Chapter 4. Models not of full rank

Chapter 3 discussed regression analysis for a model $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{0,p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, where $\boldsymbol{X}$ has full-column rank. Chapter 4 illustrated how the same equation can apply to linear models when X may not have full-column rank. We shall consider estimation and hypothesis testing for the non-full-rank case. We will follow the same sequence of development as in Chapter 3.

4.1 The Normal Equations

As was done in Chapter 3, the normal equations corresponding to the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is derived using least square criterion. As before, when $var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, the normal equations turn out to be

$$\boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}. \tag{1}$$

We will discuss later how to solve the equation when $\boldsymbol{X}^\top \boldsymbol{X}$ is not of full rank. We first present an illustrative example.

Example 1(Finding the Normal Equations). Deoxyribonucleic acid (DNA) is the hereditary material found in most organisms. A genome is an organism's complete set of DNA, including all of its genes. Each genome contains all the information needed to build and maintain that organism.

Macdonald (2015) presents an example with genome size measured in pictograms (trillionths of a gram) of DNA per haploid cell in several large groups of crustaceans. The data are taken from Gregory (2015). For purposes of illustration, we shall consider six points for three kinds of crustaceans. We shall also use these data for subsequent examples to give numerical illustrations of the computations. For the entries of Table 4.1 below, let y_{ij} denote the DNA content of the j-th crustacean of the i-th type, i taking values 1,2 and 3 for amphipods, barnacles, and branchiopods, respectively, and $j = 1, 2, \ldots, n_i$, where n_i is the number of observations of the i-th type.

Table 4.1. Amount of DNA.

	Type of Crustacean		
	Amphipods	Barnacles	Branchiopods
	27.00	0.67	0.19
	50.91	0.90	
	64.62		
Total	142.53	1.57	0.19

The objective is to estimate the effect of the type of crustacean DNA content. To do this, we assume that the observation y_{ij} is the sum of three parts

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where μ represents the population mean of the DNA content of the crustaceans, α_i is the effect of type i DNA content, and ε_{ij} is a random error term peculiar to the observation y_{ij}. $\varepsilon_{ij} \perp\!\!\!\perp$ each other.

To develop the normal equations, we write down the six observations in terms of the equation of the model

$$
\begin{aligned}
27.00 &= y_{11} = \mu + \alpha_1 + \varepsilon_{11}, \\
50.91 &= y_{12} = \mu + \alpha_1 + \varepsilon_{12}, \\
64.62 &= y_{13} = \mu + \alpha_1 + \varepsilon_{13}, \\
0.67 &= y_{21} = \mu + \alpha_2 + \varepsilon_{21}, \\
0.90 &= y_{22} = \mu + \alpha_2 + \varepsilon_{22}, \\
0.19 &= y_{31} = \mu + \alpha_3 + \varepsilon_{31}.
\end{aligned}
$$

We may write these equations in matrix form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\operatorname{rank}(X) = 3$.

where $\boldsymbol{Y} = \begin{pmatrix} 27.00 \\ 50.91 \\ 64.62 \\ 0.67 \\ 0.90 \\ 0.19 \end{pmatrix}$, $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$ and $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \end{pmatrix}$.

Here $\boldsymbol{Y}$ is the vector of observations, $\boldsymbol{X}$ is the matrix of 0's and 1's. The vector $\boldsymbol{\beta}$ is the vector of parameters. It is the vector of all of the elements of the model. The matrix $\boldsymbol{X}$ is called the *incidence matrix*, or sometimes the design matrix. This is because the location of the 0's and 1's throughout its elements represents the incidence of the terms of the model among the observations and hence of the classifications in which the observations lie. Consider the normal equation (1). They involve $\boldsymbol{X}^\top\boldsymbol{X}$ a square and symmetric matrix. Its elements are the inner products of the columns of X with each other. We have that

$$
\boldsymbol{X}^\top\boldsymbol{X} = \begin{pmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.
$$

Since $\boldsymbol{X}$ does not have full-column rank, $\boldsymbol{X}^\top\boldsymbol{X}$ is not of full rank. The normal equations also involve the vector $\boldsymbol{X}^\top\boldsymbol{Y}$. Its elements are the inner products of the column of $\boldsymbol{X}$ with the vector $\boldsymbol{y}$.

