1

The Linear Model

The application of econometrics requires more than mastering a collection of tricks. It also requires insight, intuition, and common sense.

(Jan R. Magnus, 2017, p. 31)

The natural starting point for learning about statistical data analysis is with a sample of independent and identically distributed (hereafter i.i.d.) data, say $Y = (Y_1, \ldots, Y_n)$, as was done in book III. The *linear regression model* relaxes both the identical and independent assumptions by (i) allowing the means of the Y_i to depend, in a linear way, on a set of other variables, (ii) allowing for the Y_i to have different variances, and (iii) allowing for correlation between the Y_i.

The linear regression model is not only of fundamental importance in a large variety of quantitative disciplines, but is also the basis of a large number of more complex models, such as those arising in panel data studies, time-series analysis, and generalized linear models (GLIM), the latter briefly introduced in Section 1.6. Numerous, more advanced data analysis techniques (often referred to now as algorithms) also have their roots in regression, such as the *least absolute shrinkage and selection operator* (LASSO), the *elastic net*, and *least angle regression* (LARS). Such methods are often now showcased under the heading of machine learning.

1.1 Regression, Correlation, and Causality

It is uncomfortably true, although rarely admitted in statistics texts, that many important areas of science are stubbornly impervious to experimental designs based on randomisation of treatments to experimental units. Historically, the response to this embarrassing problem has been to either ignore it or to banish the very notion of causality from the language and to claim that the shadows dancing on the screen are all that exists.

Ignoring the problem doesn't make it go away and defining a problem out of existence doesn't make it so. We need to know what we can safely infer about causes from their observational shadows, what we can't infer, and the degree of ambiguity that remains.

(Bill Shipley, 2016, p. 1)[1]

1 The metaphor to dancing shadows goes back a while, at least to Plato's Republic and the Allegory of the Cave. One can see it today in shadow theater, popular in Southeast Asia; see, e.g., Pigliucci and Kaplan (2006, p. 2).

Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH, First Edition. Marc S. Paolella.

The univariate linear regression model relates the scalar random variable Y to k other (possibly random) variables, or **regressors**, $x_1, \dots, x_k$ in a linear fashion,

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \tag{1.1}$$

where, typically, $\epsilon \sim \mathrm{N}(0, \sigma^2)$. Values $\beta_1, \dots, \beta_k$ and σ^2 are unknown, constant parameters to be estimated from the data. A more useful notation that also emphasizes that the means of the Y_i are not constant is

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i, \quad i = 1, 2, \dots, n, \tag{1.2}$$

where now a double subscript on the regressors is necessary. The ϵ_i represent the difference between the values of Y_i and the model used to represent them, $\sum_{j=1}^{k} \beta_j x_{i,j}$, and so are referred to as the **error terms**. It is important to emphasize that the error terms are i.i.d., but the Y_i are not. However, if we take $k = 1$ and $x_{i,1} \equiv 1$, then (1.2) reduces to $Y_i = \beta_1 + \epsilon_i$, which is indeed just the i.i.d. model with $Y_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\beta_1, \sigma^2)$. In fact, it is usually the case that $x_{i,1} \equiv 1$ for any $k \geqslant 1$, in which case the model is said to **include a constant** or **have an intercept term**.

We refer to Y as the **dependent** (random) variable. In other contexts, Y is also called the **endogenous** variable, while the k regressors can also be referred to as the **explanatory**, **exogenous**, or **independent** variables, although the latter term should not be taken to imply that the regressors, when viewed as random variables, are necessarily independent from one another.

The linear structure of (1.1) is one way of building a relationship between the Y_i and a set of variables that "influence" or "explain" them. The usefulness of establishing such a relationship or **conditional** model for the Y_i can be seen in a simple example: Assume a demographer is interested in the income of people living and employed in Hamburg. A random sample of n individuals could be obtained using public records or a phone book, and (rather unrealistically) their incomes Y_i, $i = 1, \dots, n$, elicited. Assuming that income is approximately normally distributed, an **unconditional** model for income could be postulated as $\mathrm{N}(\mu_u, \sigma_u^2)$, where the subscript u denotes the unconditional model and the usual estimators for the mean and variance of a normal sample could be used.

(We emphasize that this example is just an excuse to discuss some concepts. While actual incomes for certain populations can be "reasonably" approximated as Gaussian, they are, of course, not: They are strictly positive, will thus have an extended right tail, and this tail might be heavy, in the sense of being Pareto—this naming being no coincidence, as Vilfredo Pareto worked on modeling incomes, and is also the source of what is now referred to in micro-economics as Pareto optimality. An alternative type of linear model, referred to as GLIM, that uses a non-Gaussian distribution instead of the normal, is briefly discussed below in Section 1.6. Furthermore, interest might not center on modeling the mean income—which is what regression does—but rather the median, or the lower or upper quantiles. This leads to quantile regression, also briefly discussed in Section 1.6.)

A potentially much more precise description of income can be obtained by taking certain factors into consideration that are highly related to income, such as age, level of education, number of years of experience, gender, whether he or she works part or full time, etc. Before continuing this simple example, it is imperative to discuss the three Cs: correlation, causality, and control.

Observe that (simplistically here, for demonstration) age and education might be positively correlated, simply because, as the years go by, people have opportunities to further their schooling and training. As such, if one were to claim that income tends to increase as a function of age, then one cannot conclude this arises out of "seniority" at work, but rather possibly because some of the older people

have received more schooling. Another way of saying this is, while income and age are positively correlated, an increase in age is not necessarily **causal** for income; age and income may be **spuriously correlated**, meaning that their correlation is driven by other factors, such as education, which might indeed be causal for income. Likewise, if one were to claim that income tends to increase with educational levels, then one cannot claim this is due to education *per se*, but rather due simply to seniority at the workplace, possibly despite their enhanced education. Thus, it is important to include both of these variables in the regression.

In the former case, if a positive relationship is found between income and age *with education also in the regression*, then one can conclude a seniority effect. In the literature, one might say "Age appears to be a significant predictor of income, and this being concluded after having also **controlled for** education." Examples of controlling for the relevant factors when assessing causality are ubiquitous in empirical studies of all kinds, and are essential for reliable inference. As one example, in the field of "economics and religion" (which is now a fully established area in economics; see, e.g., McCleary, 2011), in the abstract of one of the highly influential papers in the field, Gruber (2005) states "Religion plays an important role in the lives of many Americans, but there is relatively little study by economists of the implications of religiosity for economic outcomes. This likely reflects the enormous difficulty inherent in separating the causal effects of religiosity from other factors that are correlated with outcomes." The paper is filled with the expression "having controlled for".

A famous example, in a famous paper, is Leamer (1983, Sec. V), showing how conclusions from a study of the factors influencing the murder rate are highly dependent on which set of variables are included in the regression. The notion of controlling for the right variables is often the vehicle for critiquing other studies in an attempt to correct potentially wrong conclusions. For example, Farkas and Vicknair (1996, p. 557) state "[Cancio et al.] claim that discrimination, measured as a residual from an earnings attainment regression, increased after 1976. Their claim depends crucially on which variables are controlled and which variables are omitted from the regression. We believe that the authors have omitted the key control variable—cognitive skill."

The concept of causality is fundamental in econometrics and other social sciences, and we have not even scratched the surface. The different ways it is addressed in popular econometrics textbooks is discussed in Chen and Pearl (2013), and debated in Swamy et al. (2015), Raunig (2017), and Swamy et al. (2017). These serve to indicate that the theoretical framework for understanding causality and its interface to statistical inference is still developing. The importance of causality for scientific inquiry cannot be overstated, and continues to grow in importance in light of artificial intelligence. As a simple example, humans understand that weather is (global warming aside) exogenous, and carrying an umbrella does not cause rain. How should a computer know this? Starting points for further reading include Pearl (2009), Shipley (2016), and the references therein.

Our development of the linear model in this chapter serves two purposes: First, it is the required theoretical statistical framework for understanding ANOVA models, as introduced in Chapters 2 and 3. As ANOVA involves designed experiments and randomization, as opposed to observational studies in the social sciences, we can avoid the delicate issues associated with assessing causality. Second, the linear model serves as the underlying structure of autoregressive time-series models as developed in Part II, and our emphasis is on statistical forecasting, as opposed to the development of structural economic models that explicitly need to address causality.

We now continue with our very simple illustration, just to introduce some terminology. Let $x_{i,2}$ denote the age of the ith person. A conditional model with a constant and age as a regressor is given by $Y_i = \beta_1 + \beta_2 x_{i,2} + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. The intercept is measured by β_1 and the slope of income

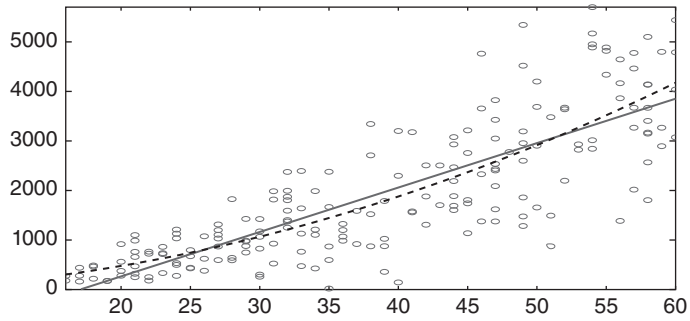

Figure 1.1 Scatterplot of age versus income overlaid with fitted regression curves.

is measured by β_2. Because age is expected to explain a considerable part of variability in income, we expect σ^2 to be significantly less than σ_u^2. A useful way of visualizing the model is with a scatterplot of $x_{i,2}$ and y_i. Figure 1.1 shows such a graph based on a fictitious set of data for 200 individuals between the ages of 16 and 60 and their monthly net income in euros. It is quite clear from the scatterplot that age and income are positively correlated. If age is neglected, then the i.i.d. normal model for income results in $\hat{\mu}_u = 1{,}797$ euros and $\hat{\sigma}_u = 1{,}320$ euros. Using the techniques discussed below, the regression model gives estimates $\hat{\beta}_1 = -1{,}465$, $\hat{\beta}_2 = 85.4$, and $\hat{\sigma} = 755$, the latter being about 43% smaller than $\hat{\sigma}_u$. The model implies that, conditional on the age x, the income Y is modeled as $\mathrm{N}(-1{,}465 + 85.4x, 755^2)$. This is valid only for $16 \leqslant x \leqslant 60$; because of the negative intercept, small values of age would erroneously imply a negative income. The fitted model $y = \hat{\beta}_1 + \hat{\beta}_2 x$ is overlaid in the figure as a solid line.

