$

and that $\overline{Y}_{1+} = c + \mathbf{a}'\mathbf{Y}$, where $c = 0$ and $\mathbf{a}' = (1/2, 1/2, 0, 0, 0, 0)$. Also,

$$\begin{aligned} E(\overline{Y}_{1+} - \overline{Y}_{2+}) &= (\mu + \alpha_1) - (\mu + \alpha_2) \\ &= \alpha_1 - \alpha_2 = \boldsymbol{\lambda}_4'\boldsymbol{\beta} \end{aligned}$$

and that $\overline{Y}_{1+} - \overline{Y}_{2+} = c + \mathbf{a}'\mathbf{Y}$, where $c = 0$ and $\mathbf{a}' = (1/2, 1/2, -1/2, -1/2, 0, 0)$. Finally,

$$\begin{aligned} E\left\{\overline{Y}_{1+} - \left(\frac{\overline{Y}_{2+} + \overline{Y}_{3+}}{2}\right)\right\} &= (\mu + \alpha_1) - \frac{1}{2}\{(\mu + \alpha_2) + (\mu + \alpha_3)\} \\ &= \alpha_1 - \frac{1}{2}(\alpha_2 + \alpha_3) = \boldsymbol{\lambda}_5'\boldsymbol{\beta}. \end{aligned}$$

Note that

$$\overline{Y}_{1+} - \left(\frac{\overline{Y}_{2+} + \overline{Y}_{3+}}{2}\right) = c + \mathbf{a}'\mathbf{Y},$$

where $c = 0$ and $\mathbf{a}' = (1/2, 1/2, -1/4, -1/4, -1/4, -1/4)$. □

REMARKS:

1. The elements of the vector $\mathbf{X}\boldsymbol{\beta}$ are estimable.

2. If $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta}$ are estimable, then any linear combination of them; i.e., $\sum_{i=1}^{k} d_i\boldsymbol{\lambda}_i'\boldsymbol{\beta}$, where $d_i \in \mathcal{R}$, is also estimable.

3. If $\mathbf{X}$ is $n \times p$ and $r(\mathbf{X}) = p$, then $\mathcal{R}(\mathbf{X}) = \mathcal{R}^p$ and $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable for all $\boldsymbol{\lambda}$.

DEFINITION: Linear functions $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta}$ are said to be **linearly independent** if $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_k$ comprise a set of linearly independent vectors; i.e., $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1 \ \boldsymbol{\lambda}_2 \ \cdots \ \boldsymbol{\lambda}_k)$ has rank k.

Result 3.2. Under the model assumptions $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, we can always find $r = r(\mathbf{X})$ linearly independent estimable functions. Moreover, no collection of estimable functions can contain more than r linearly independent functions.

Proof. Let $\boldsymbol{\zeta}_i'$ denote the ith row of $\mathbf{X}$, for $i = 1, 2, ..., n$. Clearly, $\boldsymbol{\zeta}_1'\boldsymbol{\beta}, \boldsymbol{\zeta}_2'\boldsymbol{\beta}, ..., \boldsymbol{\zeta}_n'\boldsymbol{\beta}$ are estimable. Because $r(\mathbf{X}) = r$, we can select r linearly independent rows of $\mathbf{X}$; the corresponding r functions $\boldsymbol{\zeta}_i'\boldsymbol{\beta}$ are linearly independent. Now, let $\boldsymbol{\Lambda}'\boldsymbol{\beta} = (\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta})'$ be any collection of estimable functions. Then, $\boldsymbol{\lambda}_i' \in \mathcal{R}(\mathbf{X})$, for $i = 1, 2, ..., k$, and hence there exists a matrix $\mathbf{A}$ such that $\boldsymbol{\Lambda}' = \mathbf{A}'\mathbf{X}$. Therefore, $r(\boldsymbol{\Lambda}') = r(\mathbf{A}'\mathbf{X}) \leq r(\mathbf{X}) = r$. Hence, there can be at most r linearly independent estimable functions. □

DEFINITION: A least squares estimator of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ is any solution to the normal equations.

Result 3.3. Under the model assumptions $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'\widetilde{\boldsymbol{\beta}}$ for any two solutions $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ to the normal equations.
Proof. Suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable. Then $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$, for some $\mathbf{a}$. From Result 2.5,

$$\begin{aligned} \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \mathbf{a}'\mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} \\ \boldsymbol{\lambda}'\widetilde{\boldsymbol{\beta}} = \mathbf{a}'\mathbf{X}\widetilde{\boldsymbol{\beta}} &= \mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}. \end{aligned}$$

This proves the result. □
Alternate proof. If $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ both solve the normal equations, then $\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) = \mathbf{0}$; that is, $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}} \in \mathcal{N}(\mathbf{X}'\mathbf{X}) = \mathcal{N}(\mathbf{X})$. If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X}) \Longleftrightarrow \boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}') \Longleftrightarrow \boldsymbol{\lambda} \perp \mathcal{N}(\mathbf{X})$. Thus, $\boldsymbol{\lambda}'(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) = 0$; i.e., $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'\widetilde{\boldsymbol{\beta}}$. □

IMPLICATION: Least squares estimators of (linearly) estimable functions are **invariant** to the choice of generalized inverse used to solve the normal equations.

Example 3.2. In Example 3.1, we considered the one-way fixed effects ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, for $i = 1, 2, 3$ and $j = 1, 2$. For this model, it is easy to show that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{pmatrix}$$

and $r(\mathbf{X}'\mathbf{X}) = 3$. Here are two generalized inverses of $\mathbf{X}'\mathbf{X}$:

$$(\mathbf{X}'\mathbf{X})_1^{-} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{pmatrix} \qquad (\mathbf{X}'\mathbf{X})_2^{-} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} Y_{11}+Y_{12}+Y_{21}+Y_{12}+Y_{31}+Y_{32} \\ Y_{11}+Y_{12} \\ Y_{21}+Y_{22} \\ Y_{31}+Y_{32} \end{pmatrix}.$$

Two least squares solutions (verify!) are thus

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})_1^{-}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 0 \\ \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \overline{Y}_{3+} \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})_2^{-}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \overline{Y}_{3+} \\ \overline{Y}_{1+}-\overline{Y}_{3+} \\ \overline{Y}_{2+}-\overline{Y}_{3+} \\ 0 \end{pmatrix}.$$

Recall our estimable functions from Example 3.1:

Parameter	$\boldsymbol{\lambda}'$	$\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$?	Estimable?
$\boldsymbol{\lambda}_3'\boldsymbol{\beta} = \mu + \alpha_1$	$\boldsymbol{\lambda}_3' = (1,1,0,0)$	yes	yes
$\boldsymbol{\lambda}_4'\boldsymbol{\beta} = \alpha_1 - \alpha_2$	$\boldsymbol{\lambda}_4' = (0,1,-1,0)$	yes	yes
$\boldsymbol{\lambda}_5'\boldsymbol{\beta} = \alpha_1 - (\alpha_2+\alpha_3)/2$	$\boldsymbol{\lambda}_5' = (0,1,-1/2,-1/2)$	yes	yes

Note that

- for $\boldsymbol{\lambda}_3\boldsymbol{\beta} = \mu + \alpha_1$, the (unique) least squares estimator is

$$\boldsymbol{\lambda}_3'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}_3'\widetilde{\boldsymbol{\beta}} = \overline{Y}_{1+}.$$

- for $\boldsymbol{\lambda}_4'\boldsymbol{\beta} = \alpha_1 - \alpha_2$, the (unique) least squares estimator is

$$\boldsymbol{\lambda}_4'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}_4'\widetilde{\boldsymbol{\beta}} = \overline{Y}_{1+} - \overline{Y}_{2+}.$$

- for $\boldsymbol{\lambda}_5'\boldsymbol{\beta} = \alpha_1 - (\alpha_2+\alpha_3)/2$, the (unique) least squares estimator is

$$\boldsymbol{\lambda}_5'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}_5'\widetilde{\boldsymbol{\beta}} = \overline{Y}_{1+} - \frac{1}{2}(\overline{Y}_{2+} + \overline{Y}_{3+}).$$

Finally, note that these three estimable functions are linearly independent since

$$\boldsymbol{\Lambda} = \left(\begin{array}{ccc} \boldsymbol{\lambda}_3 & \boldsymbol{\lambda}_4 & \boldsymbol{\lambda}_5 \end{array} \right) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & -1 & -1/2 \\ 0 & 0 & -1/2 \end{pmatrix}$$

has rank $r(\boldsymbol{\Lambda}) = 3$. Of course, more estimable functions $\boldsymbol{\lambda}_i'\boldsymbol{\beta}$ can be found, but we can find no more linearly independent estimable functions because $r(\mathbf{X}) = 3$. □

Result 3.4. Under the model assumptions $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, the least squares estimator $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is a linear unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.
Proof. Suppose that $\widehat{\boldsymbol{\beta}}$ solves the normal equations. We know (by definition) that $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Note that

$$\begin{aligned} \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} &= \boldsymbol{\lambda}'\{(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} + [\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}]\mathbf{z}\} \\ &= \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} + \boldsymbol{\lambda}'[\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}]\mathbf{z}. \end{aligned}$$

Also, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable by assumption, so $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X}) \iff \boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}') \iff \boldsymbol{\lambda} \perp \mathcal{N}(\mathbf{X})$. Result MAR5.2 says that $[\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}]\mathbf{z} \in \mathcal{N}(\mathbf{X}'\mathbf{X}) = \mathcal{N}(\mathbf{X})$, so $\boldsymbol{\lambda}'[\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}]\mathbf{z} = 0$. Thus, $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$, which is a linear estimator in $\mathbf{Y}$. We now show that $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is unbiased. Because $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X}) \Longrightarrow \boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$, for some $\mathbf{a}$. Thus,

$$\begin{aligned} E(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = E\{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\} &= \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'E(\mathbf{Y}) \\ &= \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{X}\boldsymbol{\beta} = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}. \ \square \end{aligned}$$

SUMMARY: Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$. From the definition, we know that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff there exists a linear unbiased estimator for it, so if we can find a linear estimator $c+\mathbf{a}'\mathbf{Y}$ whose expectation equals $\boldsymbol{\lambda}'\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$, then $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable. From Result 3.1, we know that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$. Thus, if $\boldsymbol{\lambda}'$ can be expressed as a linear combination of the rows of $\mathbf{X}$, then $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable.

IMPORTANT: Here is a commonly-used method of finding **necessary and sufficient conditions** for estimability in linear models with $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Suppose that $\mathbf{X}$ is $n \times p$ with rank $r < p$. We know that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$.

- Typically, when we find the rank of $\mathbf{X}$, we find r linearly independent columns of $\mathbf{X}$ and express the remaining $s = p - r$ columns as linear combinations of the r linearly independent columns of $\mathbf{X}$. Suppose that $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s$ satisfy $\mathbf{X}\mathbf{c}_i = \mathbf{0}$, for $i = 1, 2, ..., s$, that is, $\mathbf{c}_i \in \mathcal{N}(\mathbf{X})$, for $i = 1, 2, ..., s$. If $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s\}$ forms a basis for $\mathcal{N}(\mathbf{X})$; i.e., $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s$ are linearly independent, then

$$\begin{aligned}
\boldsymbol{\lambda}'\mathbf{c}_1 &= 0 \\
\boldsymbol{\lambda}'\mathbf{c}_2 &= 0 \\
&\vdots \\
\boldsymbol{\lambda}'\mathbf{c}_s &= 0
\end{aligned}$$

are necessary and sufficient conditions for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable.

REMARK: There are two spaces of interest: $\mathcal{C}(\mathbf{X}') = \mathcal{R}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X})$. If $\mathbf{X}$ is $n \times p$ with rank $r < p$, then $\dim\{\mathcal{C}(\mathbf{X}')\} = r$ and $\dim\{\mathcal{N}(\mathbf{X})\} = s = p - r$. Therefore, if $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s$ are linearly independent, then $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s\}$ must be a basis for $\mathcal{N}(\mathbf{X})$. But,

$$\begin{aligned}
\boldsymbol{\lambda}'\boldsymbol{\beta} \text{ estimable} \Longleftrightarrow \boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X}) &\Longleftrightarrow \boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}') \\
&\Longleftrightarrow \boldsymbol{\lambda} \text{ is orthogonal to every vector in } \mathcal{N}(\mathbf{X}) \\
&\Longleftrightarrow \boldsymbol{\lambda} \text{ is orthogonal to } \mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s \\
&\Longleftrightarrow \boldsymbol{\lambda}'\mathbf{c}_i = 0,\ i = 1, 2, ..., s.
\end{aligned}$$

Therefore, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda}'\mathbf{c}_i = 0$, for $i = 1, 2, ..., s$, where $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s$ are s linearly independent vectors satisfying $\mathbf{X}\mathbf{c}_i = \mathbf{0}$.

TERMINOLOGY: A set of linear functions $\{\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta}\}$ is said to be **jointly nonestimable** if the only linear combination of $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta}$ that is estimable is the trivial one; i.e., $\equiv 0$. These types of functions are useful in non-full-rank linear models and are associated with side conditions.

3.2.1 One-way ANOVA

GENERAL CASE: Consider the one-way fixed effects ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where $E(\epsilon_{ij}) = 0$. In matrix form, $\mathbf{X}$ and $\boldsymbol{\beta}$ are

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix},$$

where $p = a+1$ and $n = \sum_i n_i$. Note that the last a columns of $\mathbf{X}$ are linearly independent and the first column is the sum of the last a columns. Hence, $r(\mathbf{X}) = r = a$ and $s = p - r = 1$. With $\mathbf{c}_1 = (1, -\mathbf{1}_a')'$, note that $\mathbf{X}\mathbf{c}_1 = \mathbf{0}$ so $\{\mathbf{c}_1\}$ forms a basis for $\mathcal{N}(\mathbf{X})$. Thus, the necessary and sufficient condition for $\boldsymbol{\lambda}'\boldsymbol{\beta} = \lambda_0\mu + \sum_{i=1}^a \lambda_i\alpha_i$ to be estimable is

$$\boldsymbol{\lambda}'\mathbf{c}_1 = 0 \Longrightarrow \lambda_0 = \sum_{i=1}^a \lambda_i.$$

Here are some examples of **estimable** functions:

1. $\mu + \alpha_i$
2. $\alpha_i - \alpha_k$
3. any **contrast** in the α's; i.e., $\sum_{i=1}^a \lambda_i\alpha_i$, where $\sum_{i=1}^a \lambda_i = 0$.

Here are some examples of **nonestimable** functions:

1. μ
2. α_i
3. $\sum_{i=1}^a n_i\alpha_i$.

There is only $s = 1$ jointly nonestimable function. Later we will learn that jointly nonestimable functions can be used to "force" particular solutions to the normal equations.

The following are examples of sets of linearly independent estimable functions (verify!):

1. $\{\mu+\alpha_1, \mu+\alpha_2, ..., \mu+\alpha_a\}$
2. $\{\mu+\alpha_1, \alpha_1-\alpha_2, ..., \alpha_1-\alpha_a\}$.

LEAST SQUARES ESTIMATES: We now wish to calculate the least squares estimates of estimable functions. Note that $\mathbf{X}'\mathbf{X}$ and one generalized inverse of $\mathbf{X}'\mathbf{X}$ is given by

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n_1 & n_2 & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & 0 & \cdots & n_a \end{pmatrix} \quad \text{and} \quad (\mathbf{X}'\mathbf{X})^{-} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1/n_1 & 0 & \cdots & 0 \\ 0 & 0 & 1/n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/n_a \end{pmatrix}$$

For this generalized inverse, the least squares estimate is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1/n_1 & 0 & \cdots & 0 \\ 0 & 0 & 1/n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/n_a \end{pmatrix} \begin{pmatrix} \sum_i \sum_j Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \vdots \\ \sum_j Y_{aj} \end{pmatrix} = \begin{pmatrix} 0 \\ \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \vdots \\ \overline{Y}_{a+} \end{pmatrix}.$$

REMARK: We know that this solution is not unique; had we used a different generalized inverse above, we would have gotten a different least squares estimate of $\boldsymbol{\beta}$. However, least squares estimates of estimable functions $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are invariant to the choice of generalized inverse, so our choice of $(\mathbf{X}'\mathbf{X})^{-}$ above is as good as any other. From this solution, we have the unique least squares estimates:

Estimable function, $\boldsymbol{\lambda}'\boldsymbol{\beta}$	Least squares estimate, $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$
$\mu+\alpha_i$	$\overline{Y}_{i+}$
$\alpha_i-\alpha_k$	$\overline{Y}_{i+}-\overline{Y}_{k+}$
$\sum_{i=1}^a \lambda_i\alpha_i$, where $\sum_{i=1}^a \lambda_i = 0$	$\sum_{i=1}^a \lambda_i\overline{Y}_{i+}$

3.2.2 Two-way crossed ANOVA with no interaction

GENERAL CASE: Consider the two-way fixed effects (crossed) ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$, where $E(\epsilon_{ij}) = 0$. For ease of presentation, we take $n_{ij} = 1$ so there is no need for a k subscript; that is, we can rewrite the model as $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$. In matrix form, $\mathbf{X}$ and $\boldsymbol{\beta}$ are

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & \mathbf{0}_b & \cdots & \mathbf{0}_b & \mathbf{I}_b \\ \mathbf{1}_b & \mathbf{0}_b & \mathbf{1}_b & \cdots & \mathbf{0}_b & \mathbf{I}_b \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_b & \mathbf{0}_b & \mathbf{0}_b & \cdots & \mathbf{1}_b & \mathbf{I}_b \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_b \end{pmatrix},$$

where $p = a + b + 1$ and $n = ab$. Note that the first column is the sum of the last b columns. The 2nd column is the sum of the last b columns minus the sum of columns 3 through $a + 1$. The remaining columns are linearly independent. Thus, we have $s = 2$ linear dependencies so that $r(\mathbf{X}) = a + b - 1$. The dimension of $\mathcal{N}(\mathbf{X})$ is $s = 2$. Taking

$$\mathbf{c}_1 = \begin{pmatrix} 1 \\ -\mathbf{1}_a \\ \mathbf{0}_b \end{pmatrix} \quad \text{and} \quad \mathbf{c}_2 = \begin{pmatrix} 1 \\ \mathbf{0}_a \\ -\mathbf{1}_b \end{pmatrix}$$

produces $\mathbf{X}\mathbf{c}_1 = \mathbf{X}\mathbf{c}_2 = \mathbf{0}$. Since $\mathbf{c}_1$ and $\mathbf{c}_2$ are linearly independent; i.e., neither is a multiple of the other, $\{\mathbf{c}_1, \mathbf{c}_2\}$ is a basis for $\mathcal{N}(\mathbf{X})$. Thus, necessary and sufficient conditions for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable are

$$\boldsymbol{\lambda}'\mathbf{c}_1 = 0 \implies \lambda_0 = \sum_{i=1}^{a} \lambda_i$$

$$\boldsymbol{\lambda}'\mathbf{c}_2 = 0 \implies \lambda_0 = \sum_{j=1}^{b} \lambda_{a+j}.$$

Here are some examples of **estimable** functions:

1. $\mu + \alpha_i + \beta_j$
2. $\alpha_i - \alpha_k$
3. $\beta_j - \beta_k$
4. any contrast in the α's; i.e., $\sum_{i=1}^{a} \lambda_i \alpha_i$, where $\sum_{i=1}^{a} \lambda_i = 0$
5. any contrast in the β's; i.e., $\sum_{j=1}^{b} \lambda_{a+j} \beta_j$, where $\sum_{j=1}^{b} \lambda_{a+j} = 0$.

Here are some examples of **nonestimable** functions:

1. μ
2. α_i
3. β_j
4. $\sum_{i=1}^{a} \alpha_i$
5. $\sum_{j=1}^{b} \beta_j$.

We can find $s = 2$ jointly nonestimable functions. Examples of sets of jointly nonestimable functions are

1. $\{\alpha_a, \beta_b\}$
2. $\{\sum_i \alpha_i, \sum_j \beta_j\}$.

A set of linearly independent estimable functions (verify!) is

1. $\{\mu + \alpha_1 + \beta_1, \alpha_1 - \alpha_2, ..., \alpha_1 - \alpha_a, \beta_1 - \beta_2, ..., \beta_1 - \beta_b\}$.

NOTE: When **replication** occurs; i.e., when $n_{ij} > 1$, for all i and j, our estimability findings are unchanged. Replication does not change $\mathcal{R}(\mathbf{X})$. We obtain the following least squares estimates:

Estimable function, $\boldsymbol{\lambda}'\boldsymbol{\beta}$	Least squares estimate, $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$
$\mu + \alpha_i + \beta_j$	$\overline{Y}_{ij+}$
$\alpha_i - \alpha_l$	$\overline{Y}_{i++} - \overline{Y}_{l++}$
$\beta_j - \beta_l$	$\overline{Y}_{+j+} - \overline{Y}_{+l+}$
$\sum_{i=1}^{a} c_i\alpha_i$, with $\sum_{i=1}^{a} c_i = 0$	$\sum_{i=1}^{a} c_i\overline{Y}_{i++}$
$\sum_{j=1}^{b} d_i\beta_j$, with $\sum_{j=1}^{b} d_i = 0$	$\sum_{j=1}^{b} d_i\overline{Y}_{+j+}$

These formulae are still technically correct when $n_{ij} = 1$. When some $n_{ij} = 0$, i.e., there are missing cells, estimability may be affected; see Monahan, pp 46-48.

3.2.3 Two-way crossed ANOVA with interaction

GENERAL CASE: Consider the two-way fixed effects (crossed) ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$, where $E(\epsilon_{ij}) = 0$.

SPECIAL CASE: With $a = 3$, $b = 2$, and $n_{ij} = 2$, $\mathbf{X}$ and $\boldsymbol{\beta}$ are

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{31} \\ \gamma_{32} \end{pmatrix}.$$

There are $p = 12$ parameters. The last six columns of $\mathbf{X}$ are linearly independent, and the other columns can be written as linear combinations of the last six columns, so $r(\mathbf{X}) = 6$ and $s = p - r = 6$. To determine which functions $\boldsymbol{\lambda}'\boldsymbol{\beta}$ are estimable, we need to find a basis for $\mathcal{N}(\mathbf{X})$. One basis $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_6\}$ is

$$\left\{ \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

Functions $\boldsymbol{\lambda}'\boldsymbol{\beta}$ must satisfy $\boldsymbol{\lambda}'\mathbf{c}_i = 0$, for each $i = 1, 2, ..., 6$, to be estimable. It should be obvious that neither the main effect terms nor the interaction terms; i.e, α_i, β_j, γ_{ij}, are estimable on their own. The six $\alpha_i + \beta_j + \gamma_{ij}$ "cell means" terms are, but these are not that interesting. No longer are contrasts in the α's or β's estimable. Indeed, interaction makes the analysis more difficult.

3.3 Reparameterization

SETTING: Consider the general linear model

$$\text{Model GL: } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } E(\boldsymbol{\epsilon}) = \mathbf{0}.$$

Assume that $\mathbf{X}$ is $n \times p$ with rank $r \leq p$. Suppose that $\mathbf{W}$ is an $n \times t$ matrix such that $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$. Then, we know that there exist matrices $\mathbf{T}_{p\times t}$ and $\mathbf{S}_{p\times t}$ such that

$\mathbf{W} = \mathbf{X}\mathbf{T}$ and $\mathbf{X} = \mathbf{W}\mathbf{S}'$. Note that $\mathbf{X}\boldsymbol{\beta} = \mathbf{W}\mathbf{S}'\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \mathbf{S}'\boldsymbol{\beta}$. The model

$$\text{Model GL-R: } \mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \text{where } E(\boldsymbol{\epsilon}) = \mathbf{0},$$

is called a **reparameterization** of Model GL.

REMARK: Since $\mathbf{X}\boldsymbol{\beta} = \mathbf{W}\mathbf{S}'\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma} = \mathbf{X}\mathbf{T}\boldsymbol{\gamma}$, we might suspect that the estimation of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ under Model GL should be essentially the same as the estimation of $\boldsymbol{\lambda}'\mathbf{T}\boldsymbol{\gamma}$ under Model GL-R (and that estimation of an estimable function $\mathbf{q}'\boldsymbol{\gamma}$ under Model GL-R should be essentially the same as estimation of $\mathbf{q}'\mathbf{S}'\boldsymbol{\beta}$ under Model GL). The upshot of the following results is that, in determining a least squares estimate of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$, we can work with either Model GL or Model GL-R. The actual nature of these conjectured relationships is now made precise.

Result 3.5. Consider Models GL and GL-R with $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$.

1. $\mathbf{P}_{\mathbf{W}} = \mathbf{P}_{\mathbf{X}}$.

2. If $\widehat{\boldsymbol{\gamma}}$ is any solution to the normal equations $\mathbf{W}'\mathbf{W}\boldsymbol{\gamma} = \mathbf{W}'\mathbf{Y}$ associated with Model GL-R, then $\widehat{\boldsymbol{\beta}} = \mathbf{T}\widehat{\boldsymbol{\gamma}}$ is a solution to the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ associated with Model GL.

3. If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable under Model GL and if $\widehat{\boldsymbol{\gamma}}$ is any solution to the normal equations $\mathbf{W}'\mathbf{W}\boldsymbol{\gamma} = \mathbf{W}'\mathbf{Y}$ associated with Model GL-R, then $\boldsymbol{\lambda}'\mathbf{T}\widehat{\boldsymbol{\gamma}}$ is the least squares estimate of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

4. If $\mathbf{q}'\boldsymbol{\gamma}$ is estimable under Model GL-R; i.e., if $\mathbf{q}' \in \mathcal{R}(\mathbf{W})$, then $\mathbf{q}'\mathbf{S}'\boldsymbol{\beta}$ is estimable under Model GL and its least squares estimate is given by $\mathbf{q}'\widehat{\boldsymbol{\gamma}}$, where $\widehat{\boldsymbol{\gamma}}$ is any solution to the normal equations $\mathbf{W}'\mathbf{W}\boldsymbol{\gamma} = \mathbf{W}'\mathbf{Y}$.

Proof.

1. $\mathbf{P}_{\mathbf{W}} = \mathbf{P}_{\mathbf{X}}$ since perpendicular projection matrices are unique.

2. Note that

$$\mathbf{X}'\mathbf{X}\mathbf{T}\widehat{\boldsymbol{\gamma}} = \mathbf{X}'\mathbf{W}\widehat{\boldsymbol{\gamma}} = \mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{Y} = \mathbf{X}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{X}'\mathbf{Y}.$$

Hence, $\mathbf{T}\widehat{\boldsymbol{\gamma}}$ is a solution to the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$.