$$
\boldsymbol{X}^\top\boldsymbol{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{pmatrix} = \begin{pmatrix} y_{11} + y_{12} + y_{13} + y_{21} + y_{22} + y_{31} \\ y_{11} + y_{12} + y_{13} \\ y_{21} + y_{22} \\ y_{31} \end{pmatrix} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{pmatrix} = \begin{pmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \end{pmatrix}.
$$

As has already been mentioned earlier, when $\boldsymbol{X}^\top \boldsymbol{X}$ is not of full rank, the normal equation cannot be solved with one solitary solution $\hat{\boldsymbol{\beta}} = \boldsymbol{(X^\top X)^{-1} X^\top Y}$ as in Chapter 3. To emphasize this, we write the normal equations as

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{\beta}^0 = \boldsymbol{X}^\top \boldsymbol{y}. \tag{2}$$

We use the symbol $\boldsymbol{\beta}^0$ to distinguish the many solutions of (2) from the solitary solution that exists when $\boldsymbol{X}^\top \boldsymbol{X}$ is of full rank. We shall also use $\boldsymbol{\beta}^0$ to denote a solution $\boldsymbol{G}\boldsymbol{X}^\top \boldsymbol{y}$ to (2). Let $\boldsymbol{G}$ be any generalized inverse of $\boldsymbol{X}^\top \boldsymbol{X}$. Then, it can be seen that

Recall A's generalized inverse G s.t. $AGA = A$.

$$\boldsymbol{\beta}^0 = \boldsymbol{G}\boldsymbol{X}^\top \boldsymbol{y}.$$

is a solution of (2). This is because the L.H.S of (2) is

$\underline{\underline{XGX^T}}$ is invariant wrt G.

K (projection matrix).

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{G}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{y}.$$

$X^T K = X^T$

However, the solution to (2) is not unique. For example 1, one can take $X^T \underline{\underline{XGX^T}} = X'$

$$\boldsymbol{G} = \boldsymbol{G_1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{or}$$

$$\boldsymbol{G} = \boldsymbol{G_2} = \begin{pmatrix} 1 & -1 & -1 & 0 \\ -1 & \frac{4}{3} & 1 & 0 \\ -1 & 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

$$\boldsymbol{\beta}_1^0 = \boldsymbol{G_1}\boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 47.51 \\ 0.785 \\ 0.19 \end{pmatrix}.$$

$$\beta_2^0 = G_2 X^\top y = \begin{pmatrix} 1 & -1 & -1 & 0 \\ -1 & \frac{4}{3} & 1 & 0 \\ -1 & 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 144.29 \\ 142.53 \\ 1.57 \\ 0.19 \end{pmatrix}$$

$$= \begin{pmatrix} 0.19 \\ 47.32 \\ 0.595 \\ 0 \end{pmatrix}.$$

$$\Rightarrow \beta_1^0 \neq \beta_2^0$$

Notice that for both solutions $\mu + \alpha_i$ and $\alpha_i - \alpha_j$, $i, j = 1, 2, 3$ are equal.

4.2 Consequences of a solution.

We next present the properties of the LS estimate for non-full rank linear model as follows:

1. Expected value of β^0.

$$\begin{aligned} E(\beta^0) &= E(GX^\top y) \quad (G \text{ is a generalized inverse of } X^\top X) \\ &= GX^\top E(y) \\ &= GX^\top X\beta \\ &= H\beta. \quad (\text{let } H \equiv GX^\top X) \end{aligned}$$

$(GX^\top X)$

$\neq I_p$. idempotent.

$\Rightarrow \beta^0$ is an unbiased estimator of $H\beta$ but not β.

2. Variance Covariance matrix of β^0.

$$\begin{aligned} \mathrm{Var}(\beta^0) &= \mathrm{Var}(GX^\top y) \\ &= GX^\top \mathrm{Var}(y) XG^\top \\ &= GX^\top (\sigma^2 I) XG^\top \\ &= GX^\top XG^\top \sigma^2. \end{aligned}$$

$<p$.