Notice in Figure 1.1 that the linear approximation underestimates income for both low and high age groups, i.e., income does not seem perfectly linear in age, but rather somewhat quadratic. To accommodate this, we can add another regressor, $x_{i,3} = x_{i,2}^2$, into the model, i.e., $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_q^2)$ and σ_q^2 denotes the conditional variance based on the quadratic model. It is important to realize that the model is still linear (in the constant, age, and age squared). The fitted model turns out to be $Y_i = 190 - 12.5x_{i,2} + 1.29x_{i,3}$, with $\hat{\sigma}_q = 733$, which is about 3% smaller than $\hat{\sigma}$. The fitted curve is shown in Figure 1.1 as a dashed line.

One caveat still remains with the model for income based on age: The variance of income appears to increase with age. This is a typical finding with income data and agrees with economic theory. It implies that both the mean and the variance of income are functions of age. In general, when the variance of the regression error term is not constant, it is said to be **heteroskedastic**, as opposed to **homoskedastic**. The generalized least squares extension of the linear regression model discussed below can be used to address this issue when the structure of the heteroskedasticity as a function of the $\mathbf{X}$ matrix is known.

In certain applications, the ordering of the dependent variable and the regressors is important because they are observed in time, usually equally spaced. Because of this, the notation Y_t will be used, $t = 1, \ldots, T$. Thus, (1.2) becomes

$$Y_t = \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots + \beta_k x_{t,k} + \epsilon_t, \quad t = 1, 2, \ldots, T,$$

where $x_{t,i}$ indicates the tth observation of the ith explanatory variable, $i = 1, \ldots, k$, and ϵ_t is the tth error term. In standard matrix notation, the model can be compactly expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.3}$$

where $[\mathbf{X}]_{t,i} = x_{t,i}$, i.e., with $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,k})'$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ x_{T,1} & x_{T,2} & & x_{T,k} \end{bmatrix}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}),$$

$\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $T \times 1$, $\mathbf{X}$ is $T \times k$ and $\boldsymbol{\beta}$ is $k \times 1$. The first column of $\mathbf{X}$ is usually $\mathbf{1}$, the column of ones. Observe that $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$.

An important special case of (1.3) is with $k = 2$ and $x_{t,1} = 1$. Then $Y_t = \beta_1 + \beta_2 X_t + \epsilon_t$, $t = 1, \ldots, T$, is referred to as the **simple linear regression model**. See Problems 1.1 and 1.2.

1.2 Ordinary and Generalized Least Squares

1.2.1 Ordinary Least Squares Estimation

The most popular way of estimating the k parameters in $\boldsymbol{\beta}$ is the **method of least squares**,[2] which takes $\widehat{\boldsymbol{\beta}} = \arg\min \mathsf{S}(\boldsymbol{\beta})$, where

$$\mathsf{S}(\boldsymbol{\beta}) = \mathsf{S}(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{t=1}^{T} (Y_t - \mathbf{x}_t'\boldsymbol{\beta})^2, \tag{1.4}$$

and we suppress the dependency of S on $\mathbf{Y}$ and $\mathbf{X}$ when they are clear from the context.

Assume that $\mathbf{X}$ is of full rank k. One procedure to obtain the solution, commonly shown in most books on regression (see, e.g., Seber and Lee, 2003, p. 38), uses matrix calculus; it yields $\partial\, \mathsf{S}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, and setting this to zero gives the solution

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.5}$$

This is referred to as the **ordinary least squares**, or o.l.s., estimator of $\boldsymbol{\beta}$. (The adjective "ordinary" is used to distinguish it from what is called generalized least squares, addressed in Section 1.2.3 below.) Notice that $\widehat{\boldsymbol{\beta}}$ is also the solution to what are referred to as the **normal equations**, given by

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}. \tag{1.6}$$

To verify that (1.5) indeed corresponds to the minimum of $\mathsf{S}(\boldsymbol{\beta})$, the second derivative is checked for positive definiteness, yielding $\partial^2\, \mathsf{S}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = 2\mathbf{X}'\mathbf{X}$, which is necessarily positive definite when $\mathbf{X}$ is full rank. Observe that, if $\mathbf{X}$ consists only of a column of ones, which we write as $\mathbf{X} = \mathbf{1}$, then $\widehat{\boldsymbol{\beta}}$ reduces to the mean, $\bar{Y}$, of the Y_t. Also, if $k = T$ (and $\mathbf{X}$ is full rank), then $\widehat{\boldsymbol{\beta}}$ reduces to $\mathbf{X}^{-1}\mathbf{Y}$, with $\mathsf{S}(\widehat{\boldsymbol{\beta}}) = 0$.

Observe that the derivation of $\widehat{\boldsymbol{\beta}}$ in (1.5) did not involve any explicit distributional assumptions. One consequence of this is that the estimator may not have any meaning if the maximally existing moment of the $\{\epsilon_t\}$ is too low. For example, take $\mathbf{X} = \mathbf{1}$ and $\{\epsilon_t\}$ to be i.i.d. Cauchy; then $\widehat{\beta} = \bar{Y}$ is a useless estimator. If we assume that the first moment of the $\{\epsilon_t\}$ exists and is zero, then, writing $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$, we see that $\widehat{\boldsymbol{\beta}}$ is unbiased:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\beta}. \tag{1.7}$$

2 This terminology dates back to Adrien-Marie Legendre (1752–1833), though the method is most associated in its origins with Carl Friedrich Gauss, (1777–1855). See Stigler (1981) for further details.

Next, if we have existence of second moments, and $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, then $\mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2)$ is given by

$$\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \sigma^2] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{1.8}$$

It turns out that $\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators; this result is often referred to as the **Gauss–Markov Theorem**, and expressed as saying that $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator, or BLUE. We outline the usual derivation, leaving the straightforward details to the reader. Let $\widehat{\boldsymbol{\beta}}^* = \mathbf{A}'\mathbf{Y}$, where $\mathbf{A}'$ is a $k \times T$ nonstochastic matrix (it can involve $\mathbf{X}$, but not $\mathbf{Y}$). Let $\mathbf{D} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. First calculate $\mathbb{E}[\widehat{\boldsymbol{\beta}}^*]$ and show that the unbiased property implies that $\mathbf{D}'\mathbf{X} = \mathbf{0}$. Next, calculate $\mathbb{V}(\widehat{\boldsymbol{\beta}}^* \mid \sigma^2)$ and show that $\mathbb{V}(\widehat{\boldsymbol{\beta}}^* \mid \sigma^2) = \mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) + \sigma^2\mathbf{D}'\mathbf{D}$. The result follows because $\mathbf{D}'\mathbf{D}$ is obviously positive semi-definite and the variance is minimized when $\mathbf{D} = \mathbf{0}$.

In many situations, it is reasonable to assume normality for the $\{\epsilon_t\}$, in which case we may easily estimate the $k+1$ unknown parameters σ^2 and β_i, $i = 1, \ldots, k$, by maximum likelihood. In particular, with

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{1.9}$$

and log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\,\mathsf{S}(\boldsymbol{\beta}), \tag{1.10}$$

where $\mathsf{S}(\boldsymbol{\beta})$ is given in (1.4), setting

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{2}{2\sigma^2}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{and} \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\,\mathsf{S}(\boldsymbol{\beta})$$

to zero yields the same estimator for $\boldsymbol{\beta}$ as given in (1.5) and $\tilde{\sigma}^2 = \mathsf{S}(\widehat{\boldsymbol{\beta}})/T$. It will be shown in Section 1.3.2 that the maximum likelihood estimator (hereafter m.l.e.) of σ^2 is biased, while estimator

$$\hat{\sigma}^2 = \mathsf{S}(\widehat{\boldsymbol{\beta}})/(T - k) \tag{1.11}$$

is unbiased.

As $\widehat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{Y}$, $(\widehat{\boldsymbol{\beta}} \mid \sigma^2)$ is multivariate normally distributed, and thus characterized by its first two moments. From (1.7) and (1.8), it follows that $(\widehat{\boldsymbol{\beta}} \mid \sigma^2) \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

1.2.2 Further Aspects of Regression and OLS

> The coefficient of multiple determination, R^2, is a measure many statisticians love to hate. This animosity exists primarily because the widespread use of R^2 inevitably leads to at least occasional misuse.
>
> (Richard Anderson-Sprecher, 1994)

In general, the quantity $\mathsf{S}(\widehat{\boldsymbol{\beta}})$ is referred to as the **residual sum of squares**, abbreviated RSS. The **explained sum of squares**, abbreviated ESS, is defined to be $\sum_{t=1}^{T}(\widehat{Y}_t - \bar{Y})^2$, where the *fitted value* of Y_t is $\widehat{Y}_t := \mathbf{x}_t'\widehat{\boldsymbol{\beta}}$, and the **total (corrected) sum of squares**, or TSS, is $\sum_{t=1}^{T}(Y_t - \bar{Y})^2$. (Annoyingly, both words "error" and "explained" start with an "e", and some presentations define SSE to be the error sum of squares, which is our RSS; see, e.g., Ravishanker and Dey, 2002, p. 101.)

The term *corrected* in the TSS refers to the adjustment of the Y_t for their mean. This is done because the mean is a "trivial" regressor that is not considered to do any real explaining of the dependent variable. Indeed, the total *uncorrected* sum of squares, $\sum_{t=1}^{T} Y_t^2$, could be made arbitrarily large just by adding a large enough constant value to the Y_t, and the model consisting of just the mean (i.e., an $\mathbf{X}$ matrix with just a column of ones) would have the appearance of explaining an arbitrarily large amount of the variation in the data.