3. This follows from (2), since the least squares estimate is invariant to the choice of the solution to the normal equations.
4. If $\mathbf{q}' \in \mathcal{R}(\mathbf{W})$, then $\mathbf{q}' = \mathbf{a}'\mathbf{W}$, for some $\mathbf{a}$. Then, $\mathbf{q}'\mathbf{S}' = \mathbf{a}'\mathbf{W}\mathbf{S}' = \mathbf{a}'\mathbf{X} \in \mathcal{R}(\mathbf{X})$, so that $\mathbf{q}'\mathbf{S}'\boldsymbol{\beta}$ is estimable under Model GL. From (3), we know the least squares estimate of $\mathbf{q}'\mathbf{S}'\boldsymbol{\beta}$ is $\mathbf{q}'\mathbf{S}'\mathbf{T}\widehat{\boldsymbol{\gamma}}$. But,

$$\mathbf{q}'\mathbf{S}'\mathbf{T}\widehat{\boldsymbol{\gamma}} = \mathbf{a}'\mathbf{W}\mathbf{S}'\mathbf{T}\widehat{\boldsymbol{\gamma}} = \mathbf{a}'\mathbf{X}\mathbf{T}\widehat{\boldsymbol{\gamma}} = \mathbf{a}'\mathbf{W}\widehat{\boldsymbol{\gamma}} = \mathbf{q}'\widehat{\boldsymbol{\gamma}}. \ \square$$

WARNING: The converse to (4) is not true; i.e., $\mathbf{q}'\mathbf{S}'\boldsymbol{\beta}$ being estimable under Model GL doesn't necessarily imply that $\mathbf{q}'\boldsymbol{\gamma}$ is estimable under Model GL-R. See Monahan, pp 52.

TERMINOLOGY: Because $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$ and $r(\mathbf{X}) = r$, $\mathbf{W}_{n\times t}$ must have at least r columns. If $\mathbf{W}$ has exactly r columns; i.e., if $t = r$, then the reparameterization of Model GL is called a **full rank reparameterization**. If, in addition, $\mathbf{W}'\mathbf{W}$ is diagonal, the reparameterization of Model GL is called an **orthogonal reparameterization**; see, e.g., the centered linear regression model in Section 2 (notes).

NOTE: A full rank reparameterization always exists; just delete the columns of $\mathbf{X}$ that are linearly dependent on the others. In a full rank reparameterization, $(\mathbf{W}'\mathbf{W})^{-1}$ exists, so the normal equations $\mathbf{W}'\mathbf{W}\boldsymbol{\gamma} = \mathbf{W}'\mathbf{Y}$ have a unique solution; i.e., $\widehat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}$.

DISCUSSION: There are two (opposing) points of view concerning the utility of full rank reparameterizations.

- Some argue that, since making inferences about $\mathbf{q}'\boldsymbol{\gamma}$ under the full rank reparameterized model (Model GL-R) is equivalent to making inferences about $\mathbf{q}'\mathbf{S}'\boldsymbol{\beta}$ in the possibly-less-than-full rank original model (Model GL), the inclusion of the possibility that the design matrix has less than full column rank causes a needless complication in linear model theory.

- The opposing argument is that, since computations required to deal with the reparameterized model are essentially the same as those required to handle the original model, we might as well allow for less-than-full rank models in the first place.

- I tend to favor the latter point of view; to me, there is no reason not to include less-than-full rank models as long as you know what you can and can not estimate.

Example 3.3. Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where $E(\epsilon_{ij}) = 0$. In matrix form, $\mathbf{X}$ and $\boldsymbol{\beta}$ are

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix},$$

where $p = a + 1$ and $n = \sum_i n_i$. This is not a full rank model since the first column is the sum of the last a columns; i.e., $r(\mathbf{X}) = a$.

Reparameterization 1: Deleting the first column of $\mathbf{X}$, we have

$$\mathbf{W}_{n\times t} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{1}_{n_a} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma}_{t\times 1} = \begin{pmatrix} \mu + \alpha_1 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_a \end{pmatrix} \equiv \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_a \end{pmatrix},$$

where $t = a$ and $\mu_i = E(Y_{ij}) = \mu + \alpha_i$. This is called the **cell-means model** and is written $Y_{ij} = \mu_i + \epsilon_{ij}$. This is a full rank reparameterization with $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$. The least squares estimate of $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} = \begin{pmatrix} \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \vdots \\ \overline{Y}_{a+} \end{pmatrix}.$$

Exercise: What are the matrices $\mathbf{T}$ and $\mathbf{S}$ associated with this reparameterization?

Reparameterization 2: Deleting the last column of $\mathbf{X}$, we have

$$\mathbf{W}_{n\times t} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \mathbf{0}_{n_{a-1}} & \mathbf{0}_{n_{a-1}} & \cdots & \mathbf{1}_{n_{a-1}} \\ \mathbf{1}_{n_a} & \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{0}_{n_a} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma}_{t\times 1} = \begin{pmatrix} \mu + \alpha_a \\ \alpha_1 - \alpha_a \\ \alpha_2 - \alpha_a \\ \vdots \\ \alpha_{a-1} - \alpha_a \end{pmatrix},$$

where $t = a$. This is called the **cell-reference model** (what SAS uses by default). This is a full rank reparameterization with $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$. The least squares estimate of $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} = \begin{pmatrix} \overline{Y}_{a+} \\ \overline{Y}_{1+} - \overline{Y}_{a+} \\ \overline{Y}_{2+} - \overline{Y}_{a+} \\ \vdots \\ \overline{Y}_{(a-1)+} - \overline{Y}_{a+} \end{pmatrix}.$$

Reparameterization 3: Another reparameterization of the effects model uses

$$\mathbf{W}_{n\times t} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \mathbf{0}_{n_{a-1}} & \mathbf{0}_{n_{a-1}} & \cdots & \mathbf{1}_{n_{a-1}} \\ \mathbf{1}_{n_a} & -\mathbf{1}_{n_a} & -\mathbf{1}_{n_a} & \cdots & -\mathbf{1}_{n_a} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma}_{t\times 1} = \begin{pmatrix} \mu + \overline{\alpha} \\ \alpha_1 - \overline{\alpha} \\ \alpha_2 - \overline{\alpha} \\ \vdots \\ \alpha_{a-1} - \overline{\alpha} \end{pmatrix},$$

where $t = a$ and $\overline{\alpha} = a^{-1}\sum_i \alpha_i$. This is called the **deviations from the mean model**. This is a full rank reparameterization with $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$.

Example 3.4. *Two part multiple linear regression model.* Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Suppose that $\mathbf{X}$ is full rank. Write $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ so that the model can be written as

$$\text{Model GL:}\ \ \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

Now, set $\mathbf{W}_1 = \mathbf{X}_1$ and $\mathbf{W}_2 = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$, where $\mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X}_1)$. A reparameterized version of Model GL is

$$\text{Model GL-R:}\ \ \mathbf{Y} = \mathbf{W}_1\boldsymbol{\gamma}_1 + \mathbf{W}_2\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathbf{W} = (\mathbf{W}_1\ \mathbf{W}_2)$ and

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

With this reparameterization, note that

$$\mathbf{W}'\mathbf{W} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2 \end{pmatrix} \text{ and } \mathbf{W}'\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} \end{pmatrix}$$

so that

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} \\ \{\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2\}^{-1}\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} \end{pmatrix} \equiv \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_1 \\ \widehat{\boldsymbol{\gamma}}_2 \end{pmatrix}.$$

In this reparameterization, $\mathbf{W}_2$ can be thought of as the "residual" from regressing (each column of) $\mathbf{X}_2$ on $\mathbf{X}_1$. A further calculation shows that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_1 - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 \\ \widehat{\boldsymbol{\gamma}}_2 \end{pmatrix},$$

where note that $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is the estimate obtained from "regressing" $\mathbf{X}_2$ on $\mathbf{X}_1$. Furthermore, the estimate $\widehat{\boldsymbol{\gamma}}_2$ can be thought of as the estimate obtained from regressing $\mathbf{Y}$ on $\mathbf{W}_2 = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$.

APPLICATION: Consider the two part full-rank regression model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Suppose that $\mathbf{X}_2 = \mathbf{x}_2$ is $n \times 1$ and that $\boldsymbol{\beta}_2 = \beta_2$ is a scalar. Consider two different models:

$$\begin{aligned} \text{Reduced model:} &\qquad \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \\ \text{Full model:} &\qquad \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\epsilon}. \end{aligned}$$

We use the term "reduced model" since $\mathcal{C}(\mathbf{X}_1) \subset \mathcal{C}(\mathbf{X}_1, \mathbf{x}_2)$. Consider the full model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\epsilon}$ and premultiply by $\mathbf{I} - \mathbf{P}_{\mathbf{X}_1}$ to obtain

$$\begin{aligned} (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} &= (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_1\boldsymbol{\beta}_1 + b_2(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2 + (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\boldsymbol{\epsilon} \\ &= b_2(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2 + \boldsymbol{\epsilon}^*, \end{aligned}$$

where $\boldsymbol{\epsilon}^* = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\boldsymbol{\epsilon}$. Now, note that

$$(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} = \mathbf{Y} - \mathbf{P}_{\mathbf{X}_1}\mathbf{Y} \equiv \widehat{\mathbf{e}}_{\mathbf{Y}|\mathbf{X}_1},$$

say, are the residuals from regressing $\mathbf{Y}$ on $\mathbf{X}_1$. Similarly, $(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2 \equiv \widehat{\mathbf{e}}_{\mathbf{x}_2|\mathbf{X}_1}$ are the residuals from regressing $\mathbf{x}_2$ on $\mathbf{X}_1$. We thus have the following induced linear model

$$\widehat{\mathbf{e}}_{\mathbf{Y}|\mathbf{X}_1} = b_2\widehat{\mathbf{e}}_{\mathbf{x}_2|\mathbf{X}_1} + \boldsymbol{\epsilon}^*,$$

where $E(\boldsymbol{\epsilon}^*) = \mathbf{0}$. The plot of $\widehat{\mathbf{e}}_{\mathbf{Y}|\mathbf{X}_1}$ versus $\widehat{\mathbf{e}}_{\mathbf{x}_2|\mathbf{X}_1}$ is called an **added-variable plot** (or partial regression) plot. It displays the relationship between $\mathbf{Y}$ and $\mathbf{x}_2$, after adjusting for the effects of $\mathbf{X}_1$ being in the model.

- If a linear trend exists in this plot, this suggests that $\mathbf{x}_2$ enters into the (full) model linearly. This plot can also be useful for detecting outliers and high leverage points.

- On the down side, added-variable plots only look at one predictor at a time so one can not assess multicolinearity; that is, if the predictor $\mathbf{x}_2$ is "close" to $\mathcal{C}(\mathbf{X}_1)$, this may not be detected in the plot.

- The slope of the least squares regression line for the added variable plot is

$$\begin{aligned}\widehat{\beta}_2 &= [\{(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2\}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2]^{-1}\{(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2\}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} \\ &= \{\mathbf{x}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_2\}^{-1}\mathbf{x}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}.\end{aligned}$$

  This is equal to the least squares estimate of β_2 in the full model.

3.4 Forcing least squares solutions using linear constraints

REVIEW: Consider our general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\mathbf{X}$ is an $n \times p$ matrix with rank r. The normal equations are $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$.

- If $r = p$, then a unique least squares solution exists; i.e., $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

- If $r < p$, then a least squares solution is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$. This solution is not unique; its value depends on which generalized inverse $(\mathbf{X}'\mathbf{X})^{-}$ is used.

Example 3.5. Consider the one-way fixed effects ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where $E(\epsilon_{ij}) = 0$. The normal equations are

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} n & n_1 & n_2 & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & 0 & \cdots & n_a \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} = \begin{pmatrix} \sum_i \sum_j Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \vdots \\ \sum_j Y_{aj} \end{pmatrix} = \mathbf{X}'\mathbf{Y},$$

or, written another way,

$$\begin{aligned} n\mu + \sum_{i=1}^{a} n_i\alpha_i &= Y_{++} \\ n_i\mu \quad + \quad n_i\alpha_i &= Y_{i+}, \quad i = 1, 2, ..., a, \end{aligned}$$

where $Y_{i+} = \sum_j Y_{ij}$, for $i = 1, 2, ..., a$, and $Y_{++} = \sum_i \sum_j Y_{ij}$. This set of equations has no unique solution. However, from our discussion on generalized inverses (and consideration of this model), we know that

- if we set $\mu = 0$, then we get the solution $\widehat{\mu} = 0$ and $\widehat{\alpha}_i = \overline{Y}_{i+}$, for $i = 1, 2, ..., a$.
- if we set $\sum_{i=1}^{a} n_i\alpha_i = 0$, then we get the solution $\widehat{\mu} = \overline{Y}_{++}$ and $\widehat{\alpha}_i = \overline{Y}_{i+} - \overline{Y}_{++}$, for $i = 1, 2, ..., a$.
- if we set another nonestimable function equal to 0, we'll get a different solution to the normal equations.

REMARK: Equations like $\mu = 0$ and $\sum_{i=1}^{a} n_i\alpha_i = 0$ are used to "force" a particular solution to the normal equations and are called **side conditions**. Different side conditions produce different least squares solutions. We know that in the one-way ANOVA model, the parameters μ and α_i, for $i = 1, 2, ..., a$, are not estimable (individually). Imposing side conditions does not change this. My feeling is that when we attach side conditions to force a unique solution, we are doing nothing more than solving a mathematical problem that isn't relevant. After all, estimable functions $\boldsymbol{\lambda}'\boldsymbol{\beta}$ have least squares estimates that do not depend on which side condition was used, and these are the only functions we should ever be concerned with.

REMARK: We have seen similar results for the two-way crossed ANOVA model. In general, what and how many conditions should we use to "force" a particular solution to the normal equations? Mathematically, we are interested in imposing additional linear restrictions of the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ where the matrix $\mathbf{C}$ does not depend on $\mathbf{Y}$.

TERMINOLOGY: We say that the system of equations $\mathbf{Ax} = \mathbf{q}$ is **compatible** if $\mathbf{c}'\mathbf{A} = \mathbf{0} \Longrightarrow \mathbf{c}'\mathbf{q} = 0$; i.e., $\mathbf{c} \in \mathcal{N}(\mathbf{A}') \Longrightarrow \mathbf{c}'\mathbf{q} = 0$.

Result 3.6. The system $\mathbf{Ax} = \mathbf{q}$ is consistent if and only if it is compatible.
Proof. If $\mathbf{Ax} = \mathbf{q}$ is consistent, then $\mathbf{Ax}^* = \mathbf{q}$, for some $\mathbf{x}^*$. Hence, for any $\mathbf{c}$ such that $\mathbf{c}'\mathbf{A} = \mathbf{0}$, we have $\mathbf{c}'\mathbf{q} = \mathbf{c}'\mathbf{Ax}^* = 0$, so $\mathbf{Ax} = \mathbf{q}$ is compatible. If $\mathbf{Ax} = \mathbf{q}$ is compatible, then for any $\mathbf{c} \in \mathcal{N}(\mathbf{A}') = \mathcal{C}(\mathbf{I} - \mathbf{P}_\mathbf{A})$, we have $0 = \mathbf{c}'\mathbf{q} = \mathbf{q}'\mathbf{c} = \mathbf{q}'(\mathbf{I} - \mathbf{P}_\mathbf{A})\mathbf{z}$, for all $\mathbf{z}$. Successively taking $\mathbf{z}$ to be the standard unit vectors, we have $\mathbf{q}'(\mathbf{I} - \mathbf{P}_\mathbf{A}) = \mathbf{0} \Longrightarrow (\mathbf{I} - \mathbf{P}_\mathbf{A})\mathbf{q} = \mathbf{0} \Longrightarrow \mathbf{q} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{q} \Longrightarrow \mathbf{q} = \mathbf{Ax}^*$, where $\mathbf{x}^* = (\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{q}$. Thus, $\mathbf{Ax} = \mathbf{q}$ is consistent. □

AUGMENTED NORMAL EQUATIONS: We now consider adjoining the set of equations $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ to the normal equations; that is, we consider the new set of equations

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix}.$$

These are called the **augmented normal equations**. When we add the constraint $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, we want these equations to be consistent for all $\mathbf{Y}$. We now would like to find a sufficient condition for consistency. Suppose that $\mathbf{w} \in \mathcal{R}(\mathbf{X}'\mathbf{X}) \cap \mathcal{R}(\mathbf{C})$. Note that

$$\begin{aligned} \mathbf{w} \in \mathcal{R}(\mathbf{X}'\mathbf{X}) \quad &\Longrightarrow \quad \mathbf{w} = \mathbf{X}'\mathbf{X}\mathbf{v}_1, \text{ for some } \mathbf{v}_1 \\ \mathbf{w} \in \mathcal{R}(\mathbf{C}) \quad &\Longrightarrow \quad \mathbf{w} = -\mathbf{C}'\mathbf{v}_2, \text{ for some } \mathbf{v}_2. \end{aligned}$$

Thus, $\mathbf{0} = \mathbf{w} - \mathbf{w} = \mathbf{X}'\mathbf{X}\mathbf{v}_1 + \mathbf{C}'\mathbf{v}_2$

$$\begin{aligned} &\Longrightarrow \quad \mathbf{0} = \mathbf{v}_1'\mathbf{X}'\mathbf{X} + \mathbf{v}_2'\mathbf{C} \\ &\Longrightarrow \quad \mathbf{0} = (\mathbf{v}_1' \ \ \mathbf{v}_2') \begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} = \mathbf{v}' \begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix}, \end{aligned}$$

where $\mathbf{v}' = (\mathbf{v}_1' \; \mathbf{v}_2')$. We want $\mathbf{C}$ chosen so that

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix}$$

is consistent, or equivalently from Result 3.6, is compatible. Compatibility occurs when

$$\mathbf{v}' \begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} = \mathbf{0} \Longrightarrow \mathbf{v}' \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix} = 0.$$

Thus, we need $\mathbf{v}_1'\mathbf{X}'\mathbf{Y} = 0$, for all $\mathbf{Y}$. Successively taking $\mathbf{Y}$ to be standard unit vectors, for $i = 1, 2, ..., n$, convinces us that $\mathbf{v}_1'\mathbf{X}' = \mathbf{0} \Longleftrightarrow \mathbf{X}\mathbf{v}_1 = \mathbf{0} \Longleftrightarrow \mathbf{X}'\mathbf{X}\mathbf{v}_1 = \mathbf{0} \Longrightarrow \mathbf{w} = \mathbf{0}$. Thus, the augmented normal equations are consistent when $\mathcal{R}(\mathbf{X}'\mathbf{X}) \cap \mathcal{R}(\mathbf{C}) = \{\mathbf{0}\}$. Since $\mathcal{R}(\mathbf{X}'\mathbf{X}) = \mathcal{R}(\mathbf{X})$, a sufficient condition for consistency is $\mathcal{R}(\mathbf{X}) \cap \mathcal{R}(\mathbf{C}) = \{\mathbf{0}\}$. Now, consider the parametric function $\boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\beta}$, for some $\boldsymbol{\lambda}$. We know that $\boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{\lambda}'\mathbf{C} \in \mathcal{R}(\mathbf{X})$. However, clearly $\boldsymbol{\lambda}'\mathbf{C} \in \mathcal{R}(\mathbf{C})$. Thus, $\boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\beta} = 0$. In other words, writing

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_s' \end{pmatrix},$$

the set of functions $\{\mathbf{c}_1'\boldsymbol{\beta}, \mathbf{c}_2'\boldsymbol{\beta}, ..., \mathbf{c}_s'\boldsymbol{\beta}\}$ is jointly nonestimable. Therefore, we can set a collection of jointly nonestimable functions equal to zero and augment the normal equations so that they remain consistent. We get a unique solution if

$$r \begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} = p.$$

Because $\mathcal{R}(\mathbf{X}'\mathbf{X}) \cap \mathcal{R}(\mathbf{C}) = \{\mathbf{0}\}$,

$$p = r \begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} = r(\mathbf{X}'\mathbf{X}) + r(\mathbf{C}) = r + r(\mathbf{C}),$$

showing that we need $r(\mathbf{C}) = s = p - r$.

SUMMARY: To augment the normal equations, we can find a set of s jointly nonestimable functions $\{\mathbf{c}_1'\boldsymbol{\beta}, \mathbf{c}_2'\boldsymbol{\beta}, ..., \mathbf{c}_s'\boldsymbol{\beta}\}$ with

$$r(\mathbf{C}) = r\begin{pmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_s' \end{pmatrix} = s.$$

Then,

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix}$$

is consistent and has a unique solution. □

Example 3.5 (continued). Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where $E(\epsilon_{ij}) = 0$. The normal equations are

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} n & n_1 & n_2 & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & 0 & \cdots & n_a \end{pmatrix}\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} = \begin{pmatrix} \sum_i\sum_j Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \vdots \\ \sum_j Y_{aj} \end{pmatrix} = \mathbf{X}'\mathbf{Y}.$$

We know that $r(\mathbf{X}) = r = a < p$ (this system can not be solved uniquely) and that $s = p - r = (a+1) - a = 1$. Thus, to augment the normal equations, we need to find $s = 1$ (jointly) nonestimable function. Take $\mathbf{c}_1' = (1, 0, 0, ..., 0)$, which produces

$$\mathbf{c}_1'\boldsymbol{\beta} = (1\ 0\ 0\ \cdots\ 0)\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} = \mu.$$

For this choice of $\mathbf{c}_1$, the augmented normal equations are

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{c}_1 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} n & n_1 & n_2 & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_a & 0 & 0 & \cdots & n_a \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} = \begin{pmatrix} \sum_i \sum_j Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \vdots \\ \sum_j Y_{aj} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ 0 \end{pmatrix}.$$

Solving this (now full rank) system produces the unique solution

$$\begin{aligned} \widehat{\mu} &= 0 \\ \widehat{\alpha}_i &= \overline{Y}_{i+} \quad i = 1, 2, ..., a. \end{aligned}$$

You'll note that this choice of $\mathbf{c}_1$ used to augment the normal equations corresponds to specifying the side condition $\mu = 0$. □

EXERCISE. Redo this example using (a) the side condition $\sum_{i=1}^{a} n_i\alpha_i = 0$, (b) the side condition $\alpha_a = 0$ (what SAS does), and (c) using another side condition.

Example 3.6. Consider the two-way fixed effects (crossed) ANOVA model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., b$, where $E(\epsilon_{ij}) = 0$. For purposes of illustration, let's take $a = b = 3$, so that $n = ab = 9$ and $p = a + b + 1 = 7$. In matrix form, $\mathbf{X}$ and $\boldsymbol{\beta}$ are

$$\mathbf{X}_{9\times 7} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_{7\times 1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

We see that $r(\mathbf{X}) = r = 5$ so that $s = p - r = 7 - 5 = 2$. The normal equations are

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 9 & 3 & 3 & 3 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 0 & 3 & 1 & 1 & 1 \\ 3 & 1 & 1 & 1 & 3 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 \\ 3 & 1 & 1 & 1 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \sum_i \sum_j Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \sum_j Y_{3j} \\ \sum_i Y_{i1} \\ \sum_i Y_{i2} \\ \sum_i Y_{i3} \end{pmatrix} = \mathbf{X}'\mathbf{Y}.$$

This system does not have a unique solution. To augment the normal equations, we will need a set of $s = 2$ linearly independent jointly nonestimable functions. From Section 3.2.2, one example of such a set is $\{\sum_i \alpha_i, \sum_j \beta_j\}$. For this choice, our matrix $\mathbf{C}$ is

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Thus, the augmented normal equations become

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} 9 & 3 & 3 & 3 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 0 & 3 & 1 & 1 & 1 \\ 3 & 1 & 1 & 1 & 3 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 \\ 3 & 1 & 1 & 1 & 0 & 0 & 3 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \sum_i \sum_j Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \sum_j Y_{3j} \\ \sum_i Y_{i1} \\ \sum_i Y_{i2} \\ \sum_i Y_{i3} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix}.$$

Solving this system produces the "estimates" of μ, α_i and β_j under the side conditions $\sum_i \alpha_i = \sum_j \beta_j = 0$. These "estimates" are

$$\begin{aligned} \widehat{\mu} &= \overline{Y}_{++} \\ \widehat{\alpha}_i &= \overline{Y}_{i+} - \overline{Y}_{++}, \quad i = 1, 2, 3 \\ \widehat{\beta}_j &= \overline{Y}_{+j} - \overline{Y}_{++}, \quad j = 1, 2, 3. \end{aligned}$$

Exercise. Redo this example using (a) the side conditions $\alpha_a = 0$ and $\beta_b = 0$ (what SAS does) and (b) using another set of side conditions.

QUESTION: In general, can we give a mathematical form for the particular solution? Note that we are now solving

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix},$$

which is equivalent to

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} \\ \mathbf{C}'\mathbf{C} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{0} \end{pmatrix}$$

since $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ iff $\mathbf{C}'\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. Thus, any solution to this system must also satisfy

$$(\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

But,

$$r(\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C}) = r\left[(\mathbf{X}' \;\; \mathbf{C}') \begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix}\right] = r\begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix} = p,$$

that is, $\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C}$ is nonsingular. Hence, the unique solution to the augmented normal equations must be

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'\mathbf{Y}.$$

So, imposing $s = p - r$ conditions $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where the elements of $\mathbf{C}\boldsymbol{\beta}$ are jointly nonestimable, yields a particular solution to the normal equations. Finally, note that by Result 2.5 (notes),

$$\begin{aligned} \mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{P}_{\mathbf{X}}\mathbf{Y}, \end{aligned}$$

which shows that

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'$$

is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. This shows that $(\mathbf{X}'\mathbf{X} + \mathbf{C}'\mathbf{C})^{-1}$ is a (non-singular) generalized inverse of $\mathbf{X}'\mathbf{X}$.

4 The Gauss-Markov Model

Complementary reading from Monahan: Chapter 4 (except Section 4.7).

4.1 Introduction

REVIEW: Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$.

- A linear estimator $t(\mathbf{Y}) = c + \mathbf{a}'\mathbf{Y}$ is said to be unbiased for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ if and only if
$$E\{t(\mathbf{Y})\} = \boldsymbol{\lambda}'\boldsymbol{\beta},$$
for all $\boldsymbol{\beta}$. We have seen this implies that $c = 0$ and $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$; i.e., $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable.

- When $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, it is possible to find several estimators that are unbiased for $\boldsymbol{\lambda}'\boldsymbol{\beta}$. For example, in the one-way (fixed effects) ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, with $E(\epsilon_{ij}) = 0$, Y_{11}, $(Y_{11} + Y_{12})/2$, and $\overline{Y}_{1+}$ are each unbiased estimators of $\boldsymbol{\lambda}'\boldsymbol{\beta} = \mu + \alpha_1$ (there are others too).