Note that this is not an analogue of its counterpart $(X^\top X)^{-1}\sigma^2$ as would be $G\sigma^2$.

3. The fitted model

$\hat{\beta} = \underbrace{(X^\top X)^{-1}}_{G} X y$. for full rank case. for non-full rank case.

$$\begin{aligned} \hat{y} &= X\beta^0 \\ &= XGX^\top y. \end{aligned}$$

$= Ky$. K is the invariant projection matrix.

Recall that $\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top$ is invariant to the choice of $\boldsymbol{G}$, hence the fitted value $\hat{\boldsymbol{y}}$ is also invariant to different choices of $\boldsymbol{G}$.

4. Estimating $E(\mathbf{y})$.

$$\begin{aligned} E(\hat{\boldsymbol{y}}) &= E(\boldsymbol{X}\boldsymbol{\beta}^0) \\ &= \boldsymbol{X}E(\boldsymbol{\beta}^0) \\ &= \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{X}\boldsymbol{\beta}, \end{aligned}$$

which implies that $E(\hat{\boldsymbol{y}})$ is invariant to the choice of $\boldsymbol{G}$.

5. Residual error sum of squares.

$$\begin{aligned} \text{SSE} &= (\boldsymbol{y}-\hat{\boldsymbol{y}})^\top(\boldsymbol{y}-\hat{\boldsymbol{y}}) \\ &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{\mathbf{0}})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{\mathbf{0}}) \\ &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{y})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{y}) \\ &= \boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\boldsymbol{y} \\ &= \boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)^\top\boldsymbol{y}. \end{aligned}$$

Recall that $\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top$ is symmetric and is invariant to choice of $\boldsymbol{G}$; see the supplementary note of generalized inverse. As a result, the SSE is invariant to choice of $\boldsymbol{G}$. Moreover, note that

$$\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top$$

$KK = K$ idempotent and invariant.

Recall GX^TX is also an idempotent matrix.

by the definition of generalized inverse. Hence, $\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top$ is symmetric and idempotent, so is $\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top$.

6. Regression sum of squares.

$$\begin{aligned} SSR &= SST - SSE \\ &= \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\boldsymbol{y} \\ &= \boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{y} \\ &= (\boldsymbol{\beta}^{\mathbf{0}})^\top\boldsymbol{X}^\top\boldsymbol{y}. \quad (\Rightarrow \text{ invariant to } \boldsymbol{G}) \end{aligned}$$

Similar to Property 5, the SSR is also invariant to the choice of $\boldsymbol{G}$.

7. Estimating the residual error sum of squares.

$$\begin{aligned} E(SSE) &= E[\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{XGX}^\top)\boldsymbol{y}] \\ &= tr((\boldsymbol{I}-\boldsymbol{XGX}^\top)\sigma^2\boldsymbol{I}) + \boldsymbol{\beta}^\top\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{XGX}^\top)\boldsymbol{X\beta} \\ &= \sigma^2 tr(\boldsymbol{I}-\boldsymbol{XGX}^\top) \\ &= \sigma^2 r(\boldsymbol{I}-\boldsymbol{XGX}^\top) \\ &= \sigma^2(n-r(\boldsymbol{X})) \\ \Rightarrow & E(\frac{SSE}{n-r(\boldsymbol{X})}) = \sigma^2 \\ \Rightarrow & \hat{\sigma}^2 = \frac{SSE}{n-r(\boldsymbol{X})} \text{ is an unbiased estimator of } \sigma^2. \end{aligned}$$