While certainly $Y_t - \bar{Y} = (Y_t - \widehat{Y}_t) + (\widehat{Y}_t - \bar{Y})$, it is not immediately obvious that

$$\sum_{t=1}^{T} (Y_t - \bar{Y})^2 = \sum_{t=1}^{T} (Y_t - \widehat{Y}_t)^2 + \sum_{t=1}^{T} (\widehat{Y}_t - \bar{Y})^2,$$

i.e.,

$$\text{TSS} = \text{RSS} + \text{ESS}. \tag{1.12}$$

This fundamental identity is proven below in Section 1.3.2.

A popular statistic that measures the fraction of the variability of $\mathbf{Y}$ taken into account by a linear regression model that includes a constant, compared to use of just a constant (i.e., $\bar{Y}$), is the **coefficient of multiple determination**, designated as R^2, and defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\mathsf{S}(\widehat{\boldsymbol{\beta}}, \mathbf{Y}, \mathbf{X})}{\mathsf{S}(\bar{Y}, \mathbf{Y}, \mathbf{1})}, \tag{1.13}$$

where $\mathbf{1}$ is a T-length column of ones. The coefficient of multiple determination R^2 provides a measure of the extent to which the regressors "explain" the dependent variable over and above the contribution from just the constant term. It is important that $\mathbf{X}$ contain a constant or a set of variables whose linear combination yields a constant; see Becker and Kennedy (1992) and Anderson-Sprecher (1994) and the references therein for more detail on this point.

By construction, the observed R^2 is a number between zero and one. As with other quantities associated with regression (such as the nearly always reported "t-statistics" for assessing individual "significance" of the regressors), R^2 is a statistic (a function of the data but not of the unknown parameters) and thus *is a random variable*. In Section 1.4.4 we derive the F test for parameter restrictions. With J such linear restrictions, and $\widehat{\boldsymbol{\gamma}}$ referring to the restricted estimator, we will show (1.88), repeated here, as

$$F = \frac{[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})]/J}{\mathsf{S}(\widehat{\boldsymbol{\beta}})/(T-k)} \sim \text{F}(J, T-k), \tag{1.14}$$

under the null hypothesis H_0 that the J restrictions are true. Let $J = k - 1$ and $\widehat{\boldsymbol{\gamma}} = \bar{Y}$, so that the restricted model is that all regressor coefficients, *except the constant* are zero. Then, comparing (1.13) and (1.14),

$$F = \frac{T-k}{k-1} \frac{R^2}{1-R^2}, \qquad \text{or} \qquad R^2 = \frac{(k-1)F}{(T-k)+(k-1)F}. \tag{1.15}$$

Dividing the numerator and denominator of the latter expression by $T - k$ and recalling the relationship between F and beta random variables (see, e.g., Problem I.7.20), we immediately have that

$$R^2 \sim \text{Beta}\left(\frac{k-1}{2}, \frac{T-k}{2}\right), \tag{1.16}$$

so that $\mathbb{E}[R^2] = (k-1)/(T-1)$ from, for example, (I.7.12). Its variance could similarly be stated. Recall that its distribution was derived under the null hypothesis that the $k-1$ regression coefficients are zero. This implies that R^2 is upward biased, and also shows that just adding superfluous regressors will always increase the expected value of R^2. As such, choosing a set of regressors such that R^2 is maximized is not appropriate for model selection.

However, the so-called **adjusted** R^2 can be used. It is defined as

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{T-1}{T-k}. \tag{1.17}$$

Virtually all statistical software for regression will include this measure. Less well known is that it has (like so many things) its origin with Ronald Fisher; see Fisher (1925). Notice how, like the Akaike information criterion (hereafter AIC) and other penalty-based measures applied to the obtained log likelihood, when k is increased, the increase in R^2 is offset by a factor involving k in R^2_{adj}.

Measure (1.17) can be motivated in (at least) two ways. First, note that, under the null hypothesis,

$$\mathbb{E}[R^2_{\text{adj}}] = 1 - \left(1 - \frac{k-1}{T-1}\right)\frac{T-1}{T-k} = 0,$$

providing a perfect offset to R^2's expected value simply increasing in k under the null. A second way is to note that, while $R^2 = 1 - \text{RSS}/\text{TSS}$ from (1.13),

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(T-k)}{\text{TSS}/(T-1)} = 1 - \frac{\widehat{\mathbb{V}}(\widehat{\boldsymbol{\epsilon}})}{\widehat{\mathbb{V}}(\mathbf{Y})},$$

the numerator and denominator being unbiased estimators of their respective variances, recalling (1.11). The use of R^2_{adj} for model selection is very similar to use of other measures, such as the (corrected) AIC and the so-called **Mallows'** C_k; see, e.g., Seber and Lee (2003, Ch. 12) for a very good discussion of these, and other criteria, and the relationships among them.

Section 1.2.3 extends the model to the case in which $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ from (1.3), but $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a known, positive definite variance–covariance matrix. There, an appropriate expression for R^2 will be derived that generalizes (1.13). For now, the reader is encouraged to express R^2 in (1.13) as a ratio of quadratic forms, assuming $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, and compute and plot its density for a given $\mathbf{X}$ and $\boldsymbol{\Sigma}$, such as given in (1.31) for a given value of parameter a, as done in, e.g., Carrodus and Giles (1992). When $a = 0$, the density should coincide with that given by (1.16).

We end this section with an important remark, and an important example.

Remark It is often assumed that the elements of $\mathbf{X}$ are known constants. This is quite plausible in designed experiments, where $\mathbf{X}$ is chosen in such a way as to maximize the ability of the experiment to answer the questions of interest. In this case, $\mathbf{X}$ is often referred to as the **design matrix**. This will rarely hold in applications in the social sciences, where the $\mathbf{x}_t'$ reflect certain measurements and are better described as being observations of random variables from the multivariate distribution describing both $\mathbf{x}_t'$ and Y_t. Fortunately, under certain assumptions, one may ignore this issue and proceed as if $\mathbf{x}_t'$ were fixed constants and not realizations of a random variable.

Assume matrix $\mathbf{X}$ is no longer deterministic. Denote by $\mathbf{X}$ an outcome of random variable $\mathcal{X}$, with kT-variate probability density function (hereafter p.d.f.) $f_{\mathcal{X}}(\mathbf{X}\,;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector. We require the following assumption:

0. The conditional distribution $\mathbf{Y} \mid (\mathcal{X} = \mathbf{X})$ depends only on $\mathbf{X}$ and parameters $\boldsymbol{\beta}$ and σ and such that $\mathbf{Y} \mid (\mathcal{X} = \mathbf{X})$ has mean $\mathbf{X}\boldsymbol{\beta}$ and finite variance $\sigma^2\mathbf{I}$.

For example, we could have $\mathbf{Y} \mid (\mathcal{X} = \mathbf{X}) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Under the stated assumption, the joint density of $\mathbf{Y}$ and $\mathcal{X}$ can be written as

$$f_{\mathbf{Y},\mathcal{X}}(\mathbf{y}, \mathbf{X} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = f_{\mathbf{Y}|\mathcal{X}}(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\beta}, \sigma^2) \cdot f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}). \tag{1.18}$$

Now consider the following two additional assumptions:

1) The distribution of $\mathcal{X}$ does not depend on $\boldsymbol{\beta}$ or σ^2, so we can write $f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\theta})$.
2) The parameter space of $\boldsymbol{\theta}$ and that of $(\boldsymbol{\beta}, \sigma^2)$ are not related, that is, they are not restricted by one another in any way.

Then, with regard to $\boldsymbol{\beta}$ and σ^2, $f_{\mathcal{X}}$ is only a multiplicative constant and the log-likelihood corresponding to (1.18) is the same as (1.10) plus the additional term $\log f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\theta})$. As this term does not involve $\boldsymbol{\beta}$ or σ^2, the (generalized) least squares estimator still coincides with the m.l.e. When the above assumptions are satisfied, $\boldsymbol{\theta}$ and $(\boldsymbol{\beta}, \sigma^2)$ are said to be **functionally independent** (Graybill, 1976, p. 380), or **variation-free** (Poirier, 1995, p. 461). More common in the econometrics literature is to say that one assumes $\mathbf{X}$ to be **(weakly) exogenous** with respect to $\mathbf{Y}$.

The extent to which these assumptions are reasonable is open to debate. Clearly, without them, estimation of $\boldsymbol{\beta}$ and σ^2 is not so straightforward, as then $f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ must be (fully, or at least partially) specified. If they hold, then

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}_{\mathcal{X}}[\mathbb{E}[\widehat{\boldsymbol{\beta}} \mid \mathcal{X} = \mathbf{X}]] = \mathbb{E}_{\mathcal{X}}[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\epsilon} \mid \mathcal{X}]] = \mathbb{E}_{\mathcal{X}}[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

and

$$\mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) = \mathbb{E}_{\mathcal{X}}[\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \mathcal{X} = \mathbf{X}, \sigma^2]] = \sigma^2\mathbb{E}_{\mathcal{X}}[(\mathcal{X}'\mathcal{X})^{-1}],$$

the latter being obtainable only when $f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\theta})$ is known.

A discussion of the implications of falsely assuming that $\mathbf{X}$ is not stochastic is provided by Binkley and Abbott (1987).[3] ■

Example 1.1 ***Frisch–Waugh–Lovell Theorem***
It is occasionally useful to express the o.l.s. estimator of each component of the partitioned vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1$ is $k_1 \times 1, 1 \leqslant k_1 < k$. With the appropriate corresponding partition of $\mathbf{X}$, model (1.3) is then expressed as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

The normal equations (1.6) then read

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

or

$$\mathbf{X}_1'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{Y} \qquad \text{and} \qquad \mathbf{X}_2'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{Y}, \tag{1.19}$$

3 We use the tombstone, QED, or halmos, symbol ■ to denote the end of proofs of theorems, as well as examples and remarks, acknowledging that it is traditionally only used for the former, as popularized by Paul Halmos.

so that

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{Y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2) \tag{1.20}$$

and $\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1)$. To obtain an expression for $\widehat{\boldsymbol{\beta}}_2$ that does not depend on $\widehat{\boldsymbol{\beta}}_1$, let $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$, premultiply (1.20) by $\mathbf{X}_1$, and substitute $\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1$ into the second equation in (1.19) to get

$$\mathbf{X}_2'(\mathbf{I} - \mathbf{M}_1)(\mathbf{Y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2) + \mathbf{X}_2'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{Y},$$

or, expanding and solving for $\widehat{\boldsymbol{\beta}}_2$,

$$\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}. \tag{1.21}$$

A similar argument (or via symmetry) shows that

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}, \tag{1.22}$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$.