- If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then the (ordinary) least squares estimator $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is any solution to the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$, is unbiased for $\boldsymbol{\lambda}'\boldsymbol{\beta}$. To see this, recall that
$$\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} = \mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{Y},$$
where $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$, for some $\mathbf{a}$, that is, $\boldsymbol{\lambda}' \in \mathcal{R}(\mathbf{X})$, and $\mathbf{P}_{\mathbf{X}}$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. Thus,
$$E(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = E(\mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}) = \mathbf{a}'\mathbf{P}_{\mathbf{X}}E(\mathbf{Y}) = \mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{X}\boldsymbol{\beta} = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}.$$

GOAL: Among all linear unbiased estimators for $\boldsymbol{\lambda}'\boldsymbol{\beta}$, we want to find the "best" linear unbiased estimator in the sense that it has the smallest variance. We will show that the least squares estimator $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, provided that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

4.2 The Gauss-Markov Theorem

Result 4.1. Consider the Gauss-Markov model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable and let $\widehat{\boldsymbol{\beta}}$ denote any solution to the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$. The (ordinary) least squares estimator $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is the **best linear unbiased estimator (BLUE)** of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, that is, the variance of $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is uniformly less than that of any other linear unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

Proof. Suppose that $\widehat{\theta} = c + \mathbf{a}'\mathbf{Y}$ is another linear unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. From Result 3.1, we know that $c = 0$ and that $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$. Thus, $\widehat{\theta} = \mathbf{a}'\mathbf{Y}$, where $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$. Now, write $\widehat{\theta} = \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} + (\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}})$. Note that

$$\begin{aligned}\text{var}(\widehat{\theta}) &= \text{var}[\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} + (\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}})]\\ &= \text{var}(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) + \underbrace{\text{var}(\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}})}_{\geq 0} + 2\text{cov}(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}, \widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}).\end{aligned}$$

We now show that $\text{cov}(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}, \widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = 0$. Recalling that $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \mathbf{a}'\mathbf{P_X}\mathbf{Y}$, we have

$$\begin{aligned}\text{cov}(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}, \widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) &= \text{cov}(\mathbf{a}'\mathbf{P_X}\mathbf{Y}, \mathbf{a}'\mathbf{Y} - \mathbf{a}'\mathbf{P_X}\mathbf{Y})\\ &= \text{cov}[\mathbf{a}'\mathbf{P_X}\mathbf{Y}, \mathbf{a}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}]\\ &= \mathbf{a}'\mathbf{P_X}\text{cov}(\mathbf{Y}, \mathbf{Y})[\mathbf{a}'(\mathbf{I} - \mathbf{P_X})]'\\ &= \sigma^2\mathbf{I}\mathbf{a}'\mathbf{P_X}(\mathbf{I} - \mathbf{P_X})\mathbf{a} = 0,\end{aligned}$$

since $\mathbf{P_X}(\mathbf{I} - \mathbf{P_X}) = \mathbf{0}$. Thus, $\text{var}(\widehat{\theta}) \geq \text{var}(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}})$, showing that $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ has variance no larger than that of $\widehat{\theta}$. Equality results when $\text{var}(\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = 0$. However, if $\text{var}(\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = 0$, then because $E(\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = 0$ as well, $\widehat{\theta} - \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is a degenerate random variable at 0; i.e., $\text{pr}(\widehat{\theta} = \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = 1$. This establishes uniqueness. □

MULTIVARIATE CASE: Suppose that we wish to estimate simultaneously k estimable linear functions

$$\boldsymbol{\Lambda}'\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\lambda}_1'\boldsymbol{\beta} \\ \boldsymbol{\lambda}_2'\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{\lambda}_k'\boldsymbol{\beta} \end{pmatrix},$$

where $\boldsymbol{\lambda}_i' = \mathbf{a}_i'\mathbf{X}$, for some $\mathbf{a}_i$; i.e., $\boldsymbol{\lambda}_i' \in \mathcal{R}(\mathbf{X})$, for $i = 1, 2, ..., k$. We say that $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{\lambda}_i'\boldsymbol{\beta}$, $i = 1, 2, ..., k$, are each estimable. Put another way, $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{\Lambda}' = \mathbf{A}'\mathbf{X}$, for some matrix $\mathbf{A}$.

Result 4.2. Consider the Gauss-Markov model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Suppose that $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is any k-dimensional estimable vector and that $\mathbf{c}+\mathbf{A}'\mathbf{Y}$ is any vector of linear unbiased estimators of the elements of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}$ denote any solution to the normal equations. Then, the matrix $\text{cov}(\mathbf{c} + \mathbf{A}'\mathbf{Y}) - \text{cov}(\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}})$ is nonnegative definite.

Proof. It suffices to show that $\mathbf{x}'[\text{cov}(\mathbf{c} + \mathbf{A}'\mathbf{Y}) - \text{cov}(\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}})]\mathbf{x} \geq 0$, for all $\mathbf{x}$. Note that

$$\begin{aligned}\mathbf{x}'[\text{cov}(\mathbf{c} + \mathbf{A}'\mathbf{Y}) - \text{cov}(\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}})]\mathbf{x} &= \mathbf{x}'\text{cov}(\mathbf{c} + \mathbf{A}'\mathbf{Y})\mathbf{x} - \mathbf{x}'\text{cov}(\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}})\mathbf{x} \\ &= \text{var}(\mathbf{x}'\mathbf{c} + \mathbf{x}'\mathbf{A}'\mathbf{Y}) - \text{var}(\mathbf{x}'\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}).\end{aligned}$$

But $\mathbf{x}'\boldsymbol{\Lambda}'\boldsymbol{\beta} = \mathbf{x}'\mathbf{A}'\mathbf{X}\boldsymbol{\beta}$ (a scalar) is estimable since $\mathbf{x}'\mathbf{A}'\mathbf{X} \in \mathcal{R}(\mathbf{X})$. Also, $\mathbf{x}'\mathbf{c} + \mathbf{x}'\mathbf{A}'\mathbf{Y} = \mathbf{x}'(\mathbf{c}+\mathbf{A}'\mathbf{Y})$ is a linear unbiased estimator of $\mathbf{x}'\boldsymbol{\Lambda}'\boldsymbol{\beta}$. The least squares estimator of $\mathbf{x}'\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is $\mathbf{x}'\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}$. Thus, by Result 4.1, $\text{var}(\mathbf{x}'\mathbf{c} + \mathbf{x}'\mathbf{A}'\mathbf{Y}) - \text{var}(\mathbf{x}'\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}) \geq 0$. □

OBSERVATION: Consider the Gauss-Markov linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. If $\mathbf{X}$ is full rank, then $\mathbf{X}'\mathbf{X}$ is nonsingular and every linear combination of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable. The (ordinary) least squares estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It is unbiased and

$$\begin{aligned}\text{cov}(\widehat{\boldsymbol{\beta}}) = \text{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Note that this is not correct if $\mathbf{X}$ is less than full rank.

Example 4.1. Recall the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are uncorrelated random variables with mean 0 and common variance $\sigma^2 > 0$ (these are the Gauss Markov assumptions). Recall that, in matrix notation,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The least squares estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} - \widehat{\beta}_1\overline{x} \\ \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_i (x_i - \overline{x})^2} \end{pmatrix}.$$

The covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\text{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2} & -\frac{\overline{x}}{\sum_i (x_i - \overline{x})^2} \\ -\frac{\overline{x}}{\sum_i (x_i - \overline{x})^2} & \frac{1}{\sum_i (x_i - \overline{x})^2} \end{pmatrix}.$$

4.3 Estimation of σ^2 in the GM model

REVIEW: Consider the Gauss-Markov model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The best linear unbiased estimator (BLUE) for any estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is any solution to the normal equations. Clearly, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ is estimable and the BLUE of $E(\mathbf{Y})$ is

$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P_X}\mathbf{Y} = \widehat{\mathbf{Y}},$$

the perpendicular projection of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{X})$; that is, the fitted values from the least squares fit. The residuals are given by

$$\widehat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{P_X}\mathbf{Y} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y},$$

the perpendicular projection of $\mathbf{Y}$ onto $\mathcal{N}(\mathbf{X}')$. Recall that the residual sum of squares is

$$Q(\widehat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \widehat{\mathbf{e}}'\widehat{\mathbf{e}} = \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}.$$

We now turn our attention to estimating σ^2.

Result 4.3. Suppose that $\mathbf{Z}$ is a random vector with mean $E(\mathbf{Z}) = \boldsymbol{\mu}$ and covariance matrix $\text{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}$. Let $\mathbf{A}$ be nonrandom. Then

$$E(\mathbf{Z}'\mathbf{A}\mathbf{Z}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + tr(\mathbf{A}\boldsymbol{\Sigma}).$$

Proof. Note that $\mathbf{Z}'\mathbf{A}\mathbf{Z}$ is a scalar random variable; hence, $\mathbf{Z}'\mathbf{A}\mathbf{Z} = tr(\mathbf{Z}'\mathbf{A}\mathbf{Z})$. Also, recall that expectation $E(\cdot)$ and $tr(\cdot)$ are linear operators. Finally, recall that $tr(\mathbf{AB}) = tr(\mathbf{BA})$ for conformable $\mathbf{A}$ and $\mathbf{B}$. Now,

$$\begin{aligned}
E(\mathbf{Z}'\mathbf{A}\mathbf{Z}) = E[tr(\mathbf{Z}'\mathbf{A}\mathbf{Z})] &= E[tr(\mathbf{A}\mathbf{Z}\mathbf{Z}')] \\
&= tr[\mathbf{A}E(\mathbf{Z}\mathbf{Z}')] \\
&= tr[\mathbf{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')] \\
&= tr(\mathbf{A}\boldsymbol{\Sigma}) + tr(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}') \\
&= tr(\mathbf{A}\boldsymbol{\Sigma}) + tr(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \ \square
\end{aligned}$$

REMARK: Finding $\text{var}(\mathbf{Z}'\mathbf{A}\mathbf{Z})$ is more difficult; see Section 4.9 in Monahan. Considerable simplification results when $\mathbf{Z}$ follows a multivariate normal distribution.

APPLICATION: We now find an unbiased estimator of σ^2 under the GM model. Suppose that $\mathbf{Y}$ is $n \times 1$ and $\mathbf{X}$ is $n \times p$ with rank $r \leq p$. Note that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. Applying Result 4.3 directly with $\mathbf{A} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$, we have

$$\begin{aligned}
E[\mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}] &= \underbrace{(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta}}_{=\,0} + tr[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\sigma^2\mathbf{I}] \\
&= \sigma^2[tr(\mathbf{I}) - tr(\mathbf{P}_{\mathbf{X}})] \\
&= \sigma^2[n - r(\mathbf{P}_{\mathbf{X}})] = \sigma^2(n - r).
\end{aligned}$$

Thus,

$$\widehat{\sigma}^2 = (n - r)^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}$$

is an unbiased estimator of σ^2 in the GM model. In non-matrix notation,

$$\widehat{\sigma}^2 = (n - r)^{-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2,$$

where $\widehat{Y}_i$ is the least squares fitted value of Y_i.

ANOVA: Consider the Gauss-Markov model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Suppose that $\mathbf{Y}$ is $n \times 1$ and $\mathbf{X}$ is $n \times p$ with rank $r \leq p$. Recall from Chapter 2 the basic form of an ANOVA table (with corrected sums of squares):

Source	df	SS	MS	F
Model (Corrected)	$r-1$	$\text{SSR} = \mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}$	$\text{MSR} = \frac{\text{SSR}}{r-1}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Residual	$n-r$	$\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$	$\text{MSE} = \frac{\text{SSE}}{n-r}$	
Total (Corrected)	$n-1$	$\text{SST} = \mathbf{Y}'(\mathbf{I} - \mathbf{P_1})\mathbf{Y}$		

NOTES:

- The degrees of freedom associated with each SS is the rank of its appropriate perpendicular projection matrix; that is, $r(\mathbf{P_X} - \mathbf{P_1}) = r - 1$ and $r(\mathbf{I} - \mathbf{P_X}) = n - r$.

- Note that

$$\text{cov}(\widehat{\mathbf{Y}}, \widehat{\mathbf{e}}) = \text{cov}[\mathbf{P_X}\mathbf{Y}, (\mathbf{I} - \mathbf{P_X})\mathbf{Y}] = \mathbf{P_X}\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{P_X}) = \mathbf{0}.$$

  That is, the least squares fitted values are uncorrelated with the residuals.

- We have just shown that $E(\text{MSE}) = \sigma^2$. If $\mathbf{X}\boldsymbol{\beta} \notin \mathcal{C}(\mathbf{1})$, that is, the independent variables in $\mathbf{X}$ add to the model (beyond an intercept, for example), then

$$\begin{aligned} E(\text{SSR}) = E[\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}] &= (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_1})\mathbf{X}\boldsymbol{\beta} + tr[(\mathbf{P_X} - \mathbf{P_1})\sigma^2\mathbf{I}] \\ &= (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta} + \sigma^2 r(\mathbf{P_X} - \mathbf{P_1}) \\ &= (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta} + (r-1)\sigma^2. \end{aligned}$$

  Thus,

$$E(\text{MSR}) = (r-1)^{-1}E(\text{SSR}) = \sigma^2 + (r-1)^{-1}(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}.$$

- If $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{1})$, that is, the independent variables in $\mathbf{X}$ add nothing to the model, then $(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_1})\mathbf{X}\boldsymbol{\beta} = 0$ and MSR and MSE are both unbiased estimators of σ^2. If this is true, F should be close to 1. Large values of F occur when $(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_1})\mathbf{X}\boldsymbol{\beta}$ is large, that is, when $\mathbf{X}\boldsymbol{\beta}$ is "far away" from $\mathcal{C}(\mathbf{1})$, that is, when the independent variables in $\mathbf{X}$ are more relevant in explaining $E(\mathbf{Y})$.

4.4 Implications of model selection

REMARK: Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is $n \times 1$ and $\mathbf{X}$ is $n \times p$ with rank $r \leq p$. We now investigate two issues: underfitting (model misspecification) and overfitting. We assume that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, that is, our usual Gauss-Markov assumptions.

4.4.1 Underfitting (Misspecification)

UNDERFITTING: Suppose that, in truth, the "correct" model for $\mathbf{Y}$ is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The vector $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\delta}$ includes the variables and coefficients missing from $\mathbf{X}\boldsymbol{\beta}$. If the analyst uses $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ to describe the data, s/he is missing important variables that are in $\mathbf{W}$, that is, the analyst is misspecifying the true model by **underfitting**. We now examine the effect of underfitting on (a) least squares estimates of estimable functions and (b) the estimate of the error variance σ^2.

CONSEQUENCES: Suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable under $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$; i.e., $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$, for some vector $\mathbf{a}$. The least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is given by $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$. If $\mathbf{W}\boldsymbol{\delta} = \mathbf{0}$, then $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is the correct model and $E(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$. If $\mathbf{W}\boldsymbol{\delta} \neq \mathbf{0}$, then, under the correct model,

$$\begin{aligned} E(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}) = E[\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}] &= \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'E(\mathbf{Y}) \\ &= \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\delta}) \\ &= \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{W}\boldsymbol{\delta} \\ &= \mathbf{a}'\mathbf{P_X}\mathbf{X}\boldsymbol{\beta} + \mathbf{a}'\mathbf{P_X}\mathbf{W}\boldsymbol{\delta} \\ &= \boldsymbol{\lambda}'\boldsymbol{\beta} + \mathbf{a}'\mathbf{P_X}\mathbf{W}\boldsymbol{\delta}, \end{aligned}$$

showing that $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is no longer unbiased, in general. The amount of the bias depends on where $\mathbf{W}\boldsymbol{\delta}$ is. If $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\delta}$ is orthogonal to $\mathcal{C}(\mathbf{X})$, then $\mathbf{P_X}\mathbf{W}\boldsymbol{\delta} = \mathbf{0}$ and the estimation of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ with $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is unaffected. Otherwise, $\mathbf{P_X}\mathbf{W}\boldsymbol{\delta} \neq \mathbf{0}$ and the estimate of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is biased.

CONSEQUENCES: Now, let's turn to the estimation of σ^2. Under the correct model,

$$\begin{aligned} E[\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}] &= (\mathbf{X}\boldsymbol{\beta}+\mathbf{W}\boldsymbol{\delta})'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})(\mathbf{X}\boldsymbol{\beta}+\mathbf{W}\boldsymbol{\delta})+tr[(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\sigma^2\mathbf{I}] \\ &= (\mathbf{W}\boldsymbol{\delta})'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{W}\boldsymbol{\delta}+\sigma^2(n-r), \end{aligned}$$

where $r = r(\mathbf{X})$. Thus,

$$E(\text{MSE}) = \sigma^2 + (n-r)^{-1}(\mathbf{W}\boldsymbol{\delta})'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{W}\boldsymbol{\delta},$$

that is, $\widehat{\sigma}^2 = \text{MSE}$ is unbiased if and only if $\mathbf{W}\boldsymbol{\delta} \in \mathcal{C}(\mathbf{X})$.

4.4.2 Overfitting

OVERFITTING: Suppose that, in truth, the correct model for $\mathbf{Y}$ is

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, but, instead, we fit

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

that is, the extra variables in $\mathbf{X}_2$ are not needed; i.e., $\boldsymbol{\beta}_2 = \mathbf{0}$. Set $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$ and suppose that $\mathbf{X}$ and $\mathbf{X}_1$ have full column rank (i.e., a regression setting). The least squares estimator of $\boldsymbol{\beta}_1$ under the true model is $\widetilde{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$. We know that

$$\begin{aligned} E(\widetilde{\boldsymbol{\beta}}_1) &= \boldsymbol{\beta}_1 \\ \text{cov}(\widetilde{\boldsymbol{\beta}}_1) &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}. \end{aligned}$$

On the other hand, the normal equations associated with the larger (unnecessarily large) model are

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \iff \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{pmatrix}$$

and the least squares estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{pmatrix}.$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ is still unbiased in the unnecessarily large model, that is,

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_1) &= \boldsymbol{\beta}_1 \\ E(\widehat{\boldsymbol{\beta}}_2) &= \mathbf{0}. \end{aligned}$$

Thus, we assess the impact of overfitting by looking at $\text{cov}(\widehat{\boldsymbol{\beta}}_1)$. Under the unnecessarily large model,

$$\text{cov}(\widehat{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} + \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2[\mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1};$$

see Exercise A.72 (pp 268) in Monahan. Thus,

$$\text{cov}(\widehat{\boldsymbol{\beta}}_1) - \text{cov}(\widetilde{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2[\mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

- If the columns of $\mathbf{X}_2$ are each orthogonal to $\mathcal{C}(\mathbf{X}_1)$, then $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ and

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix};$$

  i.e., $\mathbf{X}'\mathbf{X}$ is block diagonal, and $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} = \widetilde{\boldsymbol{\beta}}_1$. This would mean that using the unnecessarily large model has no effect on our estimate of $\boldsymbol{\beta}_1$. However, the precision with which we can estimate σ^2 is affected since $r(\mathbf{I}-\mathbf{P}_{\mathbf{X}}) < r(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1})$; that is, we have fewer residual degrees of freedom.

- If the columns of $\mathbf{X}_2$ are not all orthogonal to $\mathcal{C}(\mathbf{X}_1)$, then

$$\text{cov}(\widehat{\boldsymbol{\beta}}_1) \neq \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

  Furthermore, as $\mathbf{X}_2$ gets "closer" to $\mathcal{C}(\mathbf{X}_1)$, then $\mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$ gets "smaller." This makes $[\mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2]^{-1}$ "larger." This makes $\text{cov}(\widehat{\boldsymbol{\beta}}_1)$ "larger."

- **Multicollinearity** occurs when $\mathbf{X}_2$ is "close" to $\mathcal{C}(\mathbf{X}_1)$. Severe multicollinearity can greatly inflate the variances of the least squares estimates. In turn, this can have a deleterious effect on inference (e.g., confidence intervals too wide, hypothesis tests with no power, predicted values with little precision, etc.). Various diagnostic measures exist to assess multicollinearity (e.g., VIFs, condition numbers, etc.); see the discussion in Monahan, pp 80-82.

4.5 The Aitken model and generalized least squares

TERMINOLOGY: The general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, $\mathbf{V}$ known, is called the **Aitken model**. It is more flexible than the Guass-Markov (GM) model, because the analyst can incorporate different correlation structures with the observed responses. The GM model is a special case of the Aitken model with $\mathbf{V} = \mathbf{I}$; that is, where responses are uncorrelated.

REMARK: In practice, $\mathbf{V}$ is rarely known; rather, it must be estimated. We will discuss this later. We assume that $\mathbf{V}$ is positive definite (pd), and, hence nonsingular, for reasons that will soon be obvious. Generalizations are possible; see Christensen (Chapter 10).

RECALL: Because $\mathbf{V}$ is symmetric, we can write $\mathbf{V}$ in its Spectral Decomposition; i.e.,

$$\mathbf{V} = \mathbf{Q}\mathbf{D}\mathbf{Q}',$$

where $\mathbf{Q}$ is orthogonal and $\mathbf{D}$ is the diagonal matrix consisting of $\lambda_1, \lambda_2, ..., \lambda_n$, the eigenvalues of $\mathbf{V}$. Because $\mathbf{V}$ is pd, we know that $\lambda_i > 0$, for each $i = 1, 2, ..., n$. The symmetric square root of $\mathbf{V}$ is

$$\mathbf{V}^{1/2} = \mathbf{Q}\mathbf{D}^{1/2}\mathbf{Q}',$$

where $\mathbf{D}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_n})$. Note that $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{V}$ and that $\mathbf{V}^{-1} = \mathbf{V}^{-1/2}\mathbf{V}^{-1/2}$, where

$$\mathbf{V}^{-1/2} = \mathbf{Q}\mathbf{D}^{-1/2}\mathbf{Q}'$$

and $\mathbf{D}^{-1/2} = \text{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, ..., 1/\sqrt{\lambda_n})$.

TRANSFORMATION: Consider the Aitken model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is known. Premultiplying by $\mathbf{V}^{-1/2}$, we get

$$\begin{aligned} & \mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\boldsymbol{\epsilon} \\ \iff \quad & \mathbf{Y}^* = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \end{aligned}$$

where $\mathbf{Y}^* = \mathbf{V}^{-1/2}\mathbf{Y}$, $\mathbf{U} = \mathbf{V}^{-1/2}\mathbf{X}$, and $\boldsymbol{\epsilon}^* = \mathbf{V}^{-1/2}\boldsymbol{\epsilon}$. It is easy to show that $\mathbf{Y}^* = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ is now a GM model. To see this, note that

$$E(\boldsymbol{\epsilon}^*) = \mathbf{V}^{-1/2}E(\boldsymbol{\epsilon}) = \mathbf{V}^{-1/2}\mathbf{0} = \mathbf{0}$$

and

$$\text{cov}(\boldsymbol{\epsilon}^*) = \mathbf{V}^{-1/2}\text{cov}(\boldsymbol{\epsilon})\mathbf{V}^{-1/2} = \mathbf{V}^{-1/2}\sigma^2\mathbf{V}\mathbf{V}^{-1/2} = \sigma^2\mathbf{I}.$$

Note also that $\mathcal{R}(\mathbf{X}) = \mathcal{R}(\mathbf{U})$, because $\mathbf{V}^{-1/2}$ is nonsingular. This means that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the Aitken model if and only if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the transformed GM model. The covariance structure on $\boldsymbol{\epsilon}$ does not affect estimability.

AITKEN EQUATIONS: In the transformed model, the normal equations are

$$\mathbf{U}'\mathbf{U}\boldsymbol{\beta} = \mathbf{U}'\mathbf{Y}^*.$$

However, note that

$$\begin{aligned} \mathbf{U}'\mathbf{U}\boldsymbol{\beta} = \mathbf{U}'\mathbf{Y}^* \quad &\iff \quad (\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} = (\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{Y} \\ &\iff \quad \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \end{aligned}$$

in the Aitken model. The equations

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

are called the **Aitken equations**. These should be compared with the normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

in the GM model. In general, we will denote by $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ and $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ the solutions to the Aitken and normal equations, respectively. "GLS" stands for generalized least squares. "OLS" stands for ordinary least squares.

GENERALIZED LEAST SQUARES: Any solution $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ to the Aitken equations is called a **generalized least squares (GLS) estimator** of $\boldsymbol{\beta}$. It is not necessarily unique (unless $\mathbf{X}$ is full rank). The solution $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ minimizes

$$Q^*(\boldsymbol{\beta}) = (\mathbf{Y}^* - \mathbf{U}\boldsymbol{\beta})'(\mathbf{Y}^* - \mathbf{U}\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

When $\mathbf{X}$ is full rank, the unique GLS estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

When $\mathbf{X}$ is not full rank, a GLS estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

NOTE: If $\mathbf{V}$ is diagonal; i.e., $\mathbf{V} = \text{diag}(v_1, v_2, ..., v_n)$, then

$$Q^*(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2,$$

where $w_i = 1/v_i$ and $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. In this situation, $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is called the **weighted least squares estimator**.

Result 4.4. Consider the Aitken model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is known. If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is the BLUE for $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

Proof. Applying the Gauss-Markov Theorem to the transformed model $\mathbf{Y}^* = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$, the GLS estimator $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is the BLUE of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ among all linear unbiased estimators involving $\mathbf{Y}^* = \mathbf{V}^{-1/2}\mathbf{Y}$. However, any linear estimator in $\mathbf{Y}$ can be obtained from $\mathbf{Y}^*$ because $\mathbf{V}^{-1/2}$ is invertible. Thus, $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is the BLUE. □

REMARK: If $\mathbf{X}$ is full rank, then estimability concerns vanish (as in the GM model) and $\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ is unique. In this case, straightforward calculations show that $E(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) = \boldsymbol{\beta}$ and

$$\text{cov}(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

Example 4.2. *Heteroscedastic regression through the origin.* Consider the regression model $Y_i = \beta x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2 g^2(x_i)$, for some real function $g(\cdot)$, and $\text{cov}(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$. For this model,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} g^2(x_1) & 0 & \cdots & 0 \\ 0 & g^2(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g^2(x_n) \end{pmatrix}.$$

The OLS estimator of $\boldsymbol{\beta} = \beta$ is given by

$$\widehat{\beta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

The GLS estimator of $\boldsymbol{\beta} = \beta$ is given by

$$\widehat{\beta}_{\text{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \frac{\sum_{i=1}^n x_i Y_i/g^2(x_i)}{\sum_{i=1}^n x_i^2/g^2(x_i)}.$$

Which one is better? Both of these estimators are unbiased, so we turn to the variances. Straightforward calculations show that

$$\begin{aligned} \text{var}(\widehat{\beta}_{\text{OLS}}) &= \frac{\sigma^2 \sum_{i=1}^n x_i^2 g^2(x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} \\ \text{var}(\widehat{\beta}_{\text{GLS}}) &= \frac{\sigma^2}{\sum_{i=1}^n x_i^2/g^2(x_i)}. \end{aligned}$$

We are thus left to compare

$$\frac{\sum_{i=1}^n x_i^2 g^2(x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} \quad \text{with} \quad \frac{1}{\sum_{i=1}^n x_i^2/g^2(x_i)}.$$

Write $x_i^2 = u_i v_i$, where $u_i = x_i g(x_i)$ and $v_i = x_i/g(x_i)$. Applying Cauchy-Schwartz's inequality, we get

$$\left(\sum_{i=1}^n x_i^2\right)^2 = \left(\sum_{i=1}^n u_i v_i\right)^2 \leq \sum_{i=1}^n u_i^2 \sum_{i=1}^n v_i^2 = \sum_{i=1}^n x_i^2 g^2(x_i) \sum_{i=1}^n x_i^2/g^2(x_i).$$

Thus,

$$\frac{1}{\sum_{i=1}^n x_i^2/g^2(x_i)} \leq \frac{\sum_{i=1}^n x_i^2 g^2(x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} \Longrightarrow \text{var}(\widehat{\beta}_{\text{GLS}}) \leq \text{var}(\widehat{\beta}_{\text{OLS}}).$$

This result should not be surprising; after all, we know that $\widehat{\beta}_{\text{GLS}}$ is BLUE. □

Result 4.5. An estimate $\widehat{\boldsymbol{\beta}}$ is a generalized least squares estimate if and only if $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{AY}$, where $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$.