8. Partitioning the Total sum of squares.
   Note that $SSM = \frac{1}{n}\boldsymbol{y}^\top\mathbf{1}\mathbf{1}^\top\boldsymbol{y}$ is the sum of squares due to fitting the general mean and

$$\begin{aligned} SSR_m &= SSR - SSM \\ &= (\boldsymbol{\beta}^{\mathbf{0}})^\top\boldsymbol{X}^\top\boldsymbol{y} - \frac{1}{n}\boldsymbol{y}^\top\mathbf{1}\mathbf{1}^\top\boldsymbol{y} \\ &= \boldsymbol{y}^\top\boldsymbol{XGX}^\top\boldsymbol{y} - \boldsymbol{y}^\top\frac{\mathbf{1}\mathbf{1}^\top}{\boldsymbol{n}}\boldsymbol{y} \\ &= \boldsymbol{y}^\top(\boldsymbol{XGX}^\top - \frac{\mathbf{1}\mathbf{1}^\top}{n})\boldsymbol{y} \end{aligned}$$

   is the sum of square for fitting the model, corrected the mean. Thus,

$$\begin{aligned} SST_m = SST - SSM &= \boldsymbol{y}^\top\boldsymbol{y} - \frac{1}{n}\boldsymbol{y}^\top\mathbf{1}\mathbf{1}^\top\boldsymbol{y} \\ &= \boldsymbol{y}^\top(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\boldsymbol{y} \end{aligned}$$

   is the total sum of squares corrected for the mean.

4.3 Distributional Properties

We now assume assume normality for the error terms. Thus, we have

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}).$$

Using the normality assumption, we shall derive the distributional properties of $\boldsymbol{y}$ in a manner similar to the full-rank case.

1. The observation vector $\boldsymbol{y}$ is normal distributed, i.e, $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$.

2. The solution to the normal equation $\boldsymbol{\beta}^0$ is normally distributed, i.e,

$$\boldsymbol{\beta}^0 = \boldsymbol{GX}^\top\boldsymbol{y} \sim N(\boldsymbol{GX}^\top\boldsymbol{X\beta}, \boldsymbol{GX}^\top\boldsymbol{XG}^\top\sigma^2).$$

   Note that the covariance matrix of $\boldsymbol{\beta}^0$ is singular. (Why?)

3. $\boldsymbol{\beta}^0$ and $\hat{\sigma}^2$ are independent.

   Since

$$\boldsymbol{\beta}^0 = \boldsymbol{G}\boldsymbol{X}^\top \boldsymbol{y}, \quad B = GX^\top$$
$$SSE = \boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\boldsymbol{y}. \quad A = I - XGX^\top$$

$\Sigma = \sigma^2 I$

   Thus,

$B\Sigma A = 0$.

$$\begin{aligned}&\boldsymbol{G}\boldsymbol{X}^\top(\boldsymbol{I}\sigma^2)(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\\ &= \sigma^2\boldsymbol{G}\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top) = \boldsymbol{0} \ \text{ because } \boldsymbol{X}^\top = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top,\\ \Rightarrow\ &\boldsymbol{\beta}^0\text{and } \hat{\sigma}^2 \text{ are independent.}\end{aligned}$$

4. The SSE divided by the population variance, SEE/σ^2 is Chi-square distributed.

   We have that

$$\frac{SSE}{\sigma^2} = \frac{\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\boldsymbol{y}}{\sigma^2}.$$

   Since

$$\frac{(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)}{\sigma^2}\boldsymbol{I}\sigma^2 = \boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top \text{ is idempotent matrix,}$$

   and

$$\begin{aligned}rank\left(\frac{\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top}{\sigma^2}\right) &= rank(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\\ &= n - r(\boldsymbol{X}),\end{aligned}$$

   we have

$$\frac{SSE}{\sigma^2} \sim \chi^2_{(n-r(\boldsymbol{X}),\frac{1}{2\sigma^2}\beta^\top\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta})}.$$

   However,

$$\frac{1}{2\sigma^2}\beta^\top\underbrace{\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta}}_{0} = 0,$$

   it reduces that

$$\frac{SSE}{\sigma^2} \sim \chi^2_{(n-r(\boldsymbol{X}))}.$$

5. Non-central χ^2's.

   With SSE/σ^2 being central Chi-square, we now show that other terms in partitioning sum of squares have non-central Chi-square distribution independent of SSE. This leads to F-statistics that have non-central F-distributions that, in turn, are central F-distributions under certain null hypothesis. Tests of hypothesis are thus established. Recall that

$$\begin{aligned}SSR &= \boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{y}\\ \Rightarrow\ \frac{SSR}{\sigma^2} &= \frac{\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^\top\boldsymbol{y}}{\sigma^2}\end{aligned}$$

and $\frac{\boldsymbol{XGX}^\top}{\sigma^2}$ is idempotent implying $rank(\frac{\boldsymbol{XGX}^\top}{\sigma^2}) = r(\boldsymbol{X})$. Therefore,

$$\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{XGX}^\top \boldsymbol{X\beta} = \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X\beta}.$$