An important special case of (1.21) discussed further in Chapter 4 is when $k_1 = k - 1$, so that $\mathbf{X}_2$ is $T \times 1$ and $\widehat{\boldsymbol{\beta}}_2$ in (1.21) reduces to the scalar

$$\widehat{\beta}_2 = \frac{\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}}{\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2}. \tag{1.23}$$

This is a ratio of a bilinear form to a quadratic form, as discussed in Appendix A.

The Frisch–Waugh–Lovell theorem has both computational value (see, e.g., Ruud, 2000, p. 66, and Example 1.9 below) and theoretical value; see Ruud (2000), Davidson and MacKinnon (2004), and also Section 5.2. Extensions of the theorem are considered in Fiebig et al. (1996). ■

1.2.3 Generalized Least Squares

Now consider the more general assumption that $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a known, positive definite variance–covariance matrix. The density of $\mathbf{Y}$ is now given by

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-T/2}|\sigma^2\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{1.24}$$

and one could use calculus to find the m.l.e. of $\boldsymbol{\beta}$. Alternatively, we could transform the model in such a way that the above results still apply. In particular, with $\boldsymbol{\Sigma}^{-1/2}$ the symmetric matrix such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$, premultiply (1.3) by $\boldsymbol{\Sigma}^{-1/2}$ so that

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{Y} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}, \qquad \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon} \sim \mathrm{N}_T(\mathbf{0}, \sigma^2\mathbf{I}). \tag{1.25}$$

Then, using the previous maximum likelihood approach as in (1.10), with

$$\mathbf{Y}_* := \boldsymbol{\Sigma}^{-1/2}\mathbf{Y} \quad \text{and} \quad \mathbf{X}_* := \boldsymbol{\Sigma}^{-1/2}\mathbf{X} \tag{1.26}$$

in place of $\mathbf{Y}$ and $\mathbf{X}$ implies the normal equations

$$(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \tag{1.27}$$

that generalize (1.6), and

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{Y}_* = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}, \tag{1.28}$$

where the notation $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ is used to indicate its dependence on knowledge of $\boldsymbol{\Sigma}$. This is known as the **generalized least squares** (g.l.s.) estimator, with variance given by

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} \mid \sigma^2) = \sigma^2(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}. \tag{1.29}$$

It is attributed to A. C. Aitken from 1934. Of course, σ^2 is unknown. The usual estimator of $(T-k)\sigma^2$ is given by

$$\mathsf{S}(\boldsymbol{\beta}; \mathbf{Y}_*, \mathbf{X}_*) = (\mathbf{Y}_* - \mathbf{X}_*\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}})'(\mathbf{Y}_* - \mathbf{X}_*\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}). \tag{1.30}$$

Example 1.2 Let $\epsilon_t \overset{\text{ind}}{\sim} \mathrm{N}(0, \sigma^2 k_t)$, where the k_t are known, positive constants, so that $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(k_1^{-1}, \ldots, k_T^{-1})$. Then $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ is referred to as the **weighted least squares** estimator. If in the Hamburg income example above, we take $k_t = x_t$, then observations $\{y_t, x_t\}$ receive weights proportional to x_t^{-1}. This has the effect of down-weighting observations with high ages, for which the uncertainty of the slope parameter is higher, and vice versa. ■

Example 1.3 Let the model be given by $Y_t = \mu + \epsilon_t$, $t = 1, \ldots, T$. With $\mathbf{X} = \mathbf{1}$, we have

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = [T^{-1}, \ldots, T^{-1}],$$

and the o.l.s. estimator of μ is just the simple average of the observations, $\bar{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Assume, however, that the ϵ_t are not i.i.d., but are given by the recursion $\epsilon_t = a\epsilon_{t-1} + U_t$, $|a| < 1$, and $U_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. This is referred to as a *stationary first order autoregressive model*, abbreviated AR(1), and is the subject of Chapter 4. There, the covariance matrix of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_T)'$ is shown to be $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{\Sigma}$ with

$$\boldsymbol{\Sigma} = \frac{1}{1-a^2}\begin{bmatrix} 1 & a & a^2 & \cdots & a^{T-1} \\ a & 1 & a & \cdots & a^{T-2} \\ a^2 & a & 1 & \cdots & a^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a^{T-1} & a^{T-2} & a^{T-3} & \cdots & 1 \end{bmatrix}. \tag{1.31}$$

The g.l.s. estimator of μ is now a weighted average of the Y_t, where the weight vector is given by $\mathbf{w} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$. Straightforward calculation shows that, for $a = 0.5$, $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} = 4/(T+2)$ and

$$\mathbf{X}'\boldsymbol{\Sigma}^{-1} = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \ldots, \frac{1}{4}, \frac{1}{2}\right]',$$

so that the first and last weights are $2/(T+2)$ and the middle $T-2$ are all $1/(T+2)$. Note that the weights sum to one. A similar pattern holds for all $|a| < 1$, with the ratio of the first and last weights to the center weights converging to $1/2$ as $a \to -1$ and to ∞ as $a \to 1$. Thus, we see that (i) for constant T, the difference between g.l.s. and o.l.s. grows as $a \to 1$ and (ii) for constant a, $|a| < 1$, the difference between g.l.s. and o.l.s. shrinks as $T \to \infty$. The latter is true because a finite number of observations, in this case only two, become negligible in the limit, and because the relative weights associated with these two values converges to a constant independent of T.

Now consider the model $Y_t = \mu + \epsilon_t$, $t = 1, \ldots, T$, with $\epsilon_t = bU_{t-1} + U_t$, $|b| < 1$, $U_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. This is referred to as an invertible *first-order moving average model*, or MA(1), and is discussed in

detail in Chapter 6. There, it is shown that $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{\Sigma}$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1+b^2 & b & 0 & \cdots & 0 \\ b & 1+b^2 & \ddots & & \vdots \\ 0 & b & \ddots & & 0 \\ \vdots & 0 & \ddots & & b \\ 0 & \cdots & 0 & b & 1+b^2 \end{bmatrix}.$$

The weight vectors $\mathbf{w} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$ for the two values, $b = -0.9$ and $b = 0.9$, are plotted in Figure 1.2 for $T = 100$. This is clearly quite a different weighting structure than for the AR(1) model.

In the limiting case $b \to 1$, we have

$$Y_1 = \mu + U_0 + U_1, \quad Y_2 = \mu + U_1 + U_2, \quad \ldots, \quad Y_T = \mu + U_{T-1} + U_T$$

so that

$$\sum_{t=1}^{T} Y_t = T\mu + U_0 + U_T + 2\sum_{t=1}^{T-1} U_t,$$

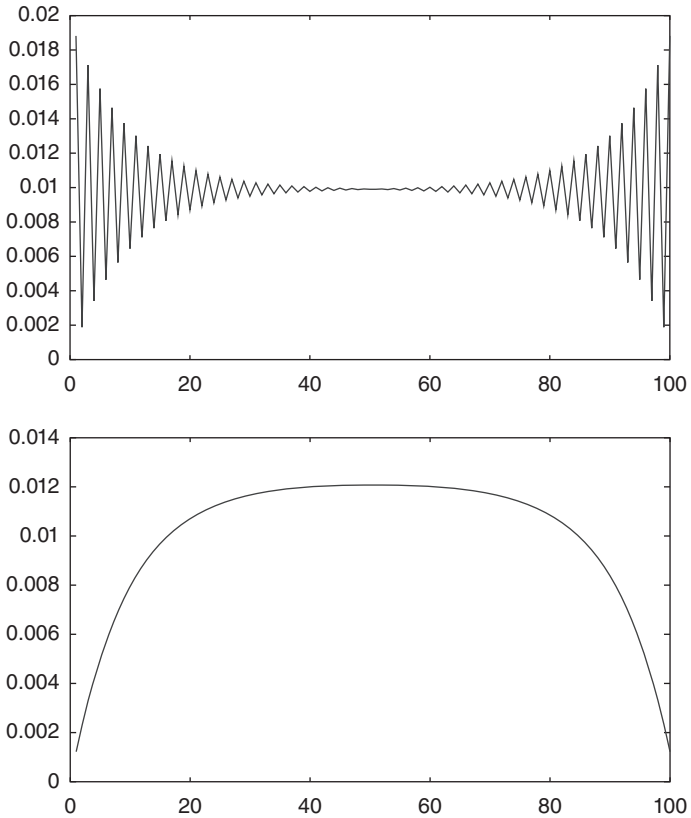

Figure 1.2 Weight vector for an MA(1) model with $T = 100$ and $b = 0.9$ (top) and $b = -0.9$ (bottom).

$\mathbb{E}[\bar{Y}] = \mu$ and

$$\mathbb{V}(\bar{Y}) = \frac{\sigma^2 + \sigma^2 + 4(T-1)\sigma^2}{T^2} = \frac{4\sigma^2}{T} - \frac{2\sigma^2}{T^2}.$$

For $T = 100$ and $\sigma^2 = 1$, $\mathbb{V}(\bar{Y} \mid b = 1) \approx 0.0398$. Similarly, for $b = -1$, $\sum_{t=1}^{T} Y_t = T\mu + U_0 + U_T$ and $\mathbb{V}(\bar{Y} \mid b = -1) = 2\sigma^2/T^2 = 0.0002$. ■

Consideration of the previous example might lead one to ponder if it is possible to specify conditions such that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ will equal $\widehat{\boldsymbol{\beta}}_{\mathbf{I}} = \widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\Sigma} \neq \mathbf{I}$. A necessary and sufficient condition for $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\beta}}$ is if the k columns of $\mathbf{X}$ are linear combinations of k of the eigenvectors of $\boldsymbol{\Sigma}$, as first established by Anderson (1948); see, e.g., Anderson (1971, p. 19 and p. 561) for proof.