Proof. The GLS estimate; i.e., the OLS estimate in the transformed model $\mathbf{Y}^* = \mathbf{U}\boldsymbol{\beta}+\boldsymbol{\epsilon}^*$, where $\mathbf{Y}^* = \mathbf{V}^{-1/2}\mathbf{Y}$, $\mathbf{U} = \mathbf{V}^{-1/2}\mathbf{X}$, and $\boldsymbol{\epsilon}^* = \mathbf{V}^{-1/2}\epsilon$, satisfies

$$\mathbf{V}^{-1/2}\mathbf{X}[(\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{X}]^{-}(\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{X}\widehat{\boldsymbol{\beta}},$$

by Result 2.5. Multiplying through by $\mathbf{V}^{1/2}$ and simplifying gives the result. □

Result 4.6. $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ is a projection matrix onto $\mathcal{C}(\mathbf{X})$.

Proof. We need to show that

(a) $\mathbf{A}$ is idempotent

(b) $\mathbf{A}\mathbf{w} \in \mathcal{C}(\mathbf{X})$, for any $\mathbf{w}$

(c) $\mathbf{A}\mathbf{z} = \mathbf{z}$, for all $\mathbf{z} \in \mathcal{C}(\mathbf{X})$.

The perpendicular projection matrix onto $\mathcal{C}(\mathbf{V}^{-1/2}\mathbf{X})$ is

$$\mathbf{V}^{-1/2}\mathbf{X}[(\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{X}]^{-}(\mathbf{V}^{-1/2}\mathbf{X})',$$

which implies that

$$\mathbf{V}^{-1/2}\mathbf{X}[(\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{X}]^{-}(\mathbf{V}^{-1/2}\mathbf{X})'\mathbf{V}^{-1/2}\mathbf{X} = \mathbf{V}^{-1/2}\mathbf{X}.$$

This can also be written as

$$\mathbf{V}^{-1/2}\mathbf{A}\mathbf{X} = \mathbf{V}^{-1/2}\mathbf{X}.$$

Premultiplying by $\mathbf{V}^{1/2}$ gives $\mathbf{A}\mathbf{X} = \mathbf{X}$. Thus,

$$\mathbf{A}\mathbf{A} = \mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{A},$$

showing that $\mathbf{A}$ is idempotent. To show (b), note $\mathbf{A}\mathbf{w} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} \in \mathcal{C}(\mathbf{X})$. To show (c), it suffices to show $\mathcal{C}(\mathbf{A}) = \mathcal{C}(\mathbf{X})$. But, $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ implies that $\mathcal{C}(\mathbf{A}) \subset \mathcal{C}(\mathbf{X})$ and $\mathbf{A}\mathbf{X} = \mathbf{X}$ implies that $\mathcal{C}(\mathbf{X}) \subset \mathcal{C}(\mathbf{A})$. □

Result 4.7. In the Aitken model, if $\mathcal{C}(\mathbf{V}\mathbf{X}) \subset \mathcal{C}(\mathbf{X})$, then the GLS and OLS estimates will be equal; i.e., OLS estimates will be BLUE in the Aitken model.

Proof. The proof proceeds by showing that $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$ when $\mathcal{C}(\mathbf{V}\mathbf{X}) \subset \mathcal{C}(\mathbf{X})$. We already know that $\mathbf{A}$ is a projection matrix onto $\mathcal{C}(\mathbf{X})$. Thus, all we have to show is that if $\mathbf{w}\perp\mathcal{C}(\mathbf{X})$, then $\mathbf{A}\mathbf{w} = \mathbf{0}$. If $\mathbf{V}$ is nonsingular, then $r(\mathbf{V}\mathbf{X}) = r(\mathbf{X})$. The only way this and $\mathcal{C}(\mathbf{V}\mathbf{X}) \subset \mathcal{C}(\mathbf{X})$ holds is if $\mathcal{C}(\mathbf{V}\mathbf{X}) = \mathcal{C}(\mathbf{X})$, in which case $\mathbf{V}\mathbf{X}\mathbf{B}_1 = \mathbf{X}$ and $\mathbf{V}\mathbf{X} = \mathbf{X}\mathbf{B}_2$, for some matrices $\mathbf{B}_1$ and $\mathbf{B}_2$. Multiplying through by $\mathbf{V}^{-1}$ gives $\mathbf{X}\mathbf{B}_1 = \mathbf{V}^{-1}\mathbf{X}$ and $\mathbf{X} = \mathbf{V}^{-1}\mathbf{X}\mathbf{B}_2$. Thus, $\mathcal{C}(\mathbf{V}^{-1}\mathbf{X}) = \mathcal{C}(\mathbf{X})$ and $\mathcal{C}(\mathbf{V}^{-1}\mathbf{X})^{\perp} = \mathcal{C}(\mathbf{X})^{\perp}$. If $\mathbf{w}\perp\mathcal{C}(\mathbf{X})$, then $\mathbf{w}\perp\mathcal{C}(\mathbf{V}^{-1}\mathbf{X})$; i.e., $\mathbf{w} \in \mathcal{N}(\mathbf{X}'\mathbf{V}^{-1})$. Since $\mathbf{A}\mathbf{w} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} = \mathbf{0}$, we are done. □

5 Distributional Theory

Complementary reading from Monahan: Chapter 5.

5.1 Introduction

PREVIEW: Consider the Gauss-Markov linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is an $n \times p$ with rank $r \leq p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\epsilon}$ is $n \times 1$ with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. In addition to the first two moment assumptions, it is common to assume that $\boldsymbol{\epsilon}$ follows a multivariate normal distribution. This additional assumption allows us to formally pursue various questions dealing with inference. In addition to the multivariate normal distribution, we will also examine noncentral distributions and quadratic forms.

RECALL: If $Z \sim \mathcal{N}(0,1)$, then the probability density function (pdf) of Z is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \mathcal{I}(z \in \mathcal{R}).$$

The $\mathcal{N}(\mu, \sigma^2)$ family is a location-scale family generated by the standard density $f_Z(z)$.

TERMINOLOGY: The collection of pdfs

$$\mathcal{LS}(f) = \left\{ f_X(\cdot|\mu,\sigma) : f_X(x|\mu,\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right); \mu \in \mathcal{R},\ \sigma > 0 \right\}$$

is a **location-scale family** generated by $f_Z(z)$; see Casella and Berger, Chapter 3. That is, if $Z \sim f_Z(z)$, then

$$X = \sigma Z + \mu \sim f_X(x|\mu,\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right).$$

APPLICATION: With the standard normal density $f_Z(z)$, it is easy to see that

$$f_X(x|\mu,\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \mathcal{I}(x \in \mathcal{R}).$$

That is, any normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ may be obtained by transforming $Z \sim \mathcal{N}(0,1)$ via $X = \sigma Z + \mu$.

5.2 Multivariate normal distribution

5.2.1 Probability density function

STARTING POINT: Suppose that $Z_1, Z_2, ..., Z_p$ are iid standard normal random variables. The joint pdf of $\mathbf{Z} = (Z_1, Z_2, ..., Z_p)'$ is given by

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= \prod_{i=1}^{p} f_Z(z_i) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\sum_{i=1}^p z_i^2/2} \prod_{i=1}^{p} \mathcal{I}(z_i \in \mathcal{R}) \\ &= (2\pi)^{-p/2} \exp(-\mathbf{z}'\mathbf{z}/2)\mathcal{I}(\mathbf{z} \in \mathcal{R}^p). \end{aligned}$$

If $\mathbf{Z}$ has pdf $f_{\mathbf{Z}}(\mathbf{z})$, we say that $\mathbf{Z}$ has a **standard multivariate normal distribution**; i.e., a multivariate normal distribution with mean $\mathbf{0}_{p\times 1}$ and covariance matrix $\mathbf{I}_p$. We write $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$.

MULTIVARIATE NORMAL DISTRIBUTION: Suppose that $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Suppose that $\mathbf{V}$ is symmetric and positive definite (and, hence, nonsingular) and let $\mathbf{V}^{1/2}$ be the symmetric square root of $\mathbf{V}$. Define the transformation

$$\mathbf{Y} = \mathbf{V}^{1/2}\mathbf{Z} + \boldsymbol{\mu},$$

where $\mathbf{Y}$ and $\boldsymbol{\mu}$ are both $p \times 1$. Note that

$$E(\mathbf{Y}) = E(\mathbf{V}^{1/2}\mathbf{Z} + \boldsymbol{\mu}) = \boldsymbol{\mu},$$

since $E(\mathbf{Z}) = \mathbf{0}$, and

$$\text{cov}(\mathbf{Y}) = \text{cov}(\mathbf{V}^{1/2}\mathbf{Z} + \boldsymbol{\mu}) = \mathbf{V}^{1/2}\text{cov}(\mathbf{Z})\mathbf{V}^{1/2} = \mathbf{V},$$

since $\text{cov}(\mathbf{Z}) = \mathbf{I}$. The transformation $\mathbf{y} = g(\mathbf{z}) = \mathbf{V}^{1/2}\mathbf{z}+\boldsymbol{\mu}$ is linear in $\mathbf{z}$ (and hence, one-to-one) and the pdf of $\mathbf{Y}$ can be found using a transformation. The inverse transformation is $\mathbf{z} = g^{-1}(\mathbf{y}) = \mathbf{V}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$. The Jacobian of the inverse transformation is

$$\left|\frac{\partial g^{-1}(\mathbf{y})}{\partial \mathbf{y}}\right| = |\mathbf{V}^{-1/2}|,$$

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. The matrix $\mathbf{V}^{-1/2}$ is pd; thus, its determinant is always positive. Thus, for $\mathbf{y} \in \mathcal{R}^p$,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= f_{\mathbf{Z}}\{g^{-1}(\mathbf{y})\}|\mathbf{V}^{-1/2}| \\ &= |\mathbf{V}|^{-1/2} f_{\mathbf{Z}}\{\mathbf{V}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})\} \\ &= (2\pi)^{-p/2}|\mathbf{V}|^{-1/2}\exp[-\{\mathbf{V}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})\}'\mathbf{V}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})/2] \\ &= (2\pi)^{-p/2}|\mathbf{V}|^{-1/2}\exp\{-(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})/2\}. \end{aligned}$$

If $\mathbf{Y} \sim f_{\mathbf{Y}}(\mathbf{y})$, we say that $\mathbf{Y}$ has a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. We write $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$.

IMPORTANT: In the preceding derivation, we assumed $\mathbf{V}$ to be pd (hence, nonsingular). If $\mathbf{V}$ is singular, then the distribution of $\mathbf{Y}$ is concentrated in a subspace of $\mathcal{R}^p$, with dimension $r(\mathbf{V})$. In this situation, the density function of $\mathbf{Y}$ does not exist.

5.2.2 Moment generating functions

REVIEW: Suppose that X is a random variable with cumulative distribution function $F_X(x) = P(X \leq x)$. If $E(e^{tX}) < \infty$ for all $|t| < \delta$, $\exists \delta > 0$, then

$$M_X(t) = E(e^{tX}) = \int_{\mathcal{R}} e^{tx} dF_X(x)$$

is defined for all t in an open neighborhood about zero. The function $M_X(t)$ is called the **moment generating function (mgf)** of X.

Result 5.1.

1. If $M_X(t)$ exists, then $E(|X|^j) < \infty$, for all $j \geq 1$, that is, the moment generating function characterizes an infinite set of moments.

2. $M_X(0) = 1$.

3. The jth moment of X is given by

$$E(X^j) = \left.\frac{\partial^j M_X(t)}{\partial t^j}\right|_{t=0}.$$

4. *Uniqueness.* If $X_1 \sim M_{X_1}(t)$, $X_2 \sim M_{X_2}(t)$, and $M_{X_1}(t) = M_{X_2}(t)$ for all t in an open neighborhood about zero, then $F_{X_1}(x) = F_{X_2}(x)$ for all x.

5. If $X_1, X_2, ..., X_n$ are independent random variables with mgfs $M_{X_i}(t)$, $i = 1, 2, ..., n$, and $Y = a_0 + \sum_{i=1}^n a_i X_i$, then
$$M_Y(t) = e^{a_0 t} \prod_{i=1}^n M_{X_i}(a_i t).$$

Result 5.2.

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $M_X(t) = \exp(\mu t + t^2\sigma^2/2)$, for all $t \in \mathcal{R}$.

2. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$.

3. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{1}{\sigma}(X - \mu) \sim \mathcal{N}(0, 1)$.

TERMINOLOGY: Define the random vector $\mathbf{X} = (X_1, X_2, ..., X_p)'$ and let $\mathbf{t} = (t_1, t_2, ..., t_p)'$. The moment generating function for $\mathbf{X}$ is given by
$$M_{\mathbf{X}}(\mathbf{t}) = E\{\exp(\mathbf{t}'\mathbf{X})\} = \int_{\mathcal{R}^p} \exp(\mathbf{t}'\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}),$$
provided that $E\{\exp(\mathbf{t}'\mathbf{X})\} < \infty$, for all $||\mathbf{t}|| < \delta$, $\exists \delta > 0$.

Result 5.3.

1. If $M_{\mathbf{X}}(\mathbf{t})$ exists, then $M_{X_i}(t_i) = M_{\mathbf{X}}(\mathbf{t}_i^*)$, where $\mathbf{t}_i^* = (0, ..., 0, t_i, 0, ..., 0)'$. This implies that $E(|X_i|^j) < \infty$, for all $j \geq 1$.

2. The expected value of $\mathbf{X}$ is
$$E(\mathbf{X}) = \left.\frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial \mathbf{t}}\right|_{\mathbf{t}=\mathbf{0}}.$$

3. The $p \times p$ second moment matrix
$$E(\mathbf{X}\mathbf{X}') = \left.\frac{\partial^2 M_{\mathbf{X}}(\mathbf{t})}{\partial \mathbf{t} \partial \mathbf{t}'}\right|_{\mathbf{t}=\mathbf{0}}.$$
Thus,
$$E(X_r X_s) = \left.\frac{\partial^2 M_{\mathbf{X}}(\mathbf{t})}{\partial t_r t_s}\right|_{t_r = t_s = 0}.$$

4. *Uniqueness.* If $\mathbf{X}_1$ and $\mathbf{X}_2$ are random vectors with $M_{\mathbf{X}_1}(\mathbf{t}) = M_{\mathbf{X}_2}(\mathbf{t})$ for all $\mathbf{t}$ in an open neighborhood about zero, then $F_{\mathbf{X}_1}(\mathbf{x}) = F_{\mathbf{X}_2}(\mathbf{x})$ for all $\mathbf{x}$.

5. If $\mathbf{X}_1, \mathbf{X}_1, ..., \mathbf{X}_n$ are independent random vectors, and
$$\mathbf{Y} = \mathbf{a}_0 + \sum_{i=1}^{n} \mathbf{A}_i \mathbf{X}_i,$$
for conformable $\mathbf{a}_0$ and $\mathbf{A}_i$; $i = 1, 2, ..., n$, then
$$M_{\mathbf{Y}}(\mathbf{t}) = \exp(\mathbf{a}_0'\mathbf{t}) \prod_{i=1}^{n} M_{\mathbf{X}_i}(\mathbf{A}_i'\mathbf{t}).$$

6. Let $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', ..., \mathbf{X}_m')'$ and suppose that $M_{\mathbf{X}}(\mathbf{t})$ exists. Let $M_{\mathbf{X}_i}(\mathbf{t}_i)$ denote the mgf of $\mathbf{X}_i$. Then, $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m$ are independent if and only if
$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n} M_{\mathbf{X}_i}(\mathbf{t}_i)$$
for all $\mathbf{t} = (\mathbf{t}_1', \mathbf{t}_2', ..., \mathbf{t}_m')'$ in an open neighborhood about zero.

Result 5.4. If $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$, then $M_{\mathbf{Y}}(\mathbf{t}) = \exp(\mathbf{t}'\boldsymbol{\mu} + \mathbf{t}'\mathbf{V}\mathbf{t}/2)$.
Proof. Exercise.

5.2.3 Properties

Result 5.5. Let $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{a}$ be $p \times 1$, $\mathbf{b}$ be $k \times 1$, and $\mathbf{A}$ be $k \times p$. Then

1. $X = \mathbf{a}'\mathbf{Y} \sim \mathcal{N}(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\mathbf{V}\mathbf{a})$.

2. $\mathbf{X} = \mathbf{A}\mathbf{Y} + \mathbf{b} \sim \mathcal{N}_k(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\mathbf{V}\mathbf{A}')$.

Result 5.6. If $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$, then any $r \times 1$ subvector of $\mathbf{Y}$ has an r-variate normal distribution with the same means, variances, and covariances as the original distribution.
Proof. Partition $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$, where $\mathbf{Y}_1$ is $r \times 1$. Partition $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ and
$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$$

accordingly. Define $\mathbf{A} = (\mathbf{I}_r\ \mathbf{0})$, where $\mathbf{0}$ is $r \times (p-r)$. Since $\mathbf{Y}_1 = \mathbf{AY}$ is a linear function of $\mathbf{Y}$, it is normally distributed. Since $\mathbf{A}\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $\mathbf{AVA}' = \mathbf{V}_{11}$, we are done. □

COROLLARY: If $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$, then $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for $i = 1, 2, ..., p$.

WARNING: Joint normality implies marginal normality. That is, if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are jointly normal, then they are marginally normal. However, if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are marginally normal, this does not necessarily mean that they are jointly normal.

APPLICATION: Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Note that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and that $\mathbf{V} = \text{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$. Furthermore, because $\mathbf{Y}$ is a linear combination of $\boldsymbol{\epsilon}$, it is also normally distributed; i.e., $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. With $\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$, we know that $\widehat{\mathbf{Y}} = \mathbf{P_X}\mathbf{Y}$ and $\widehat{\mathbf{e}} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y}$. Now,

$$E(\widehat{\mathbf{Y}}) = E(\mathbf{P_X}\mathbf{Y}) = \mathbf{P_X}E(\mathbf{Y}) = \mathbf{P_X}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$$

and

$$\text{cov}(\widehat{\mathbf{Y}}) = \text{cov}(\mathbf{P_X}\mathbf{Y}) = \mathbf{P_X}\text{cov}(\mathbf{Y})\mathbf{P}_\mathbf{X}' = \sigma^2\mathbf{P_X}\mathbf{I}\mathbf{P_X} = \sigma^2\mathbf{P_X},$$

since $\mathbf{P_X}$ is symmetric and idempotent. Also, $\widehat{\mathbf{Y}} = \mathbf{P_X}\mathbf{Y}$ is a linear combination of $\mathbf{Y}$, so it also has normal distribution. Putting everything together, we have

$$\widehat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{P_X}).$$

Exercise: Show that $\widehat{\mathbf{e}} \sim \mathcal{N}_n\{\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{P_X})\}$.

5.2.4 Less-than-full-rank normal distributions

TERMINOLOGY: The random vector $\mathbf{Y}_{p\times 1} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ is said to have a p-variate normal distribution with rank k if $\mathbf{Y}$ has the same distribution as $\boldsymbol{\mu}_{p\times 1} + \boldsymbol{\Gamma}'_{p\times k}\mathbf{Z}_{k\times 1}$, where $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \mathbf{V}$, $r(\mathbf{V}) = k < p$, and $\mathbf{Z} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I})$.

Example 5.1. Suppose that $k = 1$, $Z_1 \sim \mathcal{N}(0,1)$, and $\mathbf{Y} = (Y_1, Y_2)'$, where

$$\mathbf{Y} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} Z_1 = \begin{pmatrix} \gamma_1 Z_1 \\ \gamma_2 Z_1 \end{pmatrix},$$

where $\mathbf{\Gamma}' = (\gamma_1, \gamma_2)'$ and $r(\mathbf{\Gamma}) = 1$. Since $r(\mathbf{\Gamma}) = 1$, this means that at least one of γ_1 and γ_2 is not equal to zero. Without loss, take $\gamma_1 \neq 0$, in which case

$$Y_2 = \frac{\gamma_2}{\gamma_1} Y_1.$$

Note that $E(\mathbf{Y}) = \mathbf{0} = (0,0)'$ and

$$\text{cov}(\mathbf{Y}) = E(\mathbf{Y}\mathbf{Y}') = E\left\{ \begin{pmatrix} \gamma_1^2 Z_1^2 & \gamma_1\gamma_2 Z_1^2 \\ \gamma_1\gamma_2 Z_1^2 & \gamma_2^2 Z_1^2 \end{pmatrix} \right\} = \begin{pmatrix} \gamma_1^2 & \gamma_1\gamma_2 \\ \gamma_1\gamma_2 & \gamma_2^2 \end{pmatrix} = \mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{V}.$$

Note that $|\mathbf{V}| = 0$. Thus, $\mathbf{Y}_{2\times 1}$ is a random vector with all of its probability mass located in the linear subspace $\{(y_1, y_2) : y_2 = \gamma_2 y_1/\gamma_1\}$. Since $r(\mathbf{V}) = 1 < 2$, $\mathbf{Y}$ does not have a density function. □

5.2.5 Independence results

Result 5.7. Suppose that $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_m \end{pmatrix}, \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1m} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{m1} & \mathbf{V}_{m2} & \cdots & \mathbf{V}_{mm} \end{pmatrix}.$$

Then, $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_m$ are jointly independent if and only if $\mathbf{V}_{ij} = \mathbf{0}$, for all $i \neq j$.

Proof. Sufficiency ($\Longrightarrow$): Suppose $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_m$ are jointly independent. For all $i \neq j$,

$$\begin{aligned} \mathbf{V}_{ij} &= E\{(\mathbf{Y}_i - \boldsymbol{\mu}_i)(\mathbf{Y}_j - \boldsymbol{\mu}_j)'\} \\ &= E(\mathbf{Y}_i - \boldsymbol{\mu}_i)E\{(\mathbf{Y}_j - \boldsymbol{\mu}_j)'\} = \mathbf{0}. \end{aligned}$$

Necessity ($\Longleftarrow$): Suppose that $\mathbf{V}_{ij} = \mathbf{0}$ for all $i \neq j$, and let $\mathbf{t} = (\mathbf{t}_1', \mathbf{t}_2', ..., \mathbf{t}_m')'$. Note that

$$\mathbf{t}'\mathbf{V}\mathbf{t} = \sum_{i=1}^{m} \mathbf{t}_i'\mathbf{V}_{ii}\mathbf{t}_i \quad \text{and} \quad \mathbf{t}'\boldsymbol{\mu} = \sum_{i=1}^{m} \mathbf{t}_i'\boldsymbol{\mu}_i.$$

Thus,

$$\begin{aligned}
M_{\mathbf{Y}}(\mathbf{t}) &= \exp(\mathbf{t}'\boldsymbol{\mu} + \mathbf{t}'\mathbf{V}\mathbf{t}/2) \\
&= \exp\left(\sum_{i=1}^{m}\mathbf{t}_i'\boldsymbol{\mu}_i + \sum_{i=1}^{m}\mathbf{t}_i'\mathbf{V}_{ii}\mathbf{t}_i/2\right) \\
&= \prod_{i=1}^{m}\exp(\mathbf{t}_i'\boldsymbol{\mu}_i + \mathbf{t}_i'\mathbf{V}_{ii}\mathbf{t}_i/2) = \prod_{i=1}^{m} M_{\mathbf{Y}_i}(\mathbf{t}_i). \ \Box
\end{aligned}$$

Result 5.8. Suppose that $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{Y}_1 = \mathbf{a}_1 + \mathbf{B}_1\mathbf{X}$ and $\mathbf{Y}_2 = \mathbf{a}_2 + \mathbf{B}_2\mathbf{X}$, for nonrandom conformable $\mathbf{a}_i$ and $\mathbf{B}_i$; $i = 1, 2$. Then, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if $\mathbf{B}_1\boldsymbol{\Sigma}\mathbf{B}_2' = \mathbf{0}$.

Proof. Write $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)'$ as

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{X}.$$

Thus, $\mathbf{Y}$ is a linear combination of $\mathbf{X}$; hence, $\mathbf{Y}$ follows a multivariate normal distribution (i.e., $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are **jointly normal**). Also, $\text{cov}(\mathbf{Y}_1, \mathbf{Y}_2) = \text{cov}(\mathbf{B}_1\mathbf{X}, \mathbf{B}_2\mathbf{X}) = \mathbf{B}_1\boldsymbol{\Sigma}\mathbf{B}_2'$. Now simply apply Result 5.7. $\Box$

REMARK: If $\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, and $\text{cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, this does not necessarily mean that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent! We need $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ to be jointly normal.

APPLICATION: Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We have already seen that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Also, note that with $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$,

$$\begin{pmatrix} \widehat{\mathbf{Y}} \\ \widehat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{\mathbf{X}} \\ \mathbf{I} - \mathbf{P}_{\mathbf{X}} \end{pmatrix}\mathbf{Y},$$

a linear combination of $\mathbf{Y}$. Thus, $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{e}}$ are jointly normal. By the last result, we know that $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{e}}$ are independent since

$$\text{cov}(\widehat{\mathbf{Y}}, \widehat{\mathbf{e}}) = \mathbf{P}_{\mathbf{X}}\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})' = \mathbf{0}.$$

That is, the fitted values and residuals from the least-squares fit are independent. This explains why residual plots that display nonrandom patterns are consistent with a violation of our model assumptions.

5.2.6 Conditional distributions

RECALL: Suppose that $(X,Y)' \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

and $\rho = \text{corr}(X,Y)$. The conditional distribution of Y, given X, is also normally distributed, more precisely,

$$Y|\{X=x\} \sim \mathcal{N}\left\{\mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X), \sigma_Y^2(1-\rho^2)\right\}.$$

It is important to see that the conditional mean $E(Y|X=x)$ is a linear function of x. Note also that the conditional variance $\text{var}(Y|X=x)$ is free of x.