As a result, $\frac{\boldsymbol{XGX}^\top}{\sigma^2} \sim \chi^2_{(r(\boldsymbol{X}), \frac{1}{\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X\beta})}$.

6. SSE and SSR are independent.

   Since $(\boldsymbol{XGX}^\top)\boldsymbol{I}\sigma^2(\boldsymbol{I} - \boldsymbol{XGX}^\top) = 0$

   $\Rightarrow$ SSE and SSR are independent.

7. Let MSR=$\frac{SSR}{r(\boldsymbol{X})}$ and MSE=$\frac{SSE}{n-r(\boldsymbol{X})}$. Then

$$\frac{MSR}{MSE} = F(R) \sim F_{(r(\boldsymbol{X}), n-r(\boldsymbol{X}), \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X\beta})}.$$

8. Similarly,

$$\frac{SSM}{\sigma^2} = \frac{\boldsymbol{y}^\top \boldsymbol{n}^{-1}\boldsymbol{11}^\top \boldsymbol{y}}{\sigma^2}$$

   where $n^{-1}\boldsymbol{11}^\top$ is idempotent with rank 1,

$$\begin{aligned}\frac{SSM}{\sigma^2} &\sim \chi^2_{(1, \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{11})^\top \boldsymbol{X\beta})} \\ &= \chi^2_{(1, \frac{1}{2\sigma^2}(\boldsymbol{1}^\top \boldsymbol{X\beta})^2)}.\end{aligned}$$

   Moreover, it can be easily checked that $1^\top \boldsymbol{XGX}^\top = \boldsymbol{1}^\top$ and $n^{-1}\boldsymbol{11}^\top \boldsymbol{XGX}^\top = n^{-1}\boldsymbol{11}^\top$. Hence, the products of $n^{-1}\boldsymbol{11}^\top$ and $(\boldsymbol{I} - \boldsymbol{XGX}^\top)$ are null. Hence, SSM is distributed independently of SSE.

9. Similarly,

$$\begin{aligned}\frac{SSR_m}{\sigma^2} &= \boldsymbol{y}^\top(\boldsymbol{XGX}^\top - \frac{\boldsymbol{11}^\top}{\boldsymbol{n}})\boldsymbol{y} \\ &\sim \chi^2_{(r(\boldsymbol{X})-1, \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{I} - \frac{\boldsymbol{11}^\top}{\boldsymbol{n}})\boldsymbol{X\beta})}.\end{aligned}$$

Classic but not very useful

10. Non-central F-distributions.

    Based on above, we find that

$$F(R) = \frac{SSR/r(\boldsymbol{X})}{SSE/n - r(\boldsymbol{X})} \sim F_{(r(\boldsymbol{X}), n-r(\boldsymbol{X}), \frac{\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X\beta}}{2\sigma^2})},$$

$$F(M) = \frac{SSM}{SSE/n - r(\boldsymbol{X})} \sim F_{(1, n-r(\boldsymbol{X}), \frac{1}{\sigma^2}(\boldsymbol{1}^\top \boldsymbol{X\beta})^2)},$$

$$F(R_m) = \frac{SSR_m/(r(\boldsymbol{X}) - 1)}{SSE/n - r(\boldsymbol{X})} \sim F_{(r(\boldsymbol{X})-1, n-r(\boldsymbol{X}), \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{I} - \frac{\boldsymbol{11}^\top}{\boldsymbol{n}})\boldsymbol{X\beta})}.$$

Under certain null hypothesis, these non-central F's become central F's and so provide us with tests of hypothesis. We shall discuss these in later sections.