This question has generated a large amount of academic work, as illustrated in the survey of Puntanen and Styan (1989), which contains about 90 references (see also Krämer et al., 1996). There are several equivalent conditions for the result to hold, a rather useful and attractive one of which is that

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\beta}} \text{ if and only if } \mathbf{P}\boldsymbol{\Sigma} \text{ is symmetric,} \tag{1.32}$$

i.e., if and only if $\mathbf{P}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{P}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Another is that there exists a matrix $\mathbf{F}$ satisfying $\mathbf{XF} = \boldsymbol{\Sigma}^{-1}\mathbf{X}$, which is demonstrated in Example 1.5.

Example 1.4 With $\mathbf{X} = \mathbf{1}$ (a T-length column of ones), Anderson's condition implies that $\mathbf{1}$ needs to be an eigenvector of $\boldsymbol{\Sigma}$, or $\boldsymbol{\Sigma}1 = s\mathbf{1}$ for some nonzero scalar s. This means that the sum of each row of $\boldsymbol{\Sigma}$ must be the same value. This obviously holds when $\boldsymbol{\Sigma} = \mathbf{I}$, and clearly never holds when $\boldsymbol{\Sigma}$ is a diagonal weighting matrix with at least two weights differing.

To determine if $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\beta}}$ is possible for the AR(1) and MA(1) models from Example 1.3, we use a result of McElroy (1967), who showed that, if $\mathbf{X}$ is full rank and contains $\mathbf{1}$, then $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\beta}}$ if and only if $\boldsymbol{\Sigma}$ is full rank and can be expressed as $k_1\mathbf{I} + k_2\mathbf{1}\mathbf{1}'$, i.e., the equicorrelated case. We will see in Chapters 4 and 7 that this is never the case for AR(1) and MA(1) models or, more generally, for stationary and invertible ARMA(p, q) models. ■

Remark The previous discussion begets the question of how one could assess the extent to which o.l.s. will be inferior relative to g.l.s., notably because, in many applications, $\boldsymbol{\Sigma}$ will not be known. This turns out to be a complicated endeavor in general; see Puntanen and Styan (1989, p. 154) and the references therein for further details. Observe also how (1.28) and (1.29) assume the true $\boldsymbol{\Sigma}$. The determination of robust estimators for the variance of $\widehat{\boldsymbol{\beta}}$ for unknown $\boldsymbol{\Sigma}$ is an important and active research area in statistics and, particularly, econometrics (and for other model classes beyond the simple linear regression model studied here). The primary reference papers are White (1980, 1982), MacKinnon and White (1985), Newey and West (1987), and Andrews (1991), giving rise to the class of so-called **heteroskedastic and autocorrelation consistent** covariance matrix estimators, or HAC. With respect to computation of the HAC estimators, see Zeileis (2006), Heberle and Sattarhoff (2017), and the references therein. ■

It might come as a surprise that defining the coefficient of multiple determination R^2 in the g.l.s. context is not so trivial, and several suggestions exist. The problem stems from the definition in the o.l.s. case (1.13), with $R^2 = 1 - \mathsf{S}(\widehat{\boldsymbol{\beta}}, \mathbf{Y}, \mathbf{X})/\mathsf{S}(\bar{Y}, \mathbf{Y}, \mathbf{1})$, and observing that, if $\mathbf{1} \in C(\mathbf{X})$ (the column space of $\mathbf{X}$, as defined below), then, via the transformation in (1.26), $\mathbf{1} \notin C(\mathbf{X}_*)$.

To establish a meaningful definition, we first need the fact that, with $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ and $\widehat{\epsilon} = \mathbf{Y} - \widehat{\mathbf{Y}}$,

$$\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} = \widehat{\mathbf{Y}}'\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{Y}} + \widehat{\epsilon}'\boldsymbol{\Sigma}^{-1}\widehat{\epsilon}, \tag{1.33}$$

which is derived in (1.47). Next, from the normal equations (1.27) and letting $\mathbf{X}_i$ denote the ith column of $\mathbf{X}$, $i = 1, \ldots, k$, we have a system of k equations, the ith of which is, with $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$,

$$(\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1)\widehat{\beta}_1 + (\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_2)\widehat{\beta}_2 + \cdots + (\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_k)\widehat{\beta}_k = \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Y}.$$

Similarly, premultiplying both sides of $\mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \widehat{\mathbf{Y}}$ by $\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}$ gives

$$(\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_1)\widehat{\beta}_1 + (\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_2)\widehat{\beta}_2 + \cdots + (\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_k)\widehat{\beta}_k = \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{Y}},$$

so that

$$\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \widehat{\mathbf{Y}}) = 0,$$

which we will see again below, in the context of projection, in (1.63). In particular, with $\mathbf{X}_1 = \mathbf{1} = (1, 1, \ldots, 1)'$ the usual first regressor, $\mathbf{1}'\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{Y}} = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$. We now follow Buse (1973), and define the weighted mean to be

$$\bar{Y} := \bar{Y}_{\boldsymbol{\Sigma}} := \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} \quad \left(= \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{Y}}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}\right), \tag{1.34}$$

which obviously reduces to the simple sample mean when $\boldsymbol{\Sigma} = \mathbf{I}$. The next step is to confirm by simply multiplying out that

$$(\mathbf{Y} - \bar{Y}\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \bar{Y}\mathbf{1}) = \mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \frac{(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{Y})^2}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}},$$

and, likewise,

$$(\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}) = \widehat{\mathbf{Y}}'\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{Y}} - \frac{(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{Y})^2}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}},$$

so that (1.33) can be expressed as

$$(\mathbf{Y} - \bar{Y}\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \bar{Y}\mathbf{1}) = (\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}) + \widehat{\epsilon}'\boldsymbol{\Sigma}^{-1}\widehat{\epsilon}. \tag{1.35}$$

The definition of R^2 is now given by

$$R^2 = R^2_{\boldsymbol{\Sigma}} = 1 - \frac{\widehat{\epsilon}'\boldsymbol{\Sigma}^{-1}\widehat{\epsilon}}{(\mathbf{Y} - \bar{Y}\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \bar{Y}\mathbf{1})}, \tag{1.36}$$

which is indeed analogous to (1.13) and reduces to it when $\boldsymbol{\Sigma} = \mathbf{I}$.

Along with examples of other, less desirable, definitions, Buse (1973) discusses the benefits of this definition, which include that it is interpretable as the proportion of the generalized sum of squares of the dependent variable that is attributable to the influence of the explanatory variables, and that it lies between zero and one. It is also zero when all the estimates coefficients (except the constant) are zero, and can be related to the F test as was done above in the ordinary least squares case.

1.3 The Geometric Approach to Least Squares

> In spite of earnest prayer and the greatest desire to adhere to proper statistical behavior, I have not been able to say why the method of maximum likelihood is to be preferred over other methods, particularly the method of least squares.
>
> (Joseph Berkson, 1944, p. 359)

The following sections analyze the linear regression model using the notion of projection. This complements the purely algebraic approach to regression analysis by providing a useful terminology and geometric intuition behind least squares. Most importantly, its use often simplifies the derivation and understanding of various quantities such as point estimators and test statistics. The reader is assumed to be comfortable with the notions of linear subspaces, span, dimension, rank, and orthogonality. See the references given at the beginning of Section B.5 for detailed presentations of these and other important topics associated with linear and matrix algebra.

1.3.1 Projection

The Euclidean **dot product** or **inner product** of two vectors $\mathbf{u} = (u_1, u_2, \dots, u_T)'$ and $\mathbf{v} = (v_1, v_2, \dots, v_T)'$ is denoted by $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}'\mathbf{v} = \sum_{i=1}^{T} u_i v_i$. Observe that, for $\mathbf{y}, \mathbf{u}, \mathbf{w} \in \mathbb{R}^T$,

$$\langle \mathbf{y} - \mathbf{u}, \mathbf{w} \rangle = (\mathbf{y} - \mathbf{u})'\mathbf{w} = \mathbf{y}'\mathbf{w} - \mathbf{u}'\mathbf{w} = \langle \mathbf{y}, \mathbf{w} \rangle - \langle \mathbf{u}, \mathbf{w} \rangle. \tag{1.37}$$

The **norm** of vector $\mathbf{u}$ is $\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}$. The square matrix $\mathbf{U}$ with columns $\mathbf{u}_1, \dots, \mathbf{u}_T$ is **orthonormal** if $\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}$, i.e., $\mathbf{U}' = \mathbf{U}^{-1}$, implying $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 1$ if $i = j$ and zero otherwise.

For a fixed $T \times k$ matrix $\mathbf{X}$, $k \leqslant T$ and usually such that $k \ll T$ ("is much less than"), the **column space** of $\mathbf{X}$, denoted $\mathcal{C}(\mathbf{X})$, or the **linear span** of the k columns $\mathbf{X}$, is the set of all vectors that can be generated as a linear sum of, or *spanned by*, the columns of $\mathbf{X}$, such that the coefficient of each vector is a real number, i.e.,

$$\mathcal{C}(\mathbf{X}) = \{\mathbf{y} : \mathbf{y} = \mathbf{X}\mathbf{b}, \mathbf{b} \in \mathbb{R}^k\}. \tag{1.38}$$

In words, if $\mathbf{y} \in \mathcal{C}(\mathbf{X})$, then there exists $\mathbf{b} \in \mathbb{R}^k$ such that $\mathbf{y} = \mathbf{X}\mathbf{b}$.

It is easy to verify that $\mathcal{C}(\mathbf{X})$ is a subspace of $\mathbb{R}^T$ with **dimension** $\dim(\mathcal{C}(\mathbf{X})) = \operatorname{rank}(\mathbf{X}) \leqslant k$. If $\dim(\mathcal{C}(\mathbf{X})) = k$, then $\mathbf{X}$ is said to be a **basis matrix** (for $\mathcal{C}(\mathbf{X})$). Furthermore, if the columns of $\mathbf{X}$ are orthonormal, then $\mathbf{X}$ is an **orthonormal basis matrix** and $\mathbf{X}'\mathbf{X} = \mathbf{I}$.