EXTENSION: We wish to extend the previous result to random vectors. In particular, suppose that $\mathbf{X}$ and $\mathbf{Y}$ are jointly multivariate normal with $\boldsymbol{\Sigma}_{XY} \neq \mathbf{0}$. That is, suppose

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left\{\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_Y \end{pmatrix}\right\},$$

and assume that $\boldsymbol{\Sigma}_X$ is nonsingular. The conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is

$$\mathbf{Y}|\{\mathbf{X}=\mathbf{x}\} \sim \mathcal{N}(\boldsymbol{\mu}_{Y|X}, \boldsymbol{\Sigma}_{Y|X}),$$

where

$$\boldsymbol{\mu}_{Y|X} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}(\mathbf{x} - \boldsymbol{\mu}_X)$$

and

$$\boldsymbol{\Sigma}_{Y|X} = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}.$$

Again, the conditional mean $\boldsymbol{\mu}_{Y|X}$ is a linear function of $\mathbf{x}$ and the conditional covariance matrix $\boldsymbol{\Sigma}_{Y|X}$ is free of $\mathbf{x}$.

5.3 Noncentral χ^2 distribution

RECALL: Suppose that $U \sim \chi^2_n$; that is, U has a (central) χ^2 distribution with $n > 0$ degrees of freedom. The pdf of U is given by

$$f_U(u|n) = \frac{1}{\Gamma(\frac{n}{2})2^{n/2}} u^{\frac{n}{2}-1} e^{-u/2} I(u > 0).$$

The χ^2_n family of distributions is a gamma(α, β) subfamily with shape parameter $\alpha = n/2$ and scale parameter $\beta = 2$. Note that $E(U) = n$, var$(U) = 2n$, and $M_U(t) = (1-2t)^{-n/2}$, for $t < 1/2$.

RECALL: If $Z_1, Z_2, ..., Z_n$ are iid $\mathcal{N}(0, 1)$, then $U_1 = Z_1^2 \sim \chi^2_1$ and

$$\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

Proof. Exercise.

TERMINOLOGY: A univariate random variable V is said to have a **noncentral χ^2 distribution** with degrees of freedom $n > 0$ and noncentrality parameter $\lambda > 0$ if it has the pdf

$$f_V(v|n,\lambda) = \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^j}{j!}\right) \underbrace{\frac{1}{\Gamma(\frac{n+2j}{2})2^{(n+2j)/2}} v^{\frac{n+2j}{2}-1} e^{-v/2} I(v > 0)}_{f_U(v|n+2j)}.$$

We write $V \sim \chi^2_n(\lambda)$. When $\lambda = 0$, the $\chi^2_n(\lambda)$ distribution reduces to the central χ^2_n distribution. In the $\chi^2_n(\lambda)$ pdf, notice that $e^{-\lambda}\lambda^j/j!$ is the jth term of a Poisson pmf with parameter $\lambda > 0$.

Result 5.9. If $V|W \sim \chi^2_{n+2W}$ and $W \sim$ Poisson(λ), then $V \sim \chi^2_n(\lambda)$.

Proof. Note that

$$\begin{aligned} f_V(v) &= \sum_{j=0}^{\infty} f_{V,W}(v,j) \\ &= \sum_{j=0}^{\infty} f_W(j) f_{V|W}(v|j) \\ &= \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^j}{j!}\right) \frac{1}{\Gamma(\frac{n+2j}{2})2^{(n+2j)/2}} v^{\frac{n+2j}{2}-1} e^{-v/2} I(v > 0). \ \square \end{aligned}$$

Result 5.10. If $V \sim \chi^2_n(\lambda)$, then

$$M_V(t) = (1-2t)^{-n/2}\exp\left(\frac{2t\lambda}{1-2t}\right),$$

for $t < 1/2$.

Proof. The mgf of V, by definition, is $M_V(t) = E(e^{tV})$. Using an iterated expectation, we can write, for $t < 1/2$,

$$M_V(t) = E(e^{tV}) = E\{E(e^{tV}|W)\},$$

where $W \sim \text{Poisson}(\lambda)$. Note that $E(e^{tV}|W)$ is the conditional mgf of V, given W. We know that $V|W \sim \chi^2_{n+2W}$; thus, $E(e^{tV}|W) = (1-2t)^{-(n+2W)/2}$ and

$$\begin{aligned}
E\{E(e^{tV}|W)\} &= \sum_{j=0}^{\infty}(1-2t)^{-(n+2j)/2}\left(\frac{e^{-\lambda}\lambda^j}{j!}\right) \\
&= e^{-\lambda}(1-2t)^{-n/2}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}(1-2t)^{-j} \\
&= e^{-\lambda}(1-2t)^{-n/2}\sum_{j=0}^{\infty}\frac{\left(\frac{\lambda}{1-2t}\right)^j}{j!} \\
&= e^{-\lambda}(1-2t)^{-n/2}\exp\left(\frac{\lambda}{1-2t}\right) \\
&= (1-2t)^{-n/2}\exp\left(\frac{2t\lambda}{1-2t}\right). \ \square
\end{aligned}$$

MEAN AND VARIANCE: If $V \sim \chi^2_n(\lambda)$, then

$$E(V) = n + 2\lambda \quad \text{and} \quad \text{var}(V) = 2n + 8\lambda.$$

Result 5.11. If $Y \sim \mathcal{N}(\mu, 1)$, then $U = Y^2 \sim \chi^2_1(\lambda)$, where $\lambda = \mu^2/2$.

Outline of the proof. The proof proceeds by finding the mgf of U and showing that it equals

$$M_U(t) = (1-2t)^{-n/2}\exp\left(\frac{2t\lambda}{1-2t}\right),$$

with $n = 1$ and $\lambda = \mu^2/2$. Note that

$$M_U(t) = E(e^{tU}) = E(e^{tY^2}) = \int_{\mathcal{R}} e^{ty^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-\mu)^2}dy$$

Now, combine the exponents in the integrand, square out the $(y-\mu)^2$ term, combine like terms, complete the square, and collapse the expression to $(1-2t)^{-1/2}\exp\{\mu^2 t/(1-2t)\}$ times some normal density that is integrated over $\mathcal{R}$. □

Result 5.12. If $U_1, U_2, ..., U_m$ are independent random variables, where $U_i \sim \chi^2_{n_i}(\lambda_i)$; $i = 1, 2, ..., m$, then $U = \sum_i U_i \sim \chi^2_n(\lambda)$, where $n = \sum_i n_i$ and $\lambda = \sum_i \lambda_i$.

Result 5.13. Suppose that $V \sim \chi^2_n(\lambda)$. For fixed n and $c > 0$, the quantity $P_\lambda(V > c)$ is a strictly increasing function of λ.
Proof. See Monahan, pp 106-108. □

IMPLICATION: If $V_1 \sim \chi^2_n(\lambda_1)$ and $V_2 \sim \chi^2_n(\lambda_2)$, where $\lambda_2 > \lambda_1$, then $\text{pr}(V_2 > c) > \text{pr}(V_1 > c)$. That is, V_2 is **(strictly) stochastically greater** than V_1, written $V_2 >_{\text{st}} V_1$. Note that

$$V_2 >_{\text{st}} V_1 \Longleftrightarrow F_{V_2}(v) < F_{V_1}(v) \Longleftrightarrow S_{V_2}(v) > S_{V_1}(v),$$

for all v, where $F_{V_i}(\cdot)$ denotes the cdf of V_i and $S_{V_i}(\cdot) = 1 - F_{V_i}(\cdot)$ denotes the **survivor function** of V_i.

5.4 Noncentral F distribution

RECALL: A univariate random variable W is said to have a (central) F distribution with degrees of freedom $n_1 > 0$ and $n_2 > 0$ if it has the pdf

$$f_W(w|n_1, n_2) = \frac{\Gamma(\frac{n_1+n_2}{2})\left(\frac{n_1}{n_2}\right)^{n_1/2} w^{(n_1-2)/2}}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})\left(1+\frac{n_1 w}{n_2}\right)^{(n_1+n_2)/2}} I(w > 0).$$

We write $W \sim F_{n_1,n_2}$. The moment generating function for the F distribution does not exist in closed form.

RECALL: If U_1 and U_2 are independent central χ^2 random variables with degrees of freedom n_1 and n_2, respectively, then

$$W = \frac{U_1/n_1}{U_2/n_2} \sim F_{n_1,n_2}.$$

TERMINOLOGY: A univariate random variable W is said to have a **noncentral** F **distribution** with degrees of freedom $n_1 > 0$ and $n_2 > 0$ and noncentrality parameter $\lambda > 0$ if it has the pdf

$$f_W(w|n_1, n_2, \lambda) = \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^j}{j!} \right) \frac{\Gamma(\frac{n_1+2j+n_2}{2}) \left(\frac{n_1+2j}{n_2}\right)^{(n_1+2j)/2} w^{(n_1+2j-2)/2}}{\Gamma(\frac{n_1+2j}{2})\Gamma(\frac{n_2}{2}) \left(1 + \frac{n_1 w}{n_2}\right)^{(n_1+2j+n_2)/2}} I(w > 0).$$

We write $W \sim F_{n_1,n_2}(\lambda)$. When $\lambda = 0$, the noncentral F distribution reduces to the central F distribution.

MEAN AND VARIANCE: If $W \sim F_{n_1,n_2}(\lambda)$, then

$$E(W) = \frac{n_2}{n_2 - 2}\left(1 + \frac{2\lambda}{n_1}\right)$$

and

$$\text{var}(W) = \frac{2n_2^2}{n_1^2(n_2 - 2)} \left\{ \frac{(n_1 + 2\lambda)^2}{(n_2 - 2)(n_2 - 4)} + \frac{n_1 + 4\lambda}{n_2 - 4} \right\}.$$

$E(W)$ exists only when $n_2 > 2$ and $\text{var}(W)$ exists only when $n_2 > 4$. The moment generating function for the noncentral F distribution does not exist in closed form.

Result 5.14. If U_1 and U_2 are independent random variables with $U_1 \sim \chi^2_{n_1}(\lambda)$ and $U_2 \sim \chi^2_{n_2}$, then

$$W = \frac{U_1/n_1}{U_2/n_2} \sim F_{n_1,n_2}(\lambda).$$

Proof. See Searle, pp 51-52.

Result 5.15. Suppose that $W \sim F_{n_1,n_2}(\lambda)$. For fixed n_1, n_2, and $c > 0$, the quantity $P_\lambda(W > c)$ is a strictly increasing function of λ. That is, if $W_1 \sim F_{n_1,n_2}(\lambda_1)$ and $W_2 \sim F_{n_1,n_2}(\lambda_2)$, where $\lambda_2 > \lambda_1$, then $\text{pr}(W_2 > c) > \text{pr}(W_1 > c)$; i.e., $W_2 >_{\text{st}} W_1$.

REMARK: The fact that the noncentral F distribution tends to be larger than the central F distribution is the basis for many of the tests used in linear models. Typically, test statistics are used that have a central F distribution if the null hypothesis is true and a noncentral F distribution if the null hypothesis is not true. Since the noncentral F distribution tends to be larger, large values of the test statistic are consistent with the alternative hypothesis. Thus, the form of an appropriate rejection region is to reject H_0 for large values of the test statistic. The power is simply the probability of rejection region (defined under H_0) when the probability distribution is noncentral F. Noncentral F distributions are available in most software packages.

5.5 Distributions of quadratic forms

GOAL: We would like to find the distribution of $\mathbf{Y}'\mathbf{A}\mathbf{Y}$, where $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$. We will obtain this distribution by taking steps. Result 5.16 is a very small step. Result 5.17 is a large step, and Result 5.18 is the finish line. There is no harm in assuming that $\mathbf{A}$ is symmetric.

Result 5.16. Suppose that $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$ and define

$$W = \mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{I}\mathbf{Y} = \sum_{i=1}^{p} Y_i^2.$$

Result 5.11 says $Y_i^2 \sim \chi_1^2(\mu_i^2/2)$, for $i = 1, 2, ..., p$. Thus, from Result 5.12,

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2(\boldsymbol{\mu}'\boldsymbol{\mu}/2).$$

LEMMA: The $p \times p$ symmetric matrix $\mathbf{A}$ is idempotent of rank s if and only if there exists a $p \times s$ matrix $\mathbf{P}_1$ such that (a) $\mathbf{A} = \mathbf{P}_1\mathbf{P}_1'$ and (b) $\mathbf{P}_1'\mathbf{P}_1 = \mathbf{I}_s$.
Proof. ($\Longleftarrow$) Suppose that $\mathbf{A} = \mathbf{P}_1\mathbf{P}_1'$ and $\mathbf{P}_1'\mathbf{P}_1 = \mathbf{I}_s$. Clearly, $\mathbf{A}$ is symmetric. Also,

$$\mathbf{A}^2 = \mathbf{P}_1\mathbf{P}_1'\mathbf{P}_1\mathbf{P}_1' = \mathbf{P}_1\mathbf{P}_1' = \mathbf{A}.$$

Note also that $r(\mathbf{A}) = tr(\mathbf{A}) = tr(\mathbf{P}_1\mathbf{P}_1') = tr(\mathbf{P}_1'\mathbf{P}_1) = tr(\mathbf{I}_s) = s$. Now, to go the other way ($\Longrightarrow$), suppose that $\mathbf{A}$ is a symmetric, idempotent matrix of rank s. The spectral decomposition of $\mathbf{A}$ is given by $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}'$, where $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ and $\mathbf{Q}$ is orthogonal. Since $\mathbf{A}$ is idempotent, we know that s of the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ are equal to 1 and other $p - s$ eigenvalues are equal to 0. Thus, we can write

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}' = \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{pmatrix} = \mathbf{P}_1\mathbf{P}_1'.$$

Thus, we have shown that (a) holds. To show that (b) holds, note that because $\mathbf{Q}$ is orthogonal,

$$\mathbf{I}_p = \mathbf{Q}'\mathbf{Q} = \begin{pmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1'\mathbf{P}_1 & \mathbf{P}_1'\mathbf{P}_2 \\ \mathbf{P}_2'\mathbf{P}_1 & \mathbf{P}_2'\mathbf{P}_2 \end{pmatrix}.$$

It is easy to convince yourself that $\mathbf{P}_1'\mathbf{P}_1$ is an identity matrix. Its dimension is $s \times s$ because $tr(\mathbf{P}_1'\mathbf{P}_1) = tr(\mathbf{P}_1\mathbf{P}_1') = tr(\mathbf{A}) = r(\mathbf{A})$, which equals s by assumption. $\square$

Result 5.17. Suppose that $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$. If $\mathbf{A}$ is idempotent of rank s, then $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi_s^2(\lambda)$, where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

Proof. Suppose that $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$ and that $\mathbf{A}$ is idempotent of rank s. By the last lemma, we know that $\mathbf{A} = \mathbf{P}_1\mathbf{P}_1'$, where $\mathbf{P}_1$ is $p \times s$, and $\mathbf{P}_1'\mathbf{P}_1 = \mathbf{I}_s$. Thus,

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{Y}'\mathbf{P}_1\mathbf{P}_1'\mathbf{Y} = \mathbf{X}'\mathbf{X},$$

where $\mathbf{X} = \mathbf{P}_1'\mathbf{Y}$. Since $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I})$, and since $\mathbf{X} = \mathbf{P}_1'\mathbf{Y}$ is a linear combination of $\mathbf{Y}$, we know that

$$\mathbf{X} \sim \mathcal{N}_s(\mathbf{P}_1'\boldsymbol{\mu}, \mathbf{P}_1'\mathbf{I}\mathbf{P}_1) \sim \mathcal{N}_s(\mathbf{P}_1'\boldsymbol{\mu}, \mathbf{I}_s).$$

Result 5.16 says that $\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{X}'\mathbf{X} \sim \chi_s^2\{(\mathbf{P}_1'\boldsymbol{\mu})'(\mathbf{P}_1'\boldsymbol{\mu})/2\}$. But,

$$\lambda \equiv \frac{1}{2}(\mathbf{P}_1'\boldsymbol{\mu})'\mathbf{P}_1'\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{P}_1\mathbf{P}_1'\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \ \square$$

Result 5.18. Suppose that $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$, where $r(\mathbf{V}) = p$. If $\mathbf{A}\mathbf{V}$ is idempotent of rank s, then $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi_s^2(\lambda)$, where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

Proof. Since $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$, and since $\mathbf{X} = \mathbf{V}^{-1/2}\mathbf{Y}$ is a linear combination of $\mathbf{Y}$, we know that

$$\mathbf{X} \sim \mathcal{N}_p(\mathbf{V}^{-1/2}\boldsymbol{\mu}, \mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2}) \sim \mathcal{N}_p(\mathbf{V}^{-1/2}\boldsymbol{\mu}, \mathbf{I}_p).$$

Now,

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{X}'\mathbf{V}^{1/2}\mathbf{A}\mathbf{V}^{1/2}\mathbf{X} = \mathbf{X}'\mathbf{B}\mathbf{X},$$

where $\mathbf{B} = \mathbf{V}^{1/2}\mathbf{A}\mathbf{V}^{1/2}$. Recall that $\mathbf{V}^{1/2}$ is the symmetric square root of $\mathbf{V}$. From Result 5.17, we know that $\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{X}'\mathbf{B}\mathbf{X} \sim \chi_s^2(\lambda)$ if $\mathbf{B}$ is idempotent of rank s. However, note that

$$r(\mathbf{B}) = r(\mathbf{V}^{1/2}\mathbf{A}\mathbf{V}^{1/2}) = r(\mathbf{A}) = r(\mathbf{A}\mathbf{V}) = s,$$

since $\mathbf{AV}$ has rank s (by assumption) and $\mathbf{V}$ and $\mathbf{V}^{1/2}$ are both nonsingular. Also, $\mathbf{AV}$ is idempotent by assumption so that

$$\begin{aligned} \mathbf{AV} = \mathbf{AVAV} \quad &\Rightarrow \quad \mathbf{A} = \mathbf{AVA} \\ &\Rightarrow \quad \mathbf{V}^{1/2}\mathbf{AV}^{1/2} = \mathbf{V}^{1/2}\mathbf{AV}^{1/2}\mathbf{V}^{1/2}\mathbf{AV}^{1/2} \\ &\Rightarrow \quad \mathbf{B} = \mathbf{BB}. \end{aligned}$$

Thus, $\mathbf{B}$ is idempotent of rank s. This implies that $\mathbf{Y}'\mathbf{AY} = \mathbf{X}'\mathbf{BX} \sim \chi^2_s(\lambda)$. Noting that

$$\lambda = \frac{1}{2}(\mathbf{V}^{-1/2}\boldsymbol{\mu})'\mathbf{B}(\mathbf{V}^{-1/2}\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{AV}^{1/2}\mathbf{V}^{-1/2}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

completes the argument. □

Example 5.2. Suppose that $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)' \sim \mathcal{N}_n(\mu\mathbf{1}, \sigma^2\mathbf{I})$, so that $\boldsymbol{\mu} = \mu\mathbf{1}$ and $\mathbf{V} = \sigma^2\mathbf{I}$, where $\mathbf{1}$ is $n \times 1$ and $\mathbf{I}$ is $n \times n$. The statistic

$$(n-1)S^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{J})\mathbf{Y} = \mathbf{Y}'\mathbf{AY},$$

where $\mathbf{A} = \mathbf{I} - n^{-1}\mathbf{J}$. Thus, consider the quantity

$$\frac{(n-1)S^2}{\sigma^2} = \mathbf{Y}'\mathbf{BY},$$

where $\mathbf{B} = \sigma^{-2}\mathbf{A} = \sigma^{-2}(\mathbf{I} - n^{-1}\mathbf{J})$. Note that $\mathbf{BV} = \sigma^{-2}(\mathbf{I} - n^{-1}\mathbf{J})\sigma^2\mathbf{I} = \mathbf{I} - n^{-1}\mathbf{J} = \mathbf{A}$, which is idempotent with rank

$$r(\mathbf{BV}) = tr(\mathbf{BV}) = tr(\mathbf{A}) = tr(\mathbf{I} - n^{-1}\mathbf{J}) = n - n^{-1}n = n - 1.$$

Result 5.18 says that $(n-1)S^2/\sigma^2 = \mathbf{Y}'\mathbf{BY} \sim \chi^2_{n-1}(\lambda)$, where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\mu}$. However,

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\mu} = \frac{1}{2}(\mu\mathbf{1})'\sigma^{-2}(\mathbf{I} - n^{-1}\mathbf{J})\mu\mathbf{1} = 0,$$

since $\mu\mathbf{1} \in \mathcal{C}(\mathbf{1})$ and $\mathbf{I} - n^{-1}\mathbf{J}$ is the ppm onto $\mathcal{C}(\mathbf{1})^\perp$. Thus,

$$\frac{(n-1)S^2}{\sigma^2} = \mathbf{Y}'\mathbf{BY} \sim \chi^2_{n-1},$$

a central χ^2 distribution with $n-1$ degrees of freedom. □

Example 5.3. Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ denote the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. We know that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Consider the (uncorrected) partitioning of the sums of squares given by

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{P_X}\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}.$$

- We first consider the residual sum of squares $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$. Dividing this quantity by σ^2, we get

  $$\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2 = \mathbf{Y}'\{\sigma^{-2}(\mathbf{I} - \mathbf{P_X})\}\mathbf{Y} = \mathbf{Y}'\mathbf{A}\mathbf{Y},$$

  where $\mathbf{A} = \sigma^{-2}(\mathbf{I} - \mathbf{P_X})$. With $\mathbf{V} = \sigma^2\mathbf{I}$, note that

  $$\mathbf{AV} = \sigma^{-2}(\mathbf{I} - \mathbf{P_X})\sigma^2\mathbf{I} = \mathbf{I} - \mathbf{P_X},$$

  an idempotent matrix with rank

  $$r(\mathbf{I} - \mathbf{P_X}) = tr(\mathbf{I} - \mathbf{P_X}) = tr(\mathbf{I}) - tr(\mathbf{P_X}) = tr(\mathbf{I}) - r(\mathbf{P_X}) = n - r,$$

  since $r(\mathbf{P_X}) = r(\mathbf{X}) = r$, by assumption. Result 5.18 says that

  $$\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2 = \mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi^2_{n-r}(\lambda),$$

  where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$. However,

  $$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \frac{1}{2}(\mathbf{X}\boldsymbol{\beta})'\sigma^{-2}(\mathbf{I} - \mathbf{P_X})\mathbf{X}\boldsymbol{\beta} = 0,$$

  because $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$ and $\mathbf{I} - \mathbf{P_X}$ projects onto the orthogonal complement $\mathcal{N}(\mathbf{X}')$. Thus, we have shown that $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2 \sim \chi^2_{n-r}$, a central χ^2 distribution with $n - r$ degrees of freedom.

- Now, we turn our attention to the (uncorrected) model sum of squares $\mathbf{Y}'\mathbf{P_X}\mathbf{Y}$. Dividing this quantity by σ^2, we get

  $$\mathbf{Y}'\mathbf{P_X}\mathbf{Y}/\sigma^2 = \mathbf{Y}'(\sigma^{-2}\mathbf{P_X})\mathbf{Y} = \mathbf{Y}'\mathbf{B}\mathbf{Y},$$

where $\mathbf{B} = \sigma^{-2}\mathbf{P_X}$. With $\mathbf{V} = \sigma^2\mathbf{I}$, note that

$$\mathbf{BV} = \sigma^{-2}\mathbf{P_X}\sigma^2\mathbf{I} = \mathbf{P_X},$$

an idempotent matrix with rank $r(\mathbf{P_X}) = r(\mathbf{X}) = r$. Result 5.18 says that

$$\mathbf{Y}'\mathbf{P_X}\mathbf{Y}/\sigma^2 = \mathbf{Y}'\mathbf{BY} \sim \chi^2_r(\lambda),$$

where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\mu}$. Note that

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\mu} = \frac{1}{2}(\mathbf{X}\boldsymbol{\beta})'\sigma^{-2}\mathbf{P_X}\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2.$$

That is, $\mathbf{Y}'\mathbf{P_X}\mathbf{Y}/\sigma^2$ has a noncentral χ^2 distribution with r degrees of freedom and noncentrality parameter $\lambda = (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$.

In the last calculation, note that $\lambda = (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2 = 0$ iff $\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. In this case, both quadratic forms $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2$ and $\mathbf{Y}'\mathbf{P_X}\mathbf{Y}/\sigma^2$ have central χ^2 distributions. □

5.6 Independence of quadratic forms

GOALS: In this subsection, we consider two problems. With $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, we would like to establish sufficient conditions for

(a) $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{BY}$ to be independent, and

(b) $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ to be independent.

Result 5.19. Suppose that $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$. If $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{BY}$ are independent.

Proof. We may assume that $\mathbf{A}$ is symmetric. Write $\mathbf{A} = \mathbf{QDQ}'$, where $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ and $\mathbf{Q}$ is orthogonal. We know that $s \leq p$ of the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ are nonzero where $s = r(\mathbf{A})$. We can thus write

$$\mathbf{A} = \mathbf{QDQ}' = \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{pmatrix} = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1',$$

where $\mathbf{D}_1 = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_s)$. Thus,

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{Y}'\mathbf{P}_1\mathbf{D}_1\mathbf{P}_1'\mathbf{Y} = \mathbf{X}'\mathbf{D}_1\mathbf{X},$$

where $\mathbf{X} = \mathbf{P}_1'\mathbf{Y}$. Notice that

$$\begin{pmatrix} \mathbf{BY} \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{B} \\ \mathbf{P}_1' \end{pmatrix} \mathbf{Y} \sim \mathcal{N}\left\{ \begin{pmatrix} \mathbf{B}\boldsymbol{\mu} \\ \mathbf{P}_1'\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{BVB}' & \mathbf{BVP}_1 \\ \mathbf{P}_1'\mathbf{VB}' & \mathbf{P}_1'\mathbf{VP}_1 \end{pmatrix} \right\}.$$

Suppose that $\mathbf{BVA} = \mathbf{0}$. Then,

$$\mathbf{0} = \mathbf{BVA} = \mathbf{BVP}_1\mathbf{D}_1\mathbf{P}_1' = \mathbf{BVP}_1\mathbf{D}_1\mathbf{P}_1'\mathbf{P}_1.$$

But, because $\mathbf{Q}$ is orthogonal,

$$\mathbf{I}_p = \mathbf{Q}'\mathbf{Q} = \begin{pmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1'\mathbf{P}_1 & \mathbf{P}_1'\mathbf{P}_2 \\ \mathbf{P}_2'\mathbf{P}_1 & \mathbf{P}_2'\mathbf{P}_2 \end{pmatrix}.$$

This implies that $\mathbf{P}_1'\mathbf{P}_1 = \mathbf{I}_s$, and, thus,

$$\mathbf{0} = \mathbf{BVP}_1\mathbf{D}_1\mathbf{P}_1'\mathbf{P}_1 = \mathbf{BVP}_1\mathbf{D}_1 = \mathbf{BVP}_1\mathbf{D}_1\mathbf{D}_1^{-1} = \mathbf{BVP}_1.$$

Therefore, $\text{cov}(\mathbf{BY}, \mathbf{X}) = \mathbf{0}$, that is, $\mathbf{X}$ and $\mathbf{BY}$ are independent. But, $\mathbf{Y}'\mathbf{AY} = \mathbf{X}'\mathbf{D}_1\mathbf{X}$, a function of $\mathbf{X}$. Thus, $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{BY}$ are independent as well. □

Example 5.4. Suppose that $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)' \sim \mathcal{N}_n(\mu\mathbf{1}, \sigma^2\mathbf{I})$, where $\mathbf{1}$ is $n \times 1$ and $\mathbf{I}$ is $n \times n$, so that $\boldsymbol{\mu} = \mu\mathbf{1}$ and $\mathbf{V} = \sigma^2\mathbf{I}$. Recall that

$$(n-1)S^2 = \mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{J})\mathbf{Y} = \mathbf{YAY},$$

where $\mathbf{A} = \mathbf{I} - n^{-1}\mathbf{J}$. Also,

$$\overline{Y} = n^{-1}\mathbf{1}'\mathbf{Y} = \mathbf{BY},$$

where $\mathbf{B} = n^{-1}\mathbf{1}'$. These two statistics are independent because

$$\mathbf{BVA} = n^{-1}\mathbf{1}'\sigma^2\mathbf{I}(\mathbf{I} - n^{-1}\mathbf{J}) = \sigma^2 n^{-1}\mathbf{1}'(\mathbf{I} - n^{-1}\mathbf{J}) = \mathbf{0},$$

because $\mathbf{I} - n^{-1}\mathbf{J}$ is the ppm onto $\mathcal{C}(\mathbf{1})^\perp$. Since functions of independent statistics are also independent, $\overline{Y}$ and S^2 are also independent. □

Result 5.20. Suppose that $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$. If $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are independent.