4.4 Estimable Functions

4.4.1* Identifiability

To illustrate the definition of identifiability, we take the modeling of the mean of $\boldsymbol{Y}$ as a toy example. Generally speaking, to parametrize $\mu = E(\boldsymbol{Y})$ is to write μ as a function of some parameters $\mu = f(\beta)$. The linear model parameterize μ as $\boldsymbol{X\beta}$ since $\mu = E(\boldsymbol{Y}) = E(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta}$ when assuming $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and fixed design.

Informally, a parametrization is said to be *identifiable* if knowing μ means knowing $\boldsymbol{\beta}$.

Formally, the parameter $\boldsymbol{\beta}$ is identifiable if $f(\boldsymbol{\beta}_1) = f(\boldsymbol{\beta}_2)$ implies that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. If $\boldsymbol{\beta}$ is identifiable, the parametrization $f(\boldsymbol{\beta})$ is. More generally, the vector-valued function $g(\boldsymbol{\beta})$ is identifiable if $f(\boldsymbol{\beta}_1) = f(\boldsymbol{\beta}_2)$ implies that $g(\boldsymbol{\beta}_1) = g(\boldsymbol{\beta}_2)$.

Remark. The contrapositive

$$\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \Rightarrow \mu_1 = f(\boldsymbol{\beta}_1) \neq f(\boldsymbol{\beta}_2) = \mu_2$$

is perhaps more revealing. This says that a difference in the parameter values will show itself as a difference in the means. If a difference cannot show itself in the means, collecting even an infinite amount of data regarding the mean will not allow us to learn the parameter.

Proposition 1. *In a linear model for which $\boldsymbol{X}$ is of full rank, $\boldsymbol{\beta}$ is identifiable.*

Proof. Suppose that $\boldsymbol{X\beta_1} = \boldsymbol{X\beta_2}$. Since $\boldsymbol{X}$ is of full rank, $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsigular. Thus we can write

$\{f(\beta_1) = f(\beta_2) \Rightarrow \beta_1 = \beta_2\}$ identifiable.

$$\boldsymbol{\beta}_1 = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}_1 = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2.$$

Why we can claim if X non-full rank $\Rightarrow$ exist $\beta_1 \neq \beta_2$ such that $X\beta_1 = X\beta_2$?

But if $\boldsymbol{X}$ is not of full rank, there exist $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ for which $\boldsymbol{X\beta_1} = \boldsymbol{X\beta_2}$. Thus, $\boldsymbol{\beta}$ is not identifiable in the non-full rank case. Hence, in a linear models, it can be seen that the only function of the parameters that are identifiable are functions of $\boldsymbol{X\beta}$.

If we would like $\beta = (h \circ f)(\beta)$. which mean $f(\beta) = X\beta$ is identifiable. (no information loss).

Proposition 2. *A function $g(\boldsymbol{\beta})$ is identifiable if and only if $g(\boldsymbol{\beta}) = (h \circ f)(\boldsymbol{\beta})$ for some function h.*

Proof. (Sufficiency) Suppose that $g = h \circ f$ for some h. If $f(\boldsymbol{\beta}_1) = f(\boldsymbol{\beta}_2)$, we have

$$g(\boldsymbol{\beta}_1) = h\{f(\boldsymbol{\beta}_1)\} = h\{f(\boldsymbol{\beta}_2)\} = g(\boldsymbol{\beta}_2),$$

and we can see that $g(\boldsymbol{\beta})$ is identifiable.

(Necessity) Now suppose that g is not a function of f. Then there must exist $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ such that $f(\boldsymbol{\beta}_1) = f(\boldsymbol{\beta}_2)$ but $g(\boldsymbol{\beta}_1) \neq g(\boldsymbol{\beta}_2)$. And so $g(\boldsymbol{\beta})$ is not identifiable. We finish the proof.

Hence, it makes sense to estimate an identifiable function. In a linear model, it makes sense to estimate only functions of $\boldsymbol{X}\boldsymbol{\beta}$. However, if $\boldsymbol{X}$ is not of full rank or is rank deficient, $\boldsymbol{\beta}$ is not a function of $\boldsymbol{X}\boldsymbol{\beta}$, as shown above.

same input but different outputs.