Let $\mathbf{V}$ be a basis matrix with columns $\mathbf{v}_1, \dots, \mathbf{v}_k$. The method of **Gram–Schmidt** can be used to construct an orthonormal basis matrix $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_k]$ as follows. First set $\mathbf{u}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$ so that $\langle \mathbf{u}_1, \mathbf{u}_1 \rangle = 1$. Next, let $\mathbf{u}_2^* = \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1$, so that

$$\langle \mathbf{u}_2^*, \mathbf{u}_1 \rangle = \langle \mathbf{v}_2, \mathbf{u}_1 \rangle - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \langle \mathbf{u}_1, \mathbf{u}_1 \rangle = \langle \mathbf{v}_2, \mathbf{u}_1 \rangle - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle = 0, \tag{1.39}$$

and set $\mathbf{u}_2 = \mathbf{u}_2^*/\|\mathbf{u}_2^*\|$. By construction of $\mathbf{u}_2$, $\langle \mathbf{u}_2, \mathbf{u}_2 \rangle = 1$, and from (1.39), $\langle \mathbf{u}_2, \mathbf{u}_1 \rangle = 0$. Continue with $\mathbf{u}_3^* = \mathbf{v}_3 - \langle \mathbf{v}_3, \mathbf{u}_1 \rangle \mathbf{u}_1 - \langle \mathbf{v}_3, \mathbf{u}_2 \rangle \mathbf{u}_2$ and $\mathbf{u}_3 = \mathbf{u}_3^*/\|\mathbf{u}_3^*\|$, up to $\mathbf{u}_k^* = \mathbf{v}_k - \sum_{i=1}^{k-1} \langle \mathbf{v}_k, \mathbf{u}_i \rangle \mathbf{u}_i$ and $\mathbf{u}_k = \mathbf{u}_k^*/\|\mathbf{u}_k^*\|$. This renders $\mathbf{U}$ an orthonormal basis matrix for $\mathcal{C}(\mathbf{V})$.

The next example offers some practice with column spaces, proves a simple result, and shows how to use Matlab to investigate a special case.

Example 1.5 Consider the equality of the generalized and ordinary least squares estimators. Let $\mathbf{X}$ be a $T \times k$ regressor matrix of full rank, $\mathbf{\Sigma}$ be a $T \times T$ positive definite covariance matrix, $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$, and $\mathbf{B} = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})$ (both symmetric and full rank). Then, for all T-length column vectors $\mathbf{Y} \in \mathbb{R}^T$,

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathbf{\Sigma}} &\Leftrightarrow (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &\Leftrightarrow \mathbf{B}^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Y} = \mathbf{A}\mathbf{X}'\mathbf{Y} \\ &\Leftrightarrow \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Y} = \mathbf{B}\mathbf{A}\mathbf{X}'\mathbf{Y} \Leftrightarrow \mathbf{Y}'(\mathbf{\Sigma}^{-1}\mathbf{X}) = \mathbf{Y}'(\mathbf{X}\mathbf{A}\mathbf{B}) \\ &\Leftrightarrow \mathbf{\Sigma}^{-1}\mathbf{X} = \mathbf{X}\mathbf{A}\mathbf{B}, \end{aligned} \tag{1.40}$$

where the $\Rightarrow$ in (1.40) follows because $\mathbf{Y}$ is arbitrary. (Recall from (1.32) that equality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\mathbf{\Sigma}}$ depends only on properties of $\mathbf{X}$ and $\mathbf{\Sigma}$. Another way of confirming the $\Rightarrow$ in (1.40) is to replace $\mathbf{Y}$ in $\mathbf{Y}'(\mathbf{\Sigma}^{-1}\mathbf{X}) = \mathbf{Y}'(\mathbf{X}\mathbf{A}\mathbf{B})$ with $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and take expectations.)

Thus, if $\mathbf{z} \in C(\mathbf{\Sigma}^{-1}\mathbf{X})$, then there exists a $\mathbf{v}$ such that $\mathbf{z} = \mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{v}$. But then (1.40) implies that

$$\mathbf{z} = \mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{v} = \mathbf{X}\mathbf{A}\mathbf{B}\mathbf{v} = \mathbf{X}\mathbf{w},$$

where $\mathbf{w} = \mathbf{A}\mathbf{B}\mathbf{v}$, i.e., $\mathbf{z} \in C(\mathbf{X})$. Thus, $C(\mathbf{\Sigma}^{-1}\mathbf{X}) \subset C(\mathbf{X})$. Similarly, if $\mathbf{z} \in C(\mathbf{X})$, then there exists a $\mathbf{v}$ such that $\mathbf{z} = \mathbf{X}\mathbf{v}$, and (1.40) implies that

$$\mathbf{z} = \mathbf{X}\mathbf{v} = \mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{v} = \mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{w},$$

where $\mathbf{w} = \mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{v}$, i.e., $C(\mathbf{X}) \subset C(\mathbf{\Sigma}^{-1}\mathbf{X})$. Thus, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathbf{\Sigma}} \Leftrightarrow C(\mathbf{X}) = C(\mathbf{\Sigma}^{-1}\mathbf{X})$. This column space equality implies that there exists a $k \times k$ full rank matrix $\mathbf{F}$ such that $\mathbf{X}\mathbf{F} = \mathbf{\Sigma}^{-1}\mathbf{X}$. To compute $\mathbf{F}$, left-multiply by $\mathbf{X}'$ and, as we assumed that $\mathbf{X}$ is full rank, we can then left-multiply by $(\mathbf{X}'\mathbf{X})^{-1}$, so that $\mathbf{F} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}$.[4]

As an example, with $\mathbf{J}_T$ the $T \times T$ matrix of ones, let $\mathbf{\Sigma} = \rho\sigma^2\mathbf{J}_T + (1-\rho)\sigma^2\mathbf{I}_T$, which yields the **equi-correlated** case. Then, experimenting with $\mathbf{X}$ in the code in Listing 1.1 allows one to numerically confirm that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathbf{\Sigma}}$ when $\mathbf{1}_T \in C(\mathbf{X})$, but not when $\mathbf{1}_T \notin C(\mathbf{X})$. The fifth line checks (1.40), while the last line checks the equality of $\mathbf{X}\mathbf{F}$ and $\mathbf{\Sigma}^{-1}\mathbf{X}$. It is also easy to add code to confirm that $\mathbf{P}\mathbf{\Sigma}$ is symmetric in this case, and not when $\mathbf{1}_T \notin C(\mathbf{X})$. ■

The **orthogonal complement** of $C(\mathbf{X})$, denoted $C(\mathbf{X})^{\perp}$, is the set of all vectors in $\mathbb{R}^T$ that are orthogonal to $C(\mathbf{X})$, i.e., the set $\{\mathbf{z} : \mathbf{z}'\mathbf{y} = 0,\ \mathbf{y} \in C(\mathbf{X})\}$. From (1.38), this set can be written as $\{\mathbf{z} : \mathbf{z}'\mathbf{X}\mathbf{b} =$

```
1 s2=2; T=10; rho=0.8; Sigma=s2*( rho*ones(T,T)+(1-rho)*eye(T));
2 zeroone=[zeros(4,1);ones(6,1)]; onezero=[ones(4,1);zeros(6,1)];
3 X=[zeroone, onezero, randn(T,5)];
4 Si=inv(Sigma); A=inv(X'*X); B=X'*Si*X;
5 shouldbezeros1 = Si*X - X*A*B
6 F=inv(X'*X)*X'*Si*X; % could also use: F=X\(Si*X);
7 shouldbezeros2 = X*F - Si*X
```

Program Listing 1.1: For confirming that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathbf{\Sigma}}$ when $\mathbf{1}_T \in C(\mathbf{X})$.

4 In Matlab, one can also use the `mldivide` operator for this calculation.

0, $\mathbf{b} \in \mathbb{R}^k\}$. Taking the transpose and observing that $\mathbf{z}'\mathbf{X}\mathbf{b}$ must equal zero for all $\mathbf{b} \in \mathbb{R}^k$, we may also write

$$C(\mathbf{X})^{\perp} = \{\mathbf{z} \in \mathbb{R}^T : \mathbf{X}'\mathbf{z} = \mathbf{0}\}.$$

Finally, the shorthand notation $\mathbf{z} \perp C(\mathbf{X})$ or $\mathbf{z} \perp \mathbf{X}$ will be used to indicate that $\mathbf{z} \in C(\mathbf{X})^{\perp}$.

The usefulness of the geometric approach to least squares rests on the following fundamental result from linear algebra.

Theorem 1.1 ***Projection Theorem*** Given a subspace $\mathcal{S}$ of $\mathbb{R}^T$, there exists a unique $\mathbf{u} \in \mathcal{S}$ and $\mathbf{v} \in \mathcal{S}^{\perp}$ for every $\mathbf{y} \in \mathbb{R}^T$ such that $\mathbf{y} = \mathbf{u} + \mathbf{v}$. The vector $\mathbf{u}$ is given by

$$\mathbf{u} = \langle \mathbf{y}, \mathbf{w}_1 \rangle \mathbf{w}_1 + \langle \mathbf{y}, \mathbf{w}_2 \rangle \mathbf{w}_2 + \cdots + \langle \mathbf{y}, \mathbf{w}_k \rangle \mathbf{w}_k, \tag{1.41}$$

where $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\}$ are a set of orthonormal $T \times 1$ vectors that span $\mathcal{S}$ and k is the dimension of $\mathcal{S}$. The vector $\mathbf{v}$ is given by $\mathbf{y} - \mathbf{u}$.