Proof. Write $\mathbf{A}$ and $\mathbf{B}$ in their spectral decompositions; that is, write

$$\mathbf{A} = \mathbf{PDP}' = \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{pmatrix} = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1',$$

where $\mathbf{D}_1 = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_s)$ and $s = r(\mathbf{A})$. Similarly, write

$$\mathbf{B} = \mathbf{QRQ}' = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{pmatrix} \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_1' \\ \mathbf{Q}_2' \end{pmatrix} = \mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1',$$

where $\mathbf{R}_1 = \text{diag}(\gamma_1, \gamma_2, ..., \gamma_t)$ and $t = r(\mathbf{B})$. Since $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal, this implies that $\mathbf{P}_1'\mathbf{P}_1 = \mathbf{I}_s$ and $\mathbf{Q}_1'\mathbf{Q}_1 = \mathbf{I}_t$. Suppose that $\mathbf{BVA} = \mathbf{0}$. Then,

$$\begin{aligned} \mathbf{0} = \mathbf{BVA} &= \mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1'\mathbf{V}\mathbf{P}_1\mathbf{D}_1\mathbf{P}_1' \\ &= \mathbf{Q}_1'\mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1'\mathbf{V}\mathbf{P}_1\mathbf{D}_1\mathbf{P}_1'\mathbf{P}_1 \\ &= \mathbf{R}_1\mathbf{Q}_1'\mathbf{V}\mathbf{P}_1\mathbf{D}_1 \\ &= \mathbf{R}_1^{-1}\mathbf{R}_1\mathbf{Q}_1'\mathbf{V}\mathbf{P}_1\mathbf{D}_1\mathbf{D}_1^{-1} \\ &= \mathbf{Q}_1'\mathbf{V}\mathbf{P}_1 \\ &= \text{cov}(\mathbf{Q}_1'\mathbf{Y}, \mathbf{P}_1'\mathbf{Y}). \end{aligned}$$

Now,

$$\begin{pmatrix} \mathbf{P}_1' \\ \mathbf{Q}_1' \end{pmatrix} \mathbf{Y} \sim \mathcal{N}\left\{ \begin{pmatrix} \mathbf{P}_1'\boldsymbol{\mu} \\ \mathbf{Q}_1'\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{P}_1'\mathbf{V}\mathbf{P}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_1'\mathbf{V}\mathbf{Q}_1 \end{pmatrix} \right\}.$$

That is, $\mathbf{P}_1'\mathbf{Y}$ and $\mathbf{Q}_1'\mathbf{Y}$ are jointly normal and uncorrelated; thus, they are independent. So are $\mathbf{Y}'\mathbf{P}_1\mathbf{D}_1\mathbf{P}_1'\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1'\mathbf{Y}$. But $\mathbf{A} = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1'$ and $\mathbf{B} = \mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1'$, so we are done. □

Example 5.5. Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ denote the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$. We know that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. In

Example 5.3, we showed that

$$\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}/\sigma^2 \sim \chi^2_{n-r},$$

and that

$$\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}/\sigma^2 \sim \chi^2_r(\lambda),$$

where $\lambda = (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$. Note that

$$\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}/\sigma^2 = \mathbf{Y}'\{\sigma^{-2}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\}\mathbf{Y} = \mathbf{Y}'\mathbf{A}\mathbf{Y},$$

where $\mathbf{A} = \sigma^{-2}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})$. Also,

$$\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}/\sigma^2 = \mathbf{Y}'(\sigma^{-2}\mathbf{P}_{\mathbf{X}})\mathbf{Y} = \mathbf{Y}'\mathbf{B}\mathbf{Y},$$

where $\mathbf{B} = \sigma^{-2}\mathbf{P}_{\mathbf{X}}$. Applying Result 5.20, we have

$$\mathbf{BVA} = \sigma^{-2}\mathbf{P}_{\mathbf{X}}\sigma^2\mathbf{I}\sigma^{-2}(\mathbf{I}-\mathbf{P}_{\mathbf{X}}) = \mathbf{0},$$

that is, $\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}/\sigma^2$ and $\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}/\sigma^2$ are independent quadratic forms. Thus, the statistic

$$\begin{aligned} F &= \frac{\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}/r}{\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}/(n-r)} \\ &= \frac{\sigma^{-2}\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}/r}{\sigma^{-2}\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}/(n-r)} \sim F_{r,n-r}\left\{(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\right\}, \end{aligned}$$

a noncentral F distribution with degrees of freedom r (numerator) and $n-r$ (denominator) and noncentrality parameter $\lambda = (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$.

OBSERVATIONS:

- Note that if $\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, then $F \sim F_{r,n-r}$ since the noncentrality parameter $\lambda = 0$.

- On the other hand, as the length of $\mathbf{X}\boldsymbol{\beta}$ gets larger, so does λ. This shifts the noncentral $F_{r,n-r}\left\{(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\right\}$ distribution to the right, because the noncentral F distribution is stochastically increasing in its noncentrality parameter.

- Therefore, large values of F are consistent with large values of $||\mathbf{X}\boldsymbol{\beta}||$.

5.7 Cochran's Theorem

REMARK: An important general notion in linear models is that sums of squares like $\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Y}$ can be "broken down" into sums of squares of smaller pieces. We now discuss **Cochran's Theorem** (Result 5.21), which serves to explain why this is possible.

Result 5.21. Suppose that $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Suppose that $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_k$ are $n \times n$ symmetric and idempotent matrices, where $r(\mathbf{A}_i) = s_i$, for $i = 1, 2, ..., k$. If $\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_k = \mathbf{I}_n$, then $\mathbf{Y}'\mathbf{A}_1\mathbf{Y}/\sigma^2, \mathbf{Y}'\mathbf{A}_2\mathbf{Y}/\sigma^2, ..., \mathbf{Y}'\mathbf{A}_k\mathbf{Y}/\sigma^2$ follow independent $\chi^2_{s_i}(\lambda_i)$ distributions, where $\lambda_i = \boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}/2\sigma^2$, for $i = 1, 2, ..., k$, and $\sum_{i=1}^k s_i = n$.
Outline of proof. If $\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_k = \mathbf{I}_n$, then $\mathbf{A}_i$ idempotent implies that both (a) $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$, for $i \neq j$, and (b) $\sum_{i=1}^k s_i = n$ hold. That $\mathbf{Y}'\mathbf{A}_i\mathbf{Y}/\sigma^2 \sim \chi^2_{s_i}(\lambda_i)$ follows from Result 5.18 with $\mathbf{V} = \sigma^2\mathbf{I}$. Because $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$, Result 5.20 with $\mathbf{V} = \sigma^2\mathbf{I}$ guarantees that $\mathbf{Y}'\mathbf{A}_i\mathbf{Y}/\sigma^2$ and $\mathbf{Y}'\mathbf{A}_j\mathbf{Y}/\sigma^2$ are independent. □

IMPORTANCE: We now show how Cochran's Threorem can be used to deduce the joint distribution of the sums of squares in an analysis of variance. Suppose that we partition the design matrix $\mathbf{X}$ and the parameter vector $\boldsymbol{\beta}$ in $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ into $k+1$ parts, so that

$$\mathbf{Y} = (\mathbf{X}_0 \ \mathbf{X}_1 \ \cdots \ \mathbf{X}_k)\begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix},$$

where the dimensions of $\mathbf{X}_i$ and $\boldsymbol{\beta}_i$ are $n \times p_i$ and $p_i \times 1$, respectively, and $\sum_{i=0}^k p_i = p$. We can now write this as a $k+1$ part model (the full model):

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}.$$

Now consider fitting each of the k submodels

$$\begin{aligned}
\mathbf{Y} &= \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \\
&\vdots \\
\mathbf{Y} &= \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1} + \boldsymbol{\epsilon},
\end{aligned}$$

and let $R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_i)$ denote the regression (model) sum of squares from fitting the ith submodel, for $i = 0, 1, ..., k$; that is,

$$\begin{aligned} R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_i) &= \mathbf{Y}'(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)\left[(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)'(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)\right]^{-}(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{P}_{\mathbf{X}_i^*}\mathbf{Y}, \end{aligned}$$

where

$$\mathbf{P}_{\mathbf{X}_i^*} = (\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)\left[(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)'(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)\right]^{-}(\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)'$$

is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X}_i^*)$, where $\mathbf{X}_i^* = (\mathbf{X}_0\mathbf{X}_1 \cdots \mathbf{X}_i)$ and $i = 0, 1, ..., k$. Clearly,

$$\mathcal{C}(\mathbf{X}_0^*) \subset \mathcal{C}(\mathbf{X}_1^*) \subset \cdots \subset \mathcal{C}(\mathbf{X}_{k-1}^*) \subset \mathcal{C}(\mathbf{X}).$$

We can now partition the total sums of squares as

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} &= R(\boldsymbol{\beta}_0) + [R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) - R(\boldsymbol{\beta}_0)] + [R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)] + \cdots \\ &\quad + [R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k) - R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{k-1})] + [\mathbf{Y}'\mathbf{Y} - R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)] \\ &= \mathbf{Y}'\mathbf{A}_0\mathbf{Y} + \mathbf{Y}'\mathbf{A}_1\mathbf{Y} + \mathbf{Y}'\mathbf{A}_2\mathbf{Y} + \cdots + \mathbf{Y}'\mathbf{A}_k\mathbf{Y} + \mathbf{Y}'\mathbf{A}_{k+1}\mathbf{Y}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_0 &= \mathbf{P}_{\mathbf{X}_0^*} \\ \mathbf{A}_i &= \mathbf{P}_{\mathbf{X}_i^*} - \mathbf{P}_{\mathbf{X}_{i-1}^*} \quad i = 1, 2, ..., k-1, \\ \mathbf{A}_k &= \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_{k-1}^*} \\ \mathbf{A}_{k+1} &= \mathbf{I} - \mathbf{P}_{\mathbf{X}}. \end{aligned}$$

Note that $\mathbf{A}_0 + \mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_{k+1} = \mathbf{I}$. Note also that the $\mathbf{A}_i$ matrices are symmetric, for $i = 0, 1, ..., k+1$, and that

$$\begin{aligned} \mathbf{A}_i^2 &= (\mathbf{P}_{\mathbf{X}_i^*} - \mathbf{P}_{\mathbf{X}_{i-1}^*})(\mathbf{P}_{\mathbf{X}_i^*} - \mathbf{P}_{\mathbf{X}_{i-1}^*}) \\ &= \mathbf{P}_{\mathbf{X}_i^*}^2 - \mathbf{P}_{\mathbf{X}_{i-1}^*} - \mathbf{P}_{\mathbf{X}_{i-1}^*} + \mathbf{P}_{\mathbf{X}_{i-1}^*}^2 \\ &= \mathbf{P}_{\mathbf{X}_i^*} - \mathbf{P}_{\mathbf{X}_{i-1}^*} = \mathbf{A}_i, \end{aligned}$$

for $i = 1, 2, ..., k$, since $\mathbf{P}_{\mathbf{X}_i^*}\mathbf{P}_{\mathbf{X}_{i-1}^*} = \mathbf{P}_{\mathbf{X}_{i-1}^*}$ and $\mathbf{P}_{\mathbf{X}_{i-1}^*}\mathbf{P}_{\mathbf{X}_i^*} = \mathbf{P}_{\mathbf{X}_{i-1}^*}$. Thus, $\mathbf{A}_i$ is idempotent for $i = 1, 2, ..., k$. However, clearly $\mathbf{A}_0 = \mathbf{P}_{\mathbf{X}_0^*} = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-}\mathbf{X}_0'$ and $\mathbf{A}_{k+1} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$

are also idempotent. Let

$$\begin{aligned} s_0 = r(\mathbf{A}_0) &= r(\mathbf{X}_0) \\ s_i = r(\mathbf{A}_i) &= tr(\mathbf{A}_i) = tr(\mathbf{P}_{\mathbf{X}_i^*}) - tr(\mathbf{P}_{\mathbf{X}_{i-1}^*}) = r(\mathbf{P}_{\mathbf{X}_i^*}) - r(\mathbf{P}_{\mathbf{X}_{i-1}^*}), \quad i = 1, 2, ..., k, \\ s_{k+1} = r(\mathbf{A}_{k+1}) &= n - r(\mathbf{X}). \end{aligned}$$

It is easy to see that $\sum_{i=0}^{k+1} s_i = n$.

APPLICATION: Consider the Gauss-Markov model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$ so that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Cochran's Theorem applies and we have

$$\begin{aligned} \frac{1}{\sigma^2}\mathbf{Y}'\mathbf{A}_0\mathbf{Y} &\sim \chi^2_{s_0}[\lambda_0 = (\mathbf{X}\boldsymbol{\beta})'\mathbf{A}_0\mathbf{X}\boldsymbol{\beta}/2\sigma^2] \\ \frac{1}{\sigma^2}\mathbf{Y}'\mathbf{A}_i\mathbf{Y} &\sim \chi^2_{s_i}[\lambda_i = (\mathbf{X}\boldsymbol{\beta})'\mathbf{A}_i\mathbf{X}\boldsymbol{\beta}/2\sigma^2], \quad i = 1, 2, ..., k, \\ \frac{1}{\sigma^2}\mathbf{Y}'\mathbf{A}_{k+1}\mathbf{Y} = \frac{1}{\sigma^2}\mathbf{Y}'(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} &\sim \chi^2_{s_{k+1}}, \end{aligned}$$

where $s_{k+1} = n - r(\mathbf{X})$. Note that the last quadratic follows a central χ^2 distribution because $\lambda_{k+1} = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{X}\boldsymbol{\beta}/2\sigma^2 = 0$. Cochran's Theorem also guarantees that the quadratic forms $\mathbf{Y}'\mathbf{A}_0\mathbf{Y}/\sigma^2, \mathbf{Y}'\mathbf{A}_1\mathbf{Y}/\sigma^2, ..., \mathbf{Y}'\mathbf{A}_k\mathbf{Y}/\sigma^2, \mathbf{Y}'\mathbf{A}_{k+1}\mathbf{Y}/\sigma^2$ are independent.

ANOVA TABLE: The quadratic forms $\mathbf{Y}'\mathbf{A}_i\mathbf{Y}$, for $i = 0, 1, ..., k+1$, and the degrees of freedom $s_i = r(\mathbf{A}_i)$ are often presented in the following ANOVA table:

Source	df	SS	Noncentrality
$\boldsymbol{\beta}_0$	s_0	$R(\boldsymbol{\beta}_0)$	$\lambda_0 = (\mathbf{X}\boldsymbol{\beta})'\mathbf{A}_0\mathbf{X}\boldsymbol{\beta}/2\sigma^2$
$\boldsymbol{\beta}_1$ (after $\boldsymbol{\beta}_0$)	s_1	$R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) - R(\boldsymbol{\beta}_0)$	$\lambda_1 = (\mathbf{X}\boldsymbol{\beta})'\mathbf{A}_1\mathbf{X}\boldsymbol{\beta}/2\sigma^2$
$\boldsymbol{\beta}_2$ (after $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$)	s_2	$R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$	$\lambda_2 = (\mathbf{X}\boldsymbol{\beta})'\mathbf{A}_2\mathbf{X}\boldsymbol{\beta}/2\sigma^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{\beta}_k$ (after $\boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_{k-1}$)	s_k	$R(\boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_k) - R(\boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_{k-1})$	$\lambda_k = (\mathbf{X}\boldsymbol{\beta})'\mathbf{A}_k\mathbf{X}\boldsymbol{\beta}/2\sigma^2$
Residual	s_{k+1}	$\mathbf{Y}'\mathbf{Y} - R(\boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_k)$	$\lambda_{k+1} = 0$
Total	n	$\mathbf{Y}'\mathbf{Y}$	$(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$

Note that if $\mathbf{X}_0 = \mathbf{1}$, then $\boldsymbol{\beta}_0 = \mu$, $R(\boldsymbol{\beta}_0) = \mathbf{Y}'\mathbf{P}_\mathbf{1}\mathbf{Y} = n\overline{Y}^2$. The $R(\cdot)$ notation will come in handy when we talk about hypothesis testing later. The sums of squares

$R(\boldsymbol{\beta}_0), R(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1)-R(\boldsymbol{\beta}_0), ..., R(\boldsymbol{\beta}_0,...,\boldsymbol{\beta}_k)-R(\boldsymbol{\beta}_0,...,\boldsymbol{\beta}_{k-1})$ are called the **sequential sums of squares**. These correspond to the Type I sums of squares printed out by SAS in the ANOVA and GLM procedures. We will also use the notation

$$R(\boldsymbol{\beta}_i|\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{i-1}) = R(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_i) - R(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{i-1}).$$

Example 5.6. Consider the one-way (fixed effects) analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, ..., a$ and $j = 1, 2, ..., n_i$, where $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$ random variables. In matrix form, $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ are

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{an_a} \end{pmatrix}, \quad \mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{1}_{n_a} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix},$$

where $p = a + 1$ and $n = \sum_i n_i$. Note that we can write

$$\mathbf{X} = (\mathbf{X}_0 \ \mathbf{X}_1),$$

where $\mathbf{X}_0 = \mathbf{1}$,

$$\mathbf{X}_1 = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_a} & \mathbf{0}_{n_a} & \cdots & \mathbf{1}_{n_a} \end{pmatrix},$$

and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1')'$, where $\boldsymbol{\beta}_0 = \mu$ and $\boldsymbol{\beta}_1 = (\alpha_1, \alpha_2, ..., \alpha_a)'$. That is, we can express this model in the form

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}.$$

The submodel is

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\epsilon},$$

where $\mathbf{X}_0 = \mathbf{1}$ and $\boldsymbol{\beta}_0 = \mu$. We have

$$\begin{aligned} \mathbf{A}_0 &= \mathbf{P_1} \\ \mathbf{A}_1 &= \mathbf{P_X} - \mathbf{P_1} \\ \mathbf{A}_2 &= \mathbf{I} - \mathbf{P_X}. \end{aligned}$$

These matrices are clearly symmetric and idempotent. Also, note that

$$s_0 + s_1 + s_2 = r(\mathbf{P_1}) + r(\mathbf{P_X} - \mathbf{P_1}) + r(\mathbf{I} - \mathbf{P_X}) = 1 + (a-1) + (n-a) = n$$

and that $\mathbf{A}_0 + \mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}$. Therefore, Cochran's Theorem applies and we have

$$\begin{aligned} \frac{1}{\sigma^2}\mathbf{Y}'\mathbf{P_1}\mathbf{Y} &\sim \chi^2_1[\lambda_0 = (\mathbf{X}\boldsymbol{\beta})'\mathbf{P_1}\mathbf{X}\boldsymbol{\beta}/2\sigma^2] \\ \frac{1}{\sigma^2}\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y} &\sim \chi^2_{a-1}[\lambda_1 = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_1})\mathbf{X}\boldsymbol{\beta}/2\sigma^2] \\ \frac{1}{\sigma^2}\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y} &\sim \chi^2_{n-a}. \end{aligned}$$

Cochran's Theorem also guarantees the quadratic forms $\mathbf{Y}'\mathbf{P_1}\mathbf{Y}/\sigma^2, \mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}/\sigma^2$, and $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2$ are independent. The sums of squares, using our new notation, are

$$\begin{aligned} \mathbf{Y}'\mathbf{P_1}\mathbf{Y} &= R(\mu) \\ \mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y} &= R(\mu, \alpha_1, ..., \alpha_a) - R(\mu) \equiv R(\alpha_1, ..., \alpha_a|\mu) \\ \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y} &= \mathbf{Y}'\mathbf{Y} - R(\mu, \alpha_1, ..., \alpha_a). \end{aligned}$$

The ANOVA table for this one-way analysis of variance model is

Source	df	SS	Noncentrality
μ	1	$R(\mu)$	$\lambda_0 = (\mathbf{X}\boldsymbol{\beta})'\mathbf{P_1}\mathbf{X}\boldsymbol{\beta}/2\sigma^2$
$\alpha_1, ..., \alpha_a$ (after μ)	$a-1$	$R(\mu, \alpha_1, ..., \alpha_a) - R(\mu)$	$\lambda_1 = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_1})\mathbf{X}\boldsymbol{\beta}/2\sigma^2$
Residual	$n-a$	$\mathbf{Y}'\mathbf{Y} - R(\mu, \alpha_1, ..., \alpha_a)$	0
Total	n	$\mathbf{Y}'\mathbf{Y}$	$(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$

F STATISTIC: Because

$$\frac{1}{\sigma^2}\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y} \sim \chi^2_{a-1}(\lambda_1)$$

and

$$\frac{1}{\sigma^2}\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y} \sim \chi^2_{n-a},$$

and because these two quadratic forms are independent, it follows that

$$F = \frac{\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_1})\mathbf{Y}/(a-1)}{\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}/(n-a)} \sim F_{a-1,n-a}(\lambda_1).$$

This is the usual F statistic to test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$.

- If H_0 is true, then $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{1})$ and $\lambda_1 = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_1})\mathbf{X}\boldsymbol{\beta}/2\sigma^2 = 0$ since $\mathbf{P_X}-\mathbf{P_1}$ is the ppm onto $\mathcal{C}(\mathbf{1})^{\perp}_{\mathcal{C}(\mathbf{X})}$. In this case, $F \sim F_{a-1,n-a}$, a central F distribution. A level α rejection region is therefore $\text{RR} = \{F : F > F_{a-1,n-a,\alpha}\}$.

- If H_0 is not true, then $\mathbf{X}\boldsymbol{\beta} \notin \mathcal{C}(\mathbf{1})$ and $\lambda_1 = (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_1})\mathbf{X}\boldsymbol{\beta}/2\sigma^2 > 0$. In this case, $F \sim F_{a-1,n-a}(\lambda_1)$, which is stochastically larger than the central $F_{a-1,n-a}$ distribution. Therefore, if H_0 is not true, we would expect F to be large.

EXPECTED MEAN SQUARES: We already know that

$$\text{MSE} = (n-a)^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}$$

is an unbiased estimator of σ^2, i.e., $E(\text{MSE}) = \sigma^2$. Recall that if $V \sim \chi^2_n(\lambda)$, then $E(V) = n + 2\lambda$. Therefore,

$$E\left[\frac{1}{\sigma^2}\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_1})\mathbf{Y}\right] = (a-1) + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_1})\mathbf{X}\boldsymbol{\beta}/\sigma^2$$

and

$$\begin{aligned} E(\text{MSR}) = E\left[(a-1)^{-1}\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_1})\mathbf{Y}\right] &= \sigma^2 + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_1})\mathbf{X}\boldsymbol{\beta}/(a-1) \\ &= \sigma^2 + \sum_{i=1}^{a} n_i(\alpha_i - \overline{\alpha}_+)^2/(a-1), \end{aligned}$$

where $\overline{\alpha}_+ = a^{-1}\sum_{i=1}^{a}\alpha_i$. Again, note that if $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ is true, then MSR is also an unbiased estimator of σ^2. Therefore, values of

$$F = \frac{\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_1})\mathbf{Y}/(a-1)}{\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}/(n-a)}$$

should be close to 1 when H_0 is true and larger than 1 otherwise.

6 Statistical Inference

Complementary reading from Monahan: Chapter 6 (and revisit Sections 3.9 and 4.7).

6.1 Estimation

PREVIEW: Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Note that this is the usual Gauss-Markov model with the additional assumption of normality. With this additional assumption, we can now rigorously pursue questions that deal with statistical inference. We start by examining minimum variance unbiased estimation and maximum likelihood estimation.

SUFFICIENCY: Under the assumptions stated above, we know that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Set $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$. The pdf of $\mathbf{Y}$, for all $\mathbf{y} \in \mathcal{R}^n$, is given by

$$\begin{aligned}
f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2\} \\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{-\mathbf{y}'\mathbf{y}/2\sigma^2 + \mathbf{y}'\mathbf{X}\boldsymbol{\beta}/\sigma^2 - (\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\} \\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{-(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\}\exp\{-\mathbf{y}'\mathbf{y}/2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}/\sigma^2\} \\
&= h(\mathbf{y})c(\boldsymbol{\theta})\exp\{w_1(\boldsymbol{\theta})t_1(\mathbf{y}) + w_2(\boldsymbol{\theta})t_2(\mathbf{y})\},
\end{aligned}$$

where $h(\mathbf{y}) = (2\pi)^{-n/2}I(\mathbf{y} \in \mathcal{R}^n)$, $c(\boldsymbol{\theta}) = (\sigma^2)^{-n/2}\exp\{-(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\}$, and

$$\begin{aligned}
w_1(\boldsymbol{\theta}) &= -1/2\sigma^2 \qquad & t_1(\mathbf{y}) &= \mathbf{y}'\mathbf{y} \\
w_2(\boldsymbol{\theta}) &= \boldsymbol{\beta}/\sigma^2 \qquad & t_2(\mathbf{y}) &= \mathbf{X}'\mathbf{y},
\end{aligned}$$

that is, $\mathbf{Y}$ has pdf in the exponential family (see Casella and Berger, Chapter 3). The family is full rank (i.e., it is not curved), so we know that $\mathbf{T}(\mathbf{Y}) = (\mathbf{Y}'\mathbf{Y}, \mathbf{X}'\mathbf{Y})$ is a complete sufficient statistic for $\boldsymbol{\theta}$. We also know that minimum variance unbiased estimators (MVUEs) of functions of $\boldsymbol{\theta}$ are unbiased functions of $\mathbf{T}(\mathbf{Y})$.

Result 6.1. Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. The MVUE for an estimable function $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is given by $\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ is any solution to the normal equations. The MVUE for σ^2 is

$$\begin{aligned} \text{MSE} &= (n-r)^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y} \\ &= (n-r)^{-1}(\mathbf{Y}'\mathbf{Y}-\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}), \end{aligned}$$

where $\mathbf{P}_{\mathbf{X}}$ is the perpendicular projection matrix onto $\mathcal{C}(\mathbf{X})$.

Proof. Both $\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}$ and MSE are unbiased estimators of $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ and σ^2, respectively. These estimators are also functions of $\mathbf{T}(\mathbf{Y}) = (\mathbf{Y}'\mathbf{Y}, \mathbf{X}'\mathbf{Y})$, the complete sufficient statistic. Thus, each estimator is the MVUE for its expected value. □

MAXIMUM LIKELIHOOD: Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. The likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ is

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2\}.$$

Maximum likelihood estimators for $\boldsymbol{\beta}$ and σ^2 are found by maximizing

$$\log L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2$$

with respect to $\boldsymbol{\beta}$ and σ^2. For every value of σ^2, maximizing the loglikelihood is the same as minimizing $Q(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, that is, the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y},$$

is also an MLE. Now substitute $(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{y}$ in for $Q(\boldsymbol{\beta})$ and differentiate with respect to σ^2. The MLE of σ^2 is

$$\widehat{\sigma}^2_{\text{MLE}} = n^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Y}.$$

Note that the MLE for σ^2 is biased. The MLE is rarely used in practice; MSE is the conventional estimator for σ^2.

INVARIANCE: Under the normal GM model, the MLE for an estimable function $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is $\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is any solution to the normal equations. This is true because of the invariance property of maximum likelihood estimators (see, e.g., Casella and Berger, Chapter 7). If $\boldsymbol{\Lambda}'\boldsymbol{\beta}$ is estimable, recall that $\boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}}$ is unique even if $\widehat{\boldsymbol{\beta}}$ is not.

6.2 Testing models

PREVIEW: We now provide a general discussion on testing reduced versus full models within a Gauss Markov linear model framework. Assuming normality will allow us to derive the sampling distribution of the resulting test statistic.

PROBLEM: Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $r(\mathbf{X}) = r \leq p$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Note that these are our usual GM model assumptions. For the purposes of this discussion, we assume that this model (the full model) is a "correct" model for the data. Consider also the linear model

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ and $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. We call this a reduced model because the estimation space is smaller than in the full model. Our goal is to test whether or not the reduced model is also correct.

- If the reduced model is also correct, there is no reason not to use it. Smaller models are easier to interpret and fewer degrees of freedom are spent in estimating σ^2. Thus, there are practical and statistical advantages to using the reduced model if it is also correct.

- Hypothesis testing in linear models essentially reduces to putting a constraint on the estimation space $\mathcal{C}(\mathbf{X})$ in the full model. If $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{X})$, then the $\mathbf{W}\boldsymbol{\gamma}$ model is a reparameterization of the $\mathbf{X}\boldsymbol{\beta}$ model and there is nothing to test.

RECALL: Let $\mathbf{P}_\mathbf{W}$ and $\mathbf{P}_\mathbf{X}$ denote the perpendicular projection matrices onto $\mathcal{C}(\mathbf{W})$ and $\mathcal{C}(\mathbf{X})$, respectively. Because $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$, we know that $\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W}$ is the ppm onto $\mathcal{C}(\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{W}) = \mathcal{C}(\mathbf{W})^\perp_{\mathcal{C}(\mathbf{X})}$.

GEOMETRY: In a general reduced-versus-full model testing framework, we start by assuming the full model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is essentially "correct" so that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$.

If the reduced model is also correct, then $E(\mathbf{Y}) = \mathbf{W}\boldsymbol{\gamma} \in \mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. Geometrically, performing a reduced-versus-full model test therefore requires the analyst to decide whether $E(\mathbf{Y})$ is more likely to be in $\mathcal{C}(\mathbf{W})$ or $\mathcal{C}(\mathbf{X}) - \mathcal{C}(\mathbf{W})$. Under the full model, our estimate for $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ is $\mathbf{P_X Y}$. Under the reduced model, our estimate for $E(\mathbf{Y}) = \mathbf{W}\boldsymbol{\gamma}$ is $\mathbf{P_W Y}$.

- If the reduced model is correct, then $\mathbf{P_X Y}$ and $\mathbf{P_W Y}$ are estimates of the same thing, and $\mathbf{P_X Y} - \mathbf{P_W Y} = (\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ should be small.
- If the reduced model is not correct, then $\mathbf{P_X Y}$ and $\mathbf{P_W Y}$ are estimating different things, and $\mathbf{P_X Y} - \mathbf{P_W Y} = (\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ should be large.
- The decision about reduced model adequacy therefore hinges on assessing whether $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ is large or small. Note that $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ is the perpendicular projection of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}$.

MOTIVATION: An obvious measure of the size of $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ is its squared length, that is,

$$\{(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}\}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y} = \mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}.$$

However, the length of $(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ is also related to the sizes of $\mathcal{C}(\mathbf{X})$ and $\mathcal{C}(\mathbf{W})$. We therefore adjust for these sizes by using

$$\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r(\mathbf{P_X} - \mathbf{P_W}).$$

We now compute the expectation of this quantity when the reduced model is/is not correct. For notational simplicity, set $r^* = r(\mathbf{P_X} - \mathbf{P_W})$. When the reduced model is correct, then

$$\begin{aligned} E\{\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r^*\} &= \frac{1}{r^*}\left[(\mathbf{W}\boldsymbol{\gamma})'(\mathbf{P_X} - \mathbf{P_W})\mathbf{W}\boldsymbol{\gamma} + tr\{(\mathbf{P_X} - \mathbf{P_W})\sigma^2\mathbf{I}\}\right] \\ &= \frac{1}{r^*}\{\sigma^2 tr(\mathbf{P_X} - \mathbf{P_W})\} \\ &= \frac{1}{r^*}(r^*\sigma^2) = \sigma^2. \end{aligned}$$

This is correct because $(\mathbf{P_X} - \mathbf{P_W})\mathbf{W}\boldsymbol{\gamma} = \mathbf{0}$ and $tr(\mathbf{P_X} - \mathbf{P_W}) = r(\mathbf{P_X} - \mathbf{P_W}) = r^*$. Thus, if the reduced model is correct, $\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r^*$ is an unbiased estimator of σ^2.

When the reduced model is not correct, then

$$\begin{aligned} E\{\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_W})\mathbf{Y}/r^*\} &= \frac{1}{r^*}\left[(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_W})\mathbf{X}\boldsymbol{\beta}+tr\{(\mathbf{P_X}-\mathbf{P_W})\sigma^2\mathbf{I}\}\right] \\ &= \frac{1}{r^*}\{(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_W})\mathbf{X}\boldsymbol{\beta}+r^*\sigma^2\} \\ &= \sigma^2+(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X}-\mathbf{P_W})\mathbf{X}\boldsymbol{\beta}/r^*. \end{aligned}$$

Thus, if the reduced model is not correct, $\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_W})\mathbf{Y}/r^*$ is estimating something larger than σ^2. Of course, σ^2 is unknown, so it must be estimated. Because the full model is assumed to be correct,

$$\text{MSE} = (n-r)^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y},$$

the MSE from the full model, is an unbiased estimator of σ^2.

TEST STATISTIC: To test the reduced model versus the full model, we use

$$F = \frac{\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_W})\mathbf{Y}/r^*}{\text{MSE}}.$$

Using only our GM model assumptions (i.e., not necessarily assuming normality), we can surmise the following:

- When the reduced model is correct, the numerator and denominator of F are both unbiased estimators of σ^2, so F should be close to 1.
- When the reduced model is not correct, the numerator in F is estimating something larger than σ^2, so F should be larger than 1. Thus, values of F much larger than 1 are not consistent with the reduced model being correct.
- Values of F much smaller than 1 may mean something drastically different; see Christensen (2003).

OBSERVATIONS: In the numerator of F, note that

$$\mathbf{Y}'(\mathbf{P_X}-\mathbf{P_W})\mathbf{Y} = \mathbf{Y}'\mathbf{P_X}\mathbf{Y}-\mathbf{Y}'\mathbf{P_W}\mathbf{Y} = \mathbf{Y}'(\mathbf{P_X}-\mathbf{P_1})\mathbf{Y}-\mathbf{Y}'(\mathbf{P_W}-\mathbf{P_1})\mathbf{Y},$$

which is the difference in the regression (model) sum of squares, corrected or uncorrected, from fitting the two models. Also, the term

$$r^* = r(\mathbf{P_X}-\mathbf{P_W}) = tr(\mathbf{P_X}-\mathbf{P_W}) = tr(\mathbf{P_X})-tr(\mathbf{P_W}) = r(\mathbf{P_X})-r(\mathbf{P_W}) = r-r_0,$$

say, where $r_0 = r(\mathbf{P_W}) = r(\mathbf{W})$. Thus, $r^* = r - r_0$ is the difference in the ranks of the $\mathbf{X}$ and $\mathbf{W}$ matrices. This also equals the difference in the model degrees of freedom from the two ANOVA tables.

REMARK: You will note that we have formulated a perfectly sensible strategy for testing reduced versus full models while avoiding the question, "What is the distribution of F?" Our entire argument is based on first and second moment assumptions, that is, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, the GM assumptions. We now address the distributional question.

DISTRIBUTION OF F: To derive the sampling distribution of

$$F = \frac{\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r^*}{\text{MSE}},$$

we require that $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, from which it follows that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. First, we handle the denominator $\text{MSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/(n-r)$. In Example 5.3 (notes), we showed that

$$\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2 \sim \chi^2_{n-r}.$$

This distributional result holds regardless of whether or not the reduced model is correct. Now, we turn our attention to the numerator. Take $\mathbf{A} = \sigma^{-2}(\mathbf{P_X} - \mathbf{P_W})$ and consider the quadratic form

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/\sigma^2.$$

With $\mathbf{V} = \sigma^2\mathbf{I}$, the matrix

$$\mathbf{AV} = \sigma^{-2}(\mathbf{P_X} - \mathbf{P_W})\sigma^2\mathbf{I} = \mathbf{P_X} - \mathbf{P_W}$$

is idempotent with rank $r(\mathbf{P_X} - \mathbf{P_W}) = r^*$. Therefore, we know that $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi^2_{r^*}(\lambda)$, where

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{\beta}.$$

Now, we make the following observations:

- If the reduced model is correct and $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{W})$, then $(\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ because $\mathbf{P_X} - \mathbf{P_W}$ projects onto $\mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}$. This means that the noncentrality parameter $\lambda = 0$ and $\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/\sigma^2 \sim \chi^2_{r^*}$, a central χ^2 distribution.

- If the reduced model is not correct and $\mathbf{X}\boldsymbol{\beta} \notin \mathcal{C}(\mathbf{W})$, then $(\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{\beta} \neq \mathbf{0}$ and $\lambda > 0$. In this event, $\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/\sigma^2 \sim \chi^2_{r^*}(\lambda)$, a noncentral χ^2 distribution with noncentrality parameter λ.

- Regardless of whether or not the reduced model is correct, the quadratic forms $\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}$ and $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$ are independent since
$$(\mathbf{P_X} - \mathbf{P_W})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{P_X}) = \sigma^2(\mathbf{P_X} - \mathbf{P_W})(\mathbf{I} - \mathbf{P_X}) = \mathbf{0}.$$

CONCLUSION: Putting this all together, we have that
$$F = \frac{\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r^*}{\text{MSE}} = \frac{\sigma^{-2}\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r^*}{\sigma^{-2}\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/(n-r)} \sim F_{r^*,n-r}(\lambda),$$
where $r^* = r - r_0$ and
$$\lambda = \frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{\beta}.$$
If the reduced model is correct, that is, if $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{W})$, then $\lambda = 0$ and $F \sim F_{r^*,n-r}$, a central F distribution. Note also that if the reduced model is correct,
$$E(F) = \frac{n-r}{n-r-2} \approx 1.$$
This reaffirms our (model free) assertion that values of F close to 1 are consistent with the reduced model being correct. Because the noncentral F family is stochastically increasing in λ, larger values of F are consistent with the reduced model not being correct.

SUMMARY: Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, Suppose that we would like to test
$$H_0 : \mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$
versus
$$H_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$
where $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$. An α level rejection region is
$$\text{RR} = \{F : F > F_{r^*,n-r,\alpha}\},$$
where $r^* = r - r_0$, $r_0 = r(\mathbf{W})$, and $F_{r^*,n-r,\alpha}$ is the upper α quantile of the $F_{r^*,n-r}$ distribution.

Example 6.1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. In matrix notation,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 - \overline{x} \\ 1 & x_2 - \overline{x} \\ \vdots & \vdots \\ 1 & x_n - \overline{x} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that we would like to test whether the reduced model

$$Y_i = \beta_0 + \epsilon_i,$$

for $i = 1, 2, ..., n$, also holds. In matrix notation, the reduced model can be expressed as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{1}, \quad \gamma = \beta_0, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{1}$ is an $n \times 1$ vector of ones. Note that $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$ with $r_0 = 1$, $r = 2$, and $r^* = r - r_0 = 1$. When the reduced model is correct,

$$F = \frac{\mathbf{Y}'(\mathbf{P_X} - \mathbf{P_W})\mathbf{Y}/r^*}{\text{MSE}} \sim F_{1,n-2},$$

where MSE is the mean-squared error from the full model. When the reduced model is not correct, $F \sim F_{1,n-2}(\lambda)$, where

$$\begin{aligned} \lambda &= \frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{\beta} \\ &= \beta_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2/2\sigma^2. \end{aligned}$$

Exercises: (a) Verify that this expression for the noncentrality parameter λ is correct. (b) Suppose that n is even and the values of x_i can be selected anywhere in the interval (d_1, d_2). How should we choose the x_i values to maximize the power of a level α test?

6.3 Testing linear parametric functions

PROBLEM: Consider our usual Gauss-Markov linear model with normal errors; i.e., $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We now consider the problem of testing

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$$

versus

$$H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{m},$$

where $\mathbf{K}$ is a $p \times s$ matrix with $r(\mathbf{K}) = s$ and $\mathbf{m}$ is $s \times 1$.

Example 6.2. Consider the regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$, for $i = 1, 2, ..., n$. Express each hypothesis in the form $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$:

1. $H_0 : \beta_1 = 0$

2. $H_0 : \beta_3 = \beta_4 = 0$

3. $H_0 : \beta_1 + \beta_3 = 1$, $\beta_2 - \beta_4 = -1$

4. $H_0 : \beta_2 = \beta_3 = \beta_4$.

Example 6.3. Consider the analysis of variance model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, for $i = 1, 2, 3, 4$ and $j = 1, 2, ..., n_i$. Express each hypothesis in the form $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$:

1. $H_0 : \mu + \alpha_1 = 5$, $\alpha_3 - \alpha_4 = 1$

2. $H_0 : \alpha_1 - \alpha_2 = \alpha_3 - \alpha_4$

3. $H_0 : \alpha_1 - 2 = \frac{1}{3}(\alpha_2 + \alpha_3 + \alpha_4)$.

TERMINOLOGY: The general linear hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is said to be **testable** iff $\mathbf{K}$ has full column rank and each component of $\mathbf{K}'\boldsymbol{\beta}$ is estimable. In other words, $\mathbf{K}'\boldsymbol{\beta}$ contains s linearly independent estimable functions. Otherwise, $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is said to be **nontestable**.

GOAL: Our goal is to develop a test for $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, $\mathbf{K}'\boldsymbol{\beta}$ estimable, in the Gauss-Markov model with normal errors. We start by noting that the BLUE of $\mathbf{K}'\boldsymbol{\beta}$ is $\mathbf{K}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is a least squares estimator of $\boldsymbol{\beta}$. Also,

$$\mathbf{K}'\widehat{\boldsymbol{\beta}} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y},$$

a linear function of $\mathbf{Y}$, so $\mathbf{K}'\widehat{\boldsymbol{\beta}}$ follows an s-variate normal distribution with mean $E(\mathbf{K}'\widehat{\boldsymbol{\beta}}) = \mathbf{K}'\boldsymbol{\beta}$ and covariance matrix

$$\text{cov}(\mathbf{K}'\widehat{\boldsymbol{\beta}}) = \mathbf{K}'\text{cov}(\widehat{\boldsymbol{\beta}})\mathbf{K} = \sigma^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{K} = \sigma^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K} = \sigma^2\mathbf{H},$$

where $\mathbf{H} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}$. That is, we have shown $\mathbf{K}'\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_s(\mathbf{K}'\boldsymbol{\beta}, \sigma^2\mathbf{H})$.

NOTE: In the calculation above, note that $\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{K} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}$ only because $\mathbf{K}'\boldsymbol{\beta}$ is estimable; i.e., $\mathbf{K}' = \mathbf{A}'\mathbf{X}$ for some $\mathbf{A}$. It is also true that $\mathbf{H}$ is nonsingular.

LEMMA: If $\mathbf{K}'\boldsymbol{\beta}$ is estimable, then $\mathbf{H}$ is nonsingular.
Proof. First note that $\mathbf{H}$ is an $s \times s$ matrix. We can write

$$\mathbf{H} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K} = \mathbf{A}'\mathbf{P_X}\mathbf{A} = \mathbf{A}'\mathbf{P_X}\mathbf{P_X}\mathbf{A} = \mathbf{A}'\mathbf{P}'_\mathbf{X}\mathbf{P_X}\mathbf{A},$$

since $\mathbf{K}' = \mathbf{A}'\mathbf{X}$ for some $\mathbf{A}$ (this follows since $\mathbf{K}'\boldsymbol{\beta}$ is estimable). Therefore,

$$r(\mathbf{H}) = r(\mathbf{A}'\mathbf{P}'_\mathbf{X}\mathbf{P_X}\mathbf{A}) = r(\mathbf{P_X}\mathbf{A})$$

and

$$s = r(\mathbf{K}) = r(\mathbf{X}'\mathbf{A}) = r(\mathbf{X}'\mathbf{P_X}\mathbf{A}) \leq r(\mathbf{P_X}\mathbf{A}) = r[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{A}] \leq r(\mathbf{X}'\mathbf{A}) = s.$$

Therefore, $r(\mathbf{H}) = r(\mathbf{P_X}\mathbf{A}) = s$, showing that $\mathbf{H}$ is nonsingular. □

IMPLICATION: The lemma above is important, because it convinces us that the distribution of $\mathbf{K}'\widehat{\boldsymbol{\beta}}$ is full rank. Subtracting $\mathbf{m}$, we have

$$\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m} \sim \mathcal{N}_s(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}, \sigma^2\mathbf{H}).$$

F STATISTIC: Now, consider the quadratic form

$$(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'(\sigma^2\mathbf{H})^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m}).$$

Note that $(\sigma^2\mathbf{H})^{-1}\sigma^2\mathbf{H} = \mathbf{I}_s$, an idempotent matrix with rank s. Therefore,

$$(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'(\sigma^2\mathbf{H})^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m}) \sim \chi^2_s(\lambda),$$

where the noncentrality parameter

$$\lambda = \frac{1}{2\sigma^2}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}).$$

We have already shown that $\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2 \sim \chi^2_{n-r}$. Also,

$$(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'(\sigma^2\mathbf{H})^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m}) \quad \text{and} \quad \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2$$

are independent (verify!). Thus, the ratio

$$\begin{aligned} F &= \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})/s}{\text{MSE}} \\ &= \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})/s}{\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/(n-r)} \\ &= \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'(\sigma^2\mathbf{H})^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})/s}{\sigma^{-2}\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/(n-r)} \sim F_{s,n-r}(\lambda), \end{aligned}$$

where $\lambda = (\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})/2\sigma^2$. Note that if $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is true, the noncentrality parameter $\lambda = 0$ and $F \sim F_{s,n-r}$. Therefore, an α level rejection region for the test of $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ versus $H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{m}$ is

$$\text{RR} = \{F : F > F_{s,n-r,\alpha}\}.$$

SCALAR CASE: We now consider the special case of testing $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ when $r(\mathbf{K}) = 1$, that is, $\mathbf{K}'\boldsymbol{\beta}$ is a scalar estimable function. This function (and hypothesis) is perhaps more appropriately written as $H_0 : \mathbf{k}'\boldsymbol{\beta} = m$ to emphasize that $\mathbf{k}$ is a $p \times 1$ vector and m is a scalar. Often, $\mathbf{k}$ is chosen in a way so that $m = 0$ (e.g., testing a contrast in an ANOVA model, etc.). The hypotheses in Example 6.2 (#1) and Example 6.3 (#3) are of this form. Testing a scalar hypothesis is a mere special case of the general test we have just derived. However, additional flexibility results in the scalar case; in particular, we can test for one sided alternatives like $H_1 : \mathbf{k}'\boldsymbol{\beta} > m$ or $H_1 : \mathbf{k}'\boldsymbol{\beta} < m$. We first discuss one more noncentral distribution.

TERNINOLOGY: Suppose $Z \sim \mathcal{N}(\mu, 1)$ and $V \sim \chi^2_k$. If Z and V are independent, then

$$T = \frac{Z}{\sqrt{V/k}}$$

follows a **noncentral t distribution** with k degrees of freedom and noncentrality parameter μ. We write $T \sim t_k(\mu)$. If $\mu = 0$, the $t_k(\mu)$ distribution reduces to a central t distribution with k degrees of freedom. Note that $T \sim t_k(\mu)$ implies $T^2 \sim F_{1,k}(\mu^2/2)$.

TEST PROCEDURE: Consider the Gauss-Markov linear model with normal errors; i.e., $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that $\mathbf{k}'\boldsymbol{\beta}$ is estimable; i.e., $\mathbf{k}' \in \mathcal{R}(\mathbf{X})$, and that our goal is to test

$$H_0 : \mathbf{k}'\boldsymbol{\beta} = m$$

$$\text{versus}$$

$$H_1 : \mathbf{k}'\boldsymbol{\beta} \neq m$$

or versus a suitable one-sided alternative H_1. The BLUE for $\mathbf{k}'\boldsymbol{\beta}$ is $\mathbf{k}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is a least squares estimator. Straightforward calculations show that

$$\mathbf{k}'\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\{\mathbf{k}'\boldsymbol{\beta}, \sigma^2\mathbf{k}'(\mathbf{X}'\mathbf{X})^-\mathbf{k}\} \quad \Longrightarrow \quad Z = \frac{\mathbf{k}'\widehat{\boldsymbol{\beta}} - \mathbf{k}'\boldsymbol{\beta}}{\sqrt{\sigma^2\mathbf{k}'(\mathbf{X}'\mathbf{X})^-\mathbf{k}}} \sim \mathcal{N}(0, 1).$$

We know that $V = \mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/\sigma^2 \sim \chi^2_{n-r}$ and that Z and V are independent (verify!). Thus,

$$T = \frac{(\mathbf{k}'\widehat{\boldsymbol{\beta}} - \mathbf{k}'\boldsymbol{\beta})/\sqrt{\sigma^2\mathbf{k}'(\mathbf{X}'\mathbf{X})^-\mathbf{k}}}{\sqrt{\sigma^{-2}\mathbf{Y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Y}/(n-r)}} = \frac{\mathbf{k}'\widehat{\boldsymbol{\beta}} - \mathbf{k}'\boldsymbol{\beta}}{\sqrt{\text{MSE }\mathbf{k}'(\mathbf{X}'\mathbf{X})^-\mathbf{k}}} \sim t_{n-r}.$$

When $H_0 : \mathbf{k}'\boldsymbol{\beta} = m$ is true, the statistic

$$T = \frac{\mathbf{k}'\widehat{\boldsymbol{\beta}} - m}{\sqrt{\text{MSE }\mathbf{k}'(\mathbf{X}'\mathbf{X})^-\mathbf{k}}} \sim t_{n-r}.$$

Therefore, an α level rejection region, when H_1 is two sided, is

$$\text{RR} = \{T : T \geq t_{n-r,\alpha/2}\}.$$

One sided tests use rejection regions that are suitably adjusted. When H_0 is not true, $T \sim t_{n-r}(\mu)$, where

$$\mu = \frac{\mathbf{k}'\boldsymbol{\beta} - m}{\sqrt{\sigma^2\mathbf{k}'(\mathbf{X}'\mathbf{X})^-\mathbf{k}}}.$$

This distribution is of interest for power and sample size calculations.

6.4 Testing models versus testing linear parametric functions

SUMMARY: Under our Gauss Markov linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, we have presented F statistics to test (a) a reduced model versus a full model in Section 6.2 and (b) a hypothesis defined by $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, for $\mathbf{K}'\boldsymbol{\beta}$ estimable, in Section 6.3. In fact, testing models and testing linear parametric functions essentially is the same thing, as we now demonstrate. For simplicity, we take $\mathbf{m} = \mathbf{0}$, although the following argument can be generalized.

DISCUSSION: Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\beta} \in \mathcal{R}^p$. Consider the testable hypothesis $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$. We can write this hypothesis in the following way:

$$H_0 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \ \text{ and } \ \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}.$$

We now find a reduced model that corresponds to this hypothesis. Note that $\mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$ holds if and only if $\boldsymbol{\beta} \perp \mathcal{C}(\mathbf{K})$. To identify the reduced model, pick a matrix $\mathbf{U}$ such that $\mathcal{C}(\mathbf{U}) = \mathcal{C}(\mathbf{K})^{\perp}$. We then have

$$\mathbf{K}'\boldsymbol{\beta} = \mathbf{0} \Longleftrightarrow \boldsymbol{\beta} \perp \mathcal{C}(\mathbf{K}) \Longleftrightarrow \boldsymbol{\beta} \in \mathcal{C}(\mathbf{U}) \Longleftrightarrow \boldsymbol{\beta} = \mathbf{U}\boldsymbol{\gamma},$$

for some vector $\boldsymbol{\gamma}$. Substituting $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\gamma}$ into the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ gives the reduced model $\mathbf{Y} = \mathbf{X}\mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, or letting $\mathbf{W} = \mathbf{X}\mathbf{U}$, our hypothesis above can be written

$$H_0 : \mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\mathcal{C}(\mathbf{W}) \subset \mathcal{C}(\mathbf{X})$.

OBSERVATION: When $\mathbf{K}'\boldsymbol{\beta}$ is estimable, that is, when $\mathbf{K}' = \mathbf{D}'\mathbf{X}$ for some $n \times s$ matrix $\mathbf{D}$, we can find the perpendicular projection matrix for testing $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$ in terms of $\mathbf{D}$ and $\mathbf{P}_{\mathbf{X}}$. From Section 6.2, recall that the numerator sum of squares to test the reduced model $\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ versus the full model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is $\mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{W}})\mathbf{Y}$, where $\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{W}}$ is the ppm onto $\mathcal{C}(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{W}}) = \mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}$. For testing the estimable function $\mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$, we now show that the ppm onto $\mathcal{C}(\mathbf{P}_{\mathbf{X}}\mathbf{D})$ is also the ppm onto $\mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}$; i.e., that $\mathcal{C}(\mathbf{P}_{\mathbf{X}}\mathbf{D}) = \mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}$.