Proof: To show existence, note that, by construction, $\mathbf{u} \in \mathcal{S}$ and, from (1.37) for $i = 1, \ldots, k$,

$$\langle \mathbf{v}, \mathbf{w}_i \rangle = \langle \mathbf{y} - \mathbf{u}, \mathbf{w}_i \rangle = \langle \mathbf{y}, \mathbf{w}_i \rangle - \sum_{j=1}^{k} \langle \mathbf{y}, \mathbf{w}_j \rangle \cdot \langle \mathbf{w}_j, \mathbf{w}_i \rangle = 0,$$

so that $\mathbf{v} \perp \mathcal{S}$, as required.

To show that $\mathbf{u}$ and $\mathbf{v}$ are unique, suppose that $\mathbf{y}$ can be written as $\mathbf{y} = \mathbf{u}^* + \mathbf{v}^*$, with $\mathbf{u}^* \in \mathcal{S}$ and $\mathbf{v}^* \in \mathcal{S}^{\perp}$. It follows that $\mathbf{u}^* - \mathbf{u} = \mathbf{v} - \mathbf{v}^*$. But as the left-hand side is contained in $\mathcal{S}$ and the right-hand side in $\mathcal{S}^{\perp}$, both $\mathbf{u}^* - \mathbf{u}$ and $\mathbf{v} - \mathbf{v}^*$ must be contained in the intersection $\mathcal{S} \cap \mathcal{S}^{\perp} = \{0\}$, so that $\mathbf{u} = \mathbf{u}^*$ and $\mathbf{v} = \mathbf{v}^*$. ■

Let $\mathbf{T} = [\mathbf{w}_1 \ \mathbf{w}_2 \ \ldots \ \mathbf{w}_k]$, where the $\mathbf{w}_i$ are given in Theorem 1.1 above. From (1.41),

$$\mathbf{u} = [\mathbf{w}_1 \ \mathbf{w}_2 \ \ldots \ \mathbf{w}_k] \begin{bmatrix} \langle \mathbf{y}, \mathbf{w}_1 \rangle \\ \langle \mathbf{y}, \mathbf{w}_2 \rangle \\ \vdots \\ \langle \mathbf{y}, \mathbf{w}_k \rangle \end{bmatrix} = \mathbf{T} \begin{bmatrix} \mathbf{w}_1' \\ \mathbf{w}_2' \\ \vdots \\ \mathbf{w}_k' \end{bmatrix} \mathbf{y} = \mathbf{T}\mathbf{T}'\mathbf{y} = \mathbf{P}_{\mathcal{S}}\mathbf{y}, \tag{1.42}$$

where the matrix $\mathbf{P}_{\mathcal{S}} = \mathbf{T}\mathbf{T}'$ is referred to as the **projection matrix onto** $\mathcal{S}$. Note that $\mathbf{T}'\mathbf{T} = \mathbf{I}$. Matrix $\mathbf{P}_{\mathcal{S}}$ is unique, so that the choice of orthonormal basis is not important; see Problem 1.4. We can write the decomposition of $\mathbf{y}$ as the (algebraically obvious) identity $\mathbf{y} = \mathbf{P}_{\mathcal{S}}\mathbf{y} + (\mathbf{I}_T - \mathbf{P}_{\mathcal{S}})\mathbf{y}$. Observe that $(\mathbf{I}_T - \mathbf{P}_{\mathcal{S}})$ is itself a projection matrix onto $\mathcal{S}^{\perp}$. By construction,

$$\mathbf{P}_{\mathcal{S}}\mathbf{y} \in \mathcal{S}, \tag{1.43}$$

$$(\mathbf{I}_T - \mathbf{P}_{\mathcal{S}})\mathbf{y} \in \mathcal{S}^{\perp}. \tag{1.44}$$

This is, in fact, the definition of a projection matrix, i.e., the matrix that satisfies both (1.43) and (1.44) for a given $\mathcal{S}$ and for all $\mathbf{y} \in \mathbb{R}^T$ is the projection matrix onto $\mathcal{S}$.

From Theorem 1.1, if $\mathbf{X}$ is a $T \times k$ basis matrix, then $\text{rank}(\mathbf{P}_{C(\mathbf{X})}) = k$. This also follows from (1.42), as $\text{rank}(\mathbf{T}\mathbf{T}') = \text{rank}(\mathbf{T}) = k$, where the first equality follows from the more general result that $\text{rank}(\mathbf{K}\mathbf{B}\mathbf{B}') = \text{rank}(\mathbf{K}\mathbf{B})$ for any $n \times m$ matrix $\mathbf{B}$ and $s \times n$ matrix $\mathbf{K}$ (see, e.g., Harville, 1997, Cor. 7.4.4, p. 75).

Observe that, if $\mathbf{u} = \mathbf{P}_{\mathcal{S}}\mathbf{y}$, then $\mathbf{P}_{\mathcal{S}}\mathbf{u}$ must be equal to $\mathbf{u}$ because $\mathbf{u}$ is already in $\mathcal{S}$. This also follows algebraically from (1.42), i.e., $\mathbf{P}_{\mathcal{S}} = \mathbf{T}\mathbf{T}'$ and $\mathbf{P}_{\mathcal{S}}{}^2 = \mathbf{T}\mathbf{T}'\mathbf{T}\mathbf{T}' = \mathbf{T}\mathbf{T}' = \mathbf{P}_{\mathcal{S}}$, showing that the matrix $\mathbf{P}_{\mathcal{S}}$is **idempotent**, i.e., $\mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}} = \mathbf{P}_{\mathcal{S}}$. Therefore, if $\mathbf{w} = (\mathbf{I}_T - \mathbf{P}_{\mathcal{S}})\mathbf{y} \in \mathcal{S}^{\perp}$, then $\mathbf{P}_{\mathcal{S}}\mathbf{w} = \mathbf{P}_{\mathcal{S}}(\mathbf{I}_T - \mathbf{P}_{\mathcal{S}})\mathbf{y} = \mathbf{0}$. Another property of projection matrices is that they are symmetric, which follows directly from $\mathbf{P}_{\mathcal{S}} = \mathbf{T}\mathbf{T}'$.

Example 1.6 Let $\mathbf{y}$ be a vector in $\mathbb{R}^T$ and $\mathcal{S}$ a subspace of $\mathbb{R}^T$ with corresponding projection matrix $\mathbf{P}_{\mathcal{S}}$. Then, with $\mathbf{P}_{\mathcal{S}^{\perp}} = \mathbf{I}_T - \mathbf{P}_{\mathcal{S}}$ from (1.44),

$$\begin{aligned}\|\mathbf{P}_{\mathcal{S}\perp}\mathbf{y}\|^2 &= \|\mathbf{y} - \mathbf{P}_{\mathcal{S}}\mathbf{y}\|^2 = (\mathbf{y} - \mathbf{P}_{\mathcal{S}}\mathbf{y})'(\mathbf{y} - \mathbf{P}_{\mathcal{S}}\mathbf{y}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathcal{S}}\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathcal{S}}'\mathbf{y} + \mathbf{y}'\mathbf{P}_{\mathcal{S}}'\mathbf{P}_{\mathcal{S}}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathcal{S}}\mathbf{y} = \|\mathbf{y}\|^2 - \|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|^2,\end{aligned}$$

i.e.,

$$\|\mathbf{y}\|^2 = \|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|^2 + \|\mathbf{P}_{\mathcal{S}^{\perp}}\mathbf{y}\|^2. \tag{1.45}$$

For $\mathbf{X}$ a full-rank $T \times k$ matrix and $\mathcal{S} = C(\mathbf{X})$, this implies, for regression model (1.3) with $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned}\mathbf{Y}'\mathbf{Y} &= \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} \\ &= (\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}})'(\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}}).\end{aligned} \tag{1.46}$$

In the g.l.s. framework, use of (1.46) applied to the transformed model (1.25) and (1.26) yields, with $\widehat{\mathbf{Y}}_* = \mathbf{X}_*\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\epsilon}}_* = \mathbf{Y}_* - \widehat{\mathbf{Y}}_*$,

$$\mathbf{Y}_*'\mathbf{Y}_* = \widehat{\mathbf{Y}}_*'\widehat{\mathbf{Y}}_* + \widehat{\boldsymbol{\epsilon}}_*'\widehat{\boldsymbol{\epsilon}}_* = (\widehat{\mathbf{Y}}_* + \widehat{\boldsymbol{\epsilon}}_*)'(\widehat{\mathbf{Y}}_* + \widehat{\boldsymbol{\epsilon}}_*),$$

or, with $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$,

$$\begin{aligned}\mathbf{Y}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{Y} &= \mathbf{Y}_*'\mathbf{Y}_* \\ &= (\widehat{\mathbf{Y}}_* + \widehat{\boldsymbol{\epsilon}}_*)'(\widehat{\mathbf{Y}}_* + \widehat{\boldsymbol{\epsilon}}_*) = (\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}})'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}(\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}}),\end{aligned}$$

or, finally,

$$\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} = \widehat{\mathbf{Y}}'\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}}'\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\epsilon}}, \tag{1.47}$$

which is (1.33), as was used for determining the R^2 measure in the g.l.s. case. ■

An equivalent definition of a projection matrix $\mathbf{P}$ onto $\mathcal{S}$ is when the following are satisfied:

$$\mathbf{v} \in \mathcal{S} \Rightarrow \mathbf{P}\mathbf{v} = \mathbf{v} \quad \text{(projection)} \tag{1.48}$$

$$\mathbf{w} \perp \mathcal{S} \Rightarrow \mathbf{P}\mathbf{w} = \mathbf{0} \quad \text{(perpendicularity)}. \tag{1.49}$$

The following result is both interesting and useful; it is proven in Problem 1.8, where further comments are given.

Theorem 1.2 If $\mathbf{P}$ is symmetric and idempotent with $\text{rank}(\mathbf{P}) = k$, then (i) k of the eigenvalues of $\mathbf{P}$ are unity and the remaining $T - k$ are zero, and (ii) $\text{tr}(\mathbf{P}) = k$.