PROPOSITION: $\mathcal{C}(\mathbf{P_X}-\mathbf{P_W})=\mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}=\mathcal{C}(\mathbf{XU})^{\perp}_{\mathcal{C}(\mathbf{X})}=\mathcal{C}(\mathbf{P_X D})$.

Proof. We showed $\mathcal{C}(\mathbf{P_X}-\mathbf{P_W})=\mathcal{C}(\mathbf{W})^{\perp}_{\mathcal{C}(\mathbf{X})}$ in Chapter 2 and $\mathbf{W}=\mathbf{XU}$, so the second equality is obvious. Suppose that $\mathbf{v}\in\mathcal{C}(\mathbf{XU})^{\perp}_{\mathcal{C}(\mathbf{X})}$. Then, $\mathbf{v}'\mathbf{XU}=\mathbf{0}$, so that $\mathbf{X}'\mathbf{v}\perp\mathcal{C}(\mathbf{U})$. Because $\mathcal{C}(\mathbf{U})=\mathcal{C}(\mathbf{K})^{\perp}$, we know that $\mathbf{X}'\mathbf{v}\in\mathcal{C}(\mathbf{K})=\mathcal{C}(\mathbf{X}'\mathbf{D})$, since $\mathbf{K}'=\mathbf{D}'\mathbf{X}$. Thus,

$$\mathbf{v}=\mathbf{P_X v}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{v}\in\mathcal{C}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{D}]=\mathcal{C}(\mathbf{P_X D}).$$

Suppose that $\mathbf{v}\in\mathcal{C}(\mathbf{P_X D})$. Clearly, $\mathbf{v}\in\mathcal{C}(\mathbf{X})$. Also, $\mathbf{v}=\mathbf{P_X D d}$, for some $\mathbf{d}$ and

$$\mathbf{v}'\mathbf{XU}=\mathbf{d}'\mathbf{D}'\mathbf{P_X XU}=\mathbf{d}'\mathbf{D}'\mathbf{XU}=\mathbf{d}'\mathbf{K}'\mathbf{U}=\mathbf{0},$$

because $\mathcal{C}(\mathbf{U})=\mathcal{C}(\mathbf{K})^{\perp}$. Thus, $\mathbf{v}\in\mathcal{C}(\mathbf{XU})^{\perp}_{\mathcal{C}(\mathbf{X})}$.

IMPLICATION: It follows immediately that the numerator sum of squares for testing the reduced model $\mathbf{Y}=\mathbf{W}\boldsymbol{\gamma}+\boldsymbol{\epsilon}$ versus the full model $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ is $\mathbf{Y}'\mathbf{M_{P_X D}}\mathbf{Y}$, where

$$\mathbf{M_{P_X D}}=\mathbf{P_X D}[(\mathbf{P_X D})'(\mathbf{P_X D})]^{-}(\mathbf{P_X D})'=\mathbf{P_X D}(\mathbf{D}'\mathbf{P_X D})^{-}\mathbf{D}'\mathbf{P_X}$$

is the ppm onto $\mathcal{C}(\mathbf{P_X D})$. If $\boldsymbol{\epsilon}\sim\mathcal{N}_n(\mathbf{0},\sigma^2\mathbf{I})$, the resulting test statistic

$$F=\frac{\mathbf{Y}'\mathbf{M_{P_X D}}\mathbf{Y}/r(\mathbf{M_{P_X D}})}{\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}/r(\mathbf{I}-\mathbf{P_X})}\sim F_{r(\mathbf{M_{P_X D}}),r(\mathbf{I}-\mathbf{P_X})}(\lambda),$$

where the noncentrality parameter

$$\lambda=\frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})'\mathbf{M_{P_X D}}\mathbf{X}\boldsymbol{\beta}.$$

GOAL: Our goal now is to show that the F statistic above is the same F statistic we derived in Section 6.3 with $\mathbf{m}=\mathbf{0}$, that is,

$$F=\frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}})'\mathbf{H}^{-1}\mathbf{K}'\widehat{\boldsymbol{\beta}}/s}{\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}/(n-r)}.$$

Recall that this statistic was derived for the testable hypothesis $H_0:\mathbf{K}'\boldsymbol{\beta}=\mathbf{0}$. First, we show that $r(\mathbf{M_{P_X D}})=s$, where, recall, $s=r(\mathbf{K})$. To do this, it suffices to show that $r(\mathbf{K})=r(\mathbf{P_X D})$. Because $\mathbf{K}'\boldsymbol{\beta}$ is estimable, we know that $\mathbf{K}'=\mathbf{D}'\mathbf{X}$, for some $\mathbf{D}$. Writing $\mathbf{K}=\mathbf{X}'\mathbf{D}$, we see that for any vector $\mathbf{a}$,

$$\mathbf{X}'\mathbf{Da}=\mathbf{0}\Longleftrightarrow\mathbf{Da}\perp\mathcal{C}(\mathbf{X}),$$

which occurs iff $\mathbf{P_X Da} = \mathbf{0}$. Note that the $\mathbf{X'Da} = \mathbf{0} \iff \mathbf{P_X Da} = \mathbf{0}$ equivalence implies that

$$\mathcal{N}(\mathbf{X'D}) = \mathcal{N}(\mathbf{P_X D}) \iff \mathcal{C}(\mathbf{D'X})^{\perp} = \mathcal{C}(\mathbf{D'P_X})^{\perp} \iff \mathcal{C}(\mathbf{D'X}) = \mathcal{C}(\mathbf{D'P_X})$$

so that $r(\mathbf{D'X}) = r(\mathbf{D'P_X})$. But $\mathbf{K}' = \mathbf{D'X}$, so $r(\mathbf{K}) = r(\mathbf{D'X}) = r(\mathbf{D'P_X}) = r(\mathbf{P_X D})$, which is what we set out to prove. Now, consider the quadratic form $\mathbf{Y'M_{P_X D}Y}$, and let $\widehat{\boldsymbol{\beta}}$ denote a least squares estimator of $\boldsymbol{\beta}$. Result 2.5 says that $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P_X Y}$, so that $\mathbf{K}'\widehat{\boldsymbol{\beta}} = \mathbf{D'X}\widehat{\boldsymbol{\beta}} = \mathbf{D'P_X Y}$. Substitution gives

$$\begin{aligned}\mathbf{Y'M_{P_X D}Y} &= \mathbf{Y'P_X D(D'P_X D)^- D'P_X Y} \\ &= (\mathbf{K}'\widehat{\boldsymbol{\beta}})'[\mathbf{D'X(X'X)^- X'D}]^- \mathbf{K}'\widehat{\boldsymbol{\beta}} \\ &= (\mathbf{K}'\widehat{\boldsymbol{\beta}})'[\mathbf{K'(X'X)^- K}]^- \mathbf{K}'\widehat{\boldsymbol{\beta}}.\end{aligned}$$

Recalling that $\mathbf{H} = \mathbf{K'(X'X)^- K}$ and that $\mathbf{H}$ is nonsingular (when $\mathbf{K}'\boldsymbol{\beta}$ is estimable) should convince you that the numerator sum of squares in

$$F = \frac{\mathbf{Y'M_{P_X D}Y}/r(\mathbf{M_{P_X D}})}{\mathbf{Y'(I - P_X)Y}/r(\mathbf{I - P_X})}$$

and

$$F = \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}})'\mathbf{H}^{-1}\mathbf{K}'\widehat{\boldsymbol{\beta}}/s}{\mathbf{Y'(I - P_X)Y}/(n - r)}$$

are equal. We already showed that $r(\mathbf{M_{P_X D}}) = s$, and because $r(\mathbf{I - P_X}) = n - r$, we are done. □

6.5 Likelihood ratio tests

6.5.1 Constrained estimation

REMARK: In the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ (note the minimal assumptions), we have, up until now, allowed the $p \times 1$ parameter vector $\boldsymbol{\beta}$ to take on any value in $\mathcal{R}^p$, that is, we have made no restrictions on the parameters in $\boldsymbol{\beta}$. We now consider the case where $\boldsymbol{\beta}$ is restricted to the subspace of $\mathcal{R}^p$ consisting of values of $\boldsymbol{\beta}$ that satisfy

$\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$, where $\mathbf{P}$ is a $p \times q$ matrix and $\boldsymbol{\delta}$ is $q \times 1$. To avoid technical difficulties, we will assume that the system $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$ is consistent; i.e., $\boldsymbol{\delta} \in \mathcal{C}(\mathbf{P}')$. Otherwise, the set $\{\boldsymbol{\beta} \in \mathcal{R}^p : \mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}\}$ could be empty.

PROBLEM: In the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, we would like to minimize

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

subject to the constraint that $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$. Essentially, this requires us to find the minimum value of $Q(\boldsymbol{\beta})$ over the linear subspace $\{\boldsymbol{\beta} \in \mathcal{R}^p : \mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}\}$. This is a restricted minimization problem and standard Lagrangian methods apply; see Appendix B in Monahan. The Lagrangian $a(\boldsymbol{\beta}, \boldsymbol{\theta})$ is a function of $\boldsymbol{\beta}$ and the Lagrange multipliers in $\boldsymbol{\theta}$ and can be written as

$$a(\boldsymbol{\beta}, \boldsymbol{\theta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\theta}'(\mathbf{P}'\boldsymbol{\beta} - \boldsymbol{\delta}).$$

Taking partial derivatives, we have

$$\begin{aligned}\frac{\partial a(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\mathbf{P}\boldsymbol{\theta} \\ \frac{\partial a(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= 2(\mathbf{P}'\boldsymbol{\beta} - \boldsymbol{\delta}).\end{aligned}$$

Setting these equal to zero leads to the **restricted normal equations** (RNEs), that is,

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{P} \\ \mathbf{P}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \boldsymbol{\delta} \end{pmatrix}.$$

Denote by $\widehat{\boldsymbol{\beta}}_H$ and $\widehat{\boldsymbol{\theta}}_H$ the solutions to the RNEs, respectively. The solution $\widehat{\boldsymbol{\beta}}_H$ is called a **restricted least squares estimator**.

DISCUSSION: We now present some facts regarding this restricted linear model and its (restricted) least squares estimator. We have proven all of these facts for the unrestricted model; restricted versions of the proofs are all in Monahan.

1. The restricted normal equations are consistent; see Result 3.8, Monahan (pp 62-63).
2. A solution $\widehat{\boldsymbol{\beta}}_H$ minimizes $Q(\boldsymbol{\beta})$ over the set $\mathcal{T} \equiv \{\boldsymbol{\beta} \in \mathcal{R}^p : \mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}\}$; see Result 3.9, Monahan (pp 63).

3. The function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the restricted model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta},$$

   if and only if $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X} + \mathbf{d}'\mathbf{P}'$, for some $\mathbf{a}$ and $\mathbf{d}$, that is,

$$\boldsymbol{\lambda}' \in \mathcal{R}\begin{pmatrix} \mathbf{X} \\ \mathbf{P}' \end{pmatrix}.$$

   See Result 3.7, Monahan (pp 60).

4. If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the unrestricted model; i.e., the model without the linear restriction, then $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the restricted model. The converse is not true.

5. Under the GM model assumptions, if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the restricted model, then $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}_H$ is the BLUE of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ in the restricted model. See Result 4.5, Monahan (pp 89-90).

6.5.2 Testing procedure

SETTING: Consider the Gauss Markov linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We are interested in deriving the likelihood ratio test (LRT) for

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$$

$$\text{versus}$$

$$H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{m},$$

where $\mathbf{K}$ is a $p \times s$ matrix with $r(\mathbf{K}) = s$ and $\mathbf{m}$ is $s \times 1$. We assume that $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is testable, that is, $\mathbf{K}'\boldsymbol{\beta}$ is estimable.

RECALL: A likelihood ratio testing procedure is intuitive. One simply compares the maximized likelihood over the restricted parameter space (that is, the space under H_0) to the maximized likelihood over the entire parameter space. If the former is small when compared to the latter, then there a large amount of evidence against H_0.

DERIVATION: Under our model assumptions, we know that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. The likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ is

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-n/2}\exp\{-Q(\boldsymbol{\beta})/2\sigma^2\},$$

where $Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The unrestricted parameter space is

$$\Theta = \{\boldsymbol{\theta} : \boldsymbol{\beta} \in \mathcal{R}^p,\ \sigma^2 \in \mathcal{R}^+\}.$$

The restricted parameter space, that is, the parameter space under $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, is

$$\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\beta} \in \mathcal{R}^p,\ \mathbf{K}'\boldsymbol{\beta} = \mathbf{m},\ \sigma^2 \in \mathcal{R}^+\}.$$

The likelihood ratio statistic is

$$\lambda \equiv \lambda(\mathbf{Y}) = \frac{\sup_{\Theta_0} L(\boldsymbol{\theta}|\mathbf{Y})}{\sup_{\Theta} L(\boldsymbol{\theta}|\mathbf{Y})}.$$

We reject the null hypothesis H_0 for small values of $\lambda = \lambda(\mathbf{Y})$. Thus, to perform a level α test, reject H_0 when $\lambda < c$, where $c \in (0, 1)$ is chosen to satisfy $P_{H_0}\{\lambda(\mathbf{Y}) \leq c\} = \alpha$. We have seen (Section 6.1) that the unrestricted MLEs of $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \quad \text{and} \quad \widehat{\sigma}^2 = \frac{Q(\widehat{\boldsymbol{\beta}})}{n}.$$

Similarly, maximizing $L(\boldsymbol{\theta}|\mathbf{y})$ over Θ_0 produces the solutions $\widehat{\boldsymbol{\beta}}_H$ and $\widetilde{\sigma}^2 = Q(\widehat{\boldsymbol{\beta}}_H)/n$, where $\widehat{\boldsymbol{\beta}}_H$ is any solution to

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{K} \\ \mathbf{K}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{m} \end{pmatrix},$$

the restricted normal equations. Algebra shows that

$$\lambda = \frac{L(\widehat{\boldsymbol{\beta}}_H, \widetilde{\sigma}^2|\mathbf{Y})}{L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2|\mathbf{Y})} = \left(\frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right)^{n/2} = \left\{\frac{Q(\widehat{\boldsymbol{\beta}})}{Q(\widehat{\boldsymbol{\beta}}_H)}\right\}^{n/2}.$$

More algebra shows that

$$\left\{\frac{Q(\widehat{\boldsymbol{\beta}})}{Q(\widehat{\boldsymbol{\beta}}_H)}\right\}^{n/2} < c \iff \frac{\{Q(\widehat{\boldsymbol{\beta}}_H) - Q(\widehat{\boldsymbol{\beta}})\}/s}{Q(\widehat{\boldsymbol{\beta}})/(n-r)} > c^*,$$

where $s = r(\mathbf{K})$ and $c^* = s^{-1}(n-r)(c^{-2/n}-1)$. Furthermore, Monahan's Theorem 6.1 (pp 139-140) shows that when $\mathbf{K}'\boldsymbol{\beta}$ is estimable,

$$Q(\widehat{\boldsymbol{\beta}}_H) - Q(\widehat{\boldsymbol{\beta}}) = (\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m}),$$

where $\mathbf{H} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}$. Applying this result, and noting that $Q(\widehat{\boldsymbol{\beta}})/(n-r) = \text{MSE}$, we see that

$$\frac{\{Q(\widehat{\boldsymbol{\beta}}_H) - Q(\widehat{\boldsymbol{\beta}})\}/s}{Q(\widehat{\boldsymbol{\beta}})/(n-r)} > c^* \iff F = \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})/s}{\text{MSE}} > c^*.$$

That is, the LRT specifies that we reject H_0 when F is large. Choosing $c^* = F_{s,n-r,\alpha}$ provides a level α test. Therefore, under the Gauss Markov model with normal errors, the LRT for $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ is the same test as that in Section 6.3.

6.6 Confidence intervals

6.6.1 Single intervals

PROBLEM: Consider the Gauss Markov linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ estimable, that is, $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$, for some vector $\mathbf{a}$. Our goal is to write a $100(1-\alpha)$ percent confidence interval for $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

DERIVATION: We start with the obvious point estimator $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$, the least squares estimator (and MLE) of $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Under our model assumptions, we know that

$$\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\{\boldsymbol{\lambda}'\boldsymbol{\beta},\ \sigma^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}\}$$

and, hence,

$$Z = \frac{\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}} \sim \mathcal{N}(0,1).$$

If σ^2 was known, our work would be done as Z is a pivot. More likely, this is not the case, so we must estimate it. An obvious point estimator for σ^2 is MSE, where

$$\text{MSE} = (n-r)^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}.$$

We consider the quantity

$$T = \frac{\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\boldsymbol{\beta}}{\sqrt{\text{MSE } \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}}$$

and subsequently show that $T \sim t_{n-r}$. Note that

$$\begin{aligned} T &= \frac{\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\boldsymbol{\beta}}{\sqrt{\text{MSE } \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}} \\ &= \frac{(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\boldsymbol{\beta})/\sqrt{\sigma^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}}{\sqrt{(n-r)^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}/\sigma^2}} \sim \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\sqrt{\text{“}\chi^2_{n-r}\text{”}/(n-r)}}. \end{aligned}$$

To verify that $T \sim t_{n-r}$, it remains only to show that Z and $\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}/\sigma^2$ are independent, or equivalently, that $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ and $\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}$ are, since $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ is a function of Z and since σ^2 is not random. Note that

$$\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} = \mathbf{a}'\mathbf{P_X}\mathbf{Y},$$

a linear function of $\mathbf{Y}$. Using Result 5.19, $\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}}$ and $\mathbf{Y}'(\mathbf{I}-\mathbf{P_X})\mathbf{Y}$ are independent since $\mathbf{a}'\mathbf{P_X}\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{P_X}) = \mathbf{0}$. Thus, $T \sim t_{n-r}$, i.e., t is a pivot, so that

$$\text{pr}\left(-t_{n-r,\alpha/2} < \frac{\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\boldsymbol{\beta}}{\sqrt{\text{MSE } \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}} < t_{n-r,\alpha/2}\right) = 1-\alpha.$$

Algebra shows that this probability statement is the same as

$$\text{pr}\left(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} - t_{n-r,\alpha/2}\sqrt{\text{MSE } \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}} < \boldsymbol{\lambda}'\boldsymbol{\beta} < \boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} + t_{n-r,\alpha/2}\sqrt{\text{MSE } \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}\right) = 1-\alpha,$$

showing that

$$\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} \pm t_{n-r,\alpha/2}\sqrt{\text{MSE } \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}$$

is a $100(1-\alpha)$ percent confidence interval for $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

Example 6.4. Recall the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are iid $\mathcal{N}(0, \sigma^2)$. Recall also that the least squares estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} - \widehat{\beta}_1\overline{x} \\ \frac{\sum_i (x_i-\overline{x})(Y_i-\overline{Y})}{\sum_i (x_i-\overline{x})^2} \end{pmatrix},$$

and that the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\text{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2} & -\frac{\overline{x}}{\sum_i (x_i - \overline{x})^2} \\ -\frac{\overline{x}}{\sum_i (x_i - \overline{x})^2} & \frac{1}{\sum_i (x_i - \overline{x})^2} \end{pmatrix}.$$

We now consider the problem of writing a $100(1-\alpha)$ percent confidence interval for

$$E(Y|x = x_0) = \beta_0 + \beta_1 x_0,$$

the mean response of Y when $x = x_0$. Note that $E(Y|x = x_0) = \beta_0 + \beta_1 x_0 = \boldsymbol{\lambda}'\boldsymbol{\beta}$, where $\boldsymbol{\lambda}' = (1 \;\; x_0)$. Also, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable because this is a regression model so our previous work applies. The least squares estimator (and MLE) of $E(Y|x = x_0)$ is

$$\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Straightforward algebra (verify!) shows that

$$\begin{aligned} \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda} &= \begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} \frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2} & -\frac{\overline{x}}{\sum_i (x_i - \overline{x})^2} \\ -\frac{\overline{x}}{\sum_i (x_i - \overline{x})^2} & \frac{1}{\sum_i (x_i - \overline{x})^2} \end{pmatrix} \begin{pmatrix} 1 \\ x_0 \end{pmatrix} \\ &= \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}. \end{aligned}$$

Thus, a $100(1-\alpha)$ percent confidence interval for $\boldsymbol{\lambda}'\boldsymbol{\beta} = E(Y|x = x_0)$ is

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \pm t_{n-2,\alpha/2}\sqrt{\text{MSE}\,\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right\}}. \;\; \square$$

6.6.2 Multiple intervals

PROBLEM: Consider the Gauss Markov linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. We now consider the problem of writing **simultaneous** confidence intervals for the k estimable functions $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta}$. Let the $p \times k$ matrix $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1 \; \boldsymbol{\lambda}_2 \; \cdots \; \boldsymbol{\lambda}_k)$ so that

$$\boldsymbol{\tau} = \boldsymbol{\Lambda}'\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\lambda}_1'\boldsymbol{\beta} \\ \boldsymbol{\lambda}_2'\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{\lambda}_k'\boldsymbol{\beta} \end{pmatrix}.$$

Because $\boldsymbol{\tau} = \boldsymbol{\Lambda}'\boldsymbol{\beta}$ is estimable, it follows that

$$\widehat{\boldsymbol{\tau}} = \boldsymbol{\Lambda}'\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_k(\boldsymbol{\Lambda}'\boldsymbol{\beta}, \sigma^2\mathbf{H}),$$

where $\mathbf{H} = \boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\Lambda}$. Furthermore, because $\boldsymbol{\lambda}_1'\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda}_2'\widehat{\boldsymbol{\beta}}, ..., \boldsymbol{\lambda}_k'\widehat{\boldsymbol{\beta}}$ are jointly normal, we have that

$$\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\lambda}_j'\boldsymbol{\beta}, \sigma^2 h_{jj}),$$

where h_{jj} is the jth diagonal element of $\mathbf{H}$. Using our previous results, we know that

$$\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} \pm t_{n-r,\alpha/2}\sqrt{\widehat{\sigma}^2 h_{jj}},$$

where $\widehat{\sigma}^2 = \text{MSE}$, is a $100(1-\alpha)$ percent confidence interval for $\boldsymbol{\lambda}_j'\boldsymbol{\beta}$, that is,

$$\text{pr}\left(\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} - t_{n-r,\alpha/2}\sqrt{\widehat{\sigma}^2 h_{jj}} < \boldsymbol{\lambda}_j'\boldsymbol{\beta} < \boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} + t_{n-r,\alpha/2}\sqrt{\widehat{\sigma}^2 h_{jj}}\right) = 1-\alpha.$$

This statement is true for a single interval.

SIMULTANEOUS COVERAGE: To investigate the simultaneous coverage probability of the set of intervals

$$\{\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} \pm t_{n-r,\alpha/2}\sqrt{\widehat{\sigma}^2 h_{jj}},\ j = 1, 2, ..., k\},$$

let E_j denote the event that interval j contains $\boldsymbol{\lambda}_j'\boldsymbol{\beta}$, that is, $\text{pr}(E_j) = 1-\alpha$, for $j = 1, 2, ..., k$. The probability that each of the k intervals includes their target $\boldsymbol{\lambda}_j'\boldsymbol{\beta}$ is

$$\text{pr}\left(\bigcap_{j=1}^{k} E_j\right) = 1 - \text{pr}\left(\bigcup_{j=1}^{k} \overline{E}_j\right),$$

by DeMorgan's Law. In turn, Boole's Inequality says that

$$\text{pr}\left(\bigcup_{j=1}^{k} \overline{E}_j\right) \leq \sum_{j=1}^{k} \text{pr}(\overline{E}_j) = k\alpha.$$

Thus, the probability that each interval contains its intended target is

$$\text{pr}\left(\bigcap_{j=1}^{k} E_j\right) \geq 1 - k\alpha.$$

Obviously, this lower bound $1 - k\alpha$ can be quite a bit lower than $1 - \alpha$, that is, the simultaneous coverage probability of the set of

$$\{\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} \pm t_{n-r,\alpha/2}\sqrt{\widehat{\sigma}^2 h_{jj}},\ j = 1, 2, ..., k\}$$

can be much lower than the single interval coverage probability.

GOAL: We would like the set of intervals $\{\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} \pm d\sqrt{\widehat{\sigma}^2 h_{jj}},\ j = 1, 2, ..., k\}$ to have a **simultaneous coverage probability** of at least $1 - \alpha$. Here, d represents a probability point that guarantees the desired simultaneous coverage. Because taking $d = t_{n-r,\alpha/2}$ does not guarantee this minimum, we need to take d to be larger.

BONFERRONI: From the argument on the last page, it is clear that if one takes $d = t_{n-r,\alpha/2k}$, then

$$\text{pr}\left(\bigcap_{j=1}^{k} E_j\right) \geq 1 - k(\alpha/k) = 1 - \alpha.$$

Thus, $100(1 - \alpha)$ percent simultaneous confidence intervals for $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, ..., \boldsymbol{\lambda}_k'\boldsymbol{\beta}$ are

$$\boldsymbol{\lambda}_j'\widehat{\boldsymbol{\beta}} \pm t_{n-r,\alpha/2k}\sqrt{\widehat{\sigma}^2 h_{jj}}$$

for $j = 1, 2, ..., k$.

SCHEFFÉ: The idea behind Scheffé's approach is to consider an arbitrary linear combination of $\boldsymbol{\tau} = \boldsymbol{\Lambda}'\boldsymbol{\beta}$, say, $\mathbf{u}'\boldsymbol{\tau} = \mathbf{u}'\boldsymbol{\Lambda}'\boldsymbol{\beta}$ and construct a confidence interval

$$C(\mathbf{u}, d) = (\mathbf{u}'\widehat{\boldsymbol{\tau}} - d\sqrt{\widehat{\sigma}^2\mathbf{u}'\mathbf{H}\mathbf{u}},\ \mathbf{u}'\widehat{\boldsymbol{\tau}} + d\sqrt{\widehat{\sigma}^2\mathbf{u}'\mathbf{H}\mathbf{u}}),$$

where d is chosen so that

$$\text{pr}\{\mathbf{u}'\boldsymbol{\tau} \in C(\mathbf{u}, d),\ \text{for all } \mathbf{u}\} = 1 - \alpha.$$

Since d is chosen in this way, one guarantees the necessary simultaneous coverage probability for all possible linear combinations of $\boldsymbol{\tau} = \boldsymbol{\Lambda}'\boldsymbol{\beta}$ (an infinite number of combinations). Clearly, the desired simultaneous coverage is then conferred for the k functions of interest $\tau_j = \boldsymbol{\lambda}_j'\boldsymbol{\beta}$, $j = 1, 2, ..., k$; these functions result from taking $\mathbf{u}$ to be the standard unit vectors. The argument in Monahan (pp 144) shows that $d = (kF_{k,n-r,\alpha})^{1/2}$.