This is understood as follows: If $T \times T$ matrix $\mathbf{P}$ is such that $\text{rank}(\mathbf{P}) = \text{tr}(\mathbf{P}) = k$ and k of the eigenvalues of $\mathbf{P}$ are unity and the remaining $T - k$ are zero, then it is not necessarily the case that $\mathbf{P}$ is symmetric and idempotent. However, if $\mathbf{P}$ is symmetric and idempotent, then $\text{tr}(\mathbf{P}) = k \iff \text{rank}(\mathbf{P}) = k$.

```
function G=makeG(X)     % G is such that M=G'G and I=GG'
k=size(X,2);            % could also use k = rank(X).
M=makeM(X);             % M=eye(T)-X*inv(X'*X)*X', where X is size TXk
[V,D]=eig(0.5*(M+M')); % V are eigenvectors, D eigenvalues
e=diag(D);
[e,I]=sort(e);          % I is a permutation index of the sorting
G=V(:,I(k+1:end)); G=G';
```

Program Listing 1.2: Computes matrix **G** in Theorem 1.3. Function `makeM` is given in Listing B.2.

Let $\mathbf{M} = \mathbf{I}_T - \mathbf{P}_S$ with $\dim(S) = k$, $k \in \{1, 2, \ldots, T-1\}$. As **M** is itself a projection matrix, then, similar to (1.42), it can be expressed as $\mathbf{VV}'$, where **V** is a $T \times (T-k)$ matrix with orthonormal columns. We state this obvious, but important, result as a theorem because it will be useful elsewhere (and it is slightly more convenient to use $\mathbf{V}'\mathbf{V}$ instead of $\mathbf{VV}'$).

Theorem 1.3 Let **X** be a full-rank $T \times k$ matrix, $k \in \{1, 2, \ldots, T-1\}$, and $S = \mathcal{C}(\mathbf{X})$ with $\dim(S) = k$. Let $\mathbf{M} = \mathbf{I}_T - \mathbf{P}_S$. The projection matrix **M** may be written as $\mathbf{M} = \mathbf{G}'\mathbf{G}$, where **G** is $(T-k) \times T$ and such that $\mathbf{GG}' = \mathbf{I}_{T-k}$ and $\mathbf{GX} = \mathbf{0}$.

A less direct, but instructive, method for proving Theorem 1.3 is given in Problem 1.5. Matrix **G** can be computed by taking its rows to be the $T-k$ eigenvectors of **M** that correspond to the unit eigenvalues. The small program in Listing 1.2 performs this computation. Alternatively, **G** can be computed by applying Gram–Schmidt orthogonalization to the columns of **M** and keeping the nonzero vectors.[5] Matrix **G** is not unique and the two methods just stated often result in different values.

It turns out that any symmetric, idempotent matrix is a projection matrix:

Theorem 1.4 The symmetry and idempotency of a matrix **P** are necessary and sufficient conditions for it to be the projection matrix onto the space spanned by its columns.

Proof: Sufficiency: We assume **P** is a symmetric and idempotent $T \times T$ matrix, and must show that (1.43) and (1.44) are satisfied for all $\mathbf{y} \in \mathbb{R}^T$. Let **y** be an element of $\mathbb{R}^T$ and let $S = \mathcal{C}(\mathbf{P})$. By the definition of column space, $\mathbf{Py} \in S$, which is (1.43). To see that (1.44) is satisfied, we must show that $(\mathbf{I}-\mathbf{P})\mathbf{y}$ is perpendicular to every vector in S, or that $(\mathbf{I}-\mathbf{P})\mathbf{y} \perp \mathbf{Pw}$ for all $\mathbf{w} \in \mathbb{R}^T$. But

$$((\mathbf{I}-\mathbf{P})\mathbf{y})'\mathbf{Pw} = \mathbf{y}'\mathbf{Pw} - \mathbf{y}'\mathbf{P}'\mathbf{Pw} = \mathbf{0}$$

because, by assumption, $\mathbf{P}'\mathbf{P} = \mathbf{P}$.

For necessity, following Christensen (1987, p. 335), write $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$, where $\mathbf{y} \in \mathbb{R}^T$, $\mathbf{y}_1 \in S$ and $\mathbf{y}_2 \in S^\perp$. Then, using only (1.48) and (1.49), $\mathbf{Py} = \mathbf{Py}_1 + \mathbf{Py}_2 = \mathbf{Py}_1 = \mathbf{y}_1$ and

$$\mathbf{P}^2\mathbf{y} = \mathbf{P}^2\mathbf{y}_1 + \mathbf{P}^2\mathbf{y}_2 = \mathbf{Py}_1 = \mathbf{Py},$$

so that **P** is idempotent. Next, as $\mathbf{Py}_1 = \mathbf{y}_1$ and $(\mathbf{I}-\mathbf{P})\mathbf{y} = \mathbf{y}_2$,

$$\mathbf{y}'\mathbf{P}'(\mathbf{I}-\mathbf{P})\mathbf{y} = \mathbf{y}_1'\mathbf{y}_2 = 0,$$

5 In Matlab, the `orth` function can be used. The implementation uses the singular value decomposition (svd) and attempts to determine the number of nonzero singular values. Because of numerical imprecision, this latter step can choose too many. Instead, just use `[U,S,V]=svd(M); dim=sum(round(diag(S))==1); G=U(:,1:dim)';`, where `dim` will equal $T-k$ for full rank **X** matrices.

because $\mathbf{y}_1$ and $\mathbf{y}_2$ are orthogonal. As $\mathbf{y}$ is arbitrary, $\mathbf{P}'(\mathbf{I}-\mathbf{P})$ must be $\mathbf{0}$, or $\mathbf{P}' = \mathbf{P}'\mathbf{P}$. From this and the symmetry of $\mathbf{P}'\mathbf{P}$, it follows that $\mathbf{P}$ is also symmetric. ■

The following fact will be the key to obtaining the o.l.s. estimator in a linear regression model, as discussed in Section 1.3.2.

Theorem 1.5 Vector $\mathbf{u}$ in $\mathcal{S}$ is the closest to $\mathbf{y}$ in the sense that

$$\|\mathbf{y}-\mathbf{u}\|^2 = \min_{\tilde{\boldsymbol{u}}\in\mathcal{S}} \|\mathbf{y}-\tilde{\boldsymbol{u}}\|^2.$$

Proof: Let $\mathbf{y} = \mathbf{u} + \mathbf{v}$, where $\mathbf{u} \in \mathcal{S}$ and $\mathbf{v} \in \mathcal{S}^\perp$. We have, for any $\tilde{\boldsymbol{u}} \in \mathcal{S}$,

$$\|\mathbf{y}-\tilde{\boldsymbol{u}}\|^2 = \|\mathbf{u}+\mathbf{v}-\tilde{\boldsymbol{u}}\|^2 = \|\mathbf{u}-\tilde{\boldsymbol{u}}\|^2 + \|\mathbf{v}\|^2 \geqslant \|\mathbf{v}\|^2 = \|\mathbf{y}-\mathbf{u}\|^2,$$

where the second equality holds because $\mathbf{v} \perp (\mathbf{u}-\tilde{\mathbf{u}})$. ■

The next theorem will be useful for testing whether the mean vector of a linear model lies in a subspace of $C(\mathbf{X})$, as developed in Section 1.4.

Theorem 1.6 Let $\mathcal{S}_0 \subset \mathcal{S}$ be subspaces of $\mathbb{R}^T$ with respective integer dimensions r and s, such that $0 < r < s < T$. Further, let $\mathcal{S}\backslash\mathcal{S}_0$ denote the subspace $\mathcal{S} \cap \mathcal{S}_0^\perp$ with dimension $s-r$, i.e., $\mathcal{S}\backslash\mathcal{S}_0 = \{\mathbf{s} : \mathbf{s} \in \mathcal{S};\ \mathbf{s} \perp \mathcal{S}_0\}$. Then

a. $\mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}_0}$ and $\mathbf{P}_{\mathcal{S}_0}\mathbf{P}_{\mathcal{S}} = \mathbf{P}_{\mathcal{S}_0}$.
b. $\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}_0}$.
c. $\|\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}\mathbf{y}\|^2 = \|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|^2 - \|\mathbf{P}_{\mathcal{S}_0}\mathbf{y}\|^2$.
d. $\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}_0^\perp\backslash\mathcal{S}^\perp} = \mathbf{P}_{\mathcal{S}_0^\perp} - \mathbf{P}_{\mathcal{S}^\perp}$.
e. $\mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}\mathbf{P}_{\mathcal{S}} = \mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}$.
f. $\|\mathbf{P}_{\mathcal{S}_0^\perp\backslash\mathcal{S}^\perp}\mathbf{y}\|^2 = \|\mathbf{P}_{\mathcal{S}_0^\perp}\mathbf{y}\|^2 - \|\mathbf{P}_{\mathcal{S}^\perp}\mathbf{y}^2\|$.

Proof: (part a) For all $\mathbf{y} \in \mathbb{R}^T$, as $\mathbf{P}_{\mathcal{S}_0}\mathbf{y} \in \mathcal{S}$, $\mathbf{P}_{\mathcal{S}}(\mathbf{P}_{\mathcal{S}_0}\mathbf{y}) = \mathbf{P}_{\mathcal{S}_0}\mathbf{y}$. Transposing yields the second result.

Another way of seeing this (and which is useful for proving the other results) is to partition $\mathbb{R}^T$ into subspaces $\mathcal{S}$ and $\mathcal{S}^\perp$, and then $\mathcal{S}$ into subspaces $\mathcal{S}_0$ and $\mathcal{S}\backslash\mathcal{S}_0$. Take as a basis for $\mathbb{R}^T$ the vectors

$$\underbrace{\overbrace{\mathbf{r}_1,\ldots,\mathbf{r}_r}^{\mathcal{S}_0 \text{ basis}},\overbrace{\mathbf{s}_{r+1},\ldots,\mathbf{s}_s}^{\mathcal{S}\backslash\mathcal{S}_0 \text{ basis}}}_{\mathcal{S} \text{ basis}},\underbrace{\mathbf{z}_{s+1},\ldots,\mathbf{z}_T}_{\mathcal{S}^\perp \text{ basis}} \tag{1.50}$$

and let $\mathbf{y} = \mathbf{r} + \mathbf{s} + \mathbf{z}$, where $\mathbf{r} \in \mathcal{S}_0$, $\mathbf{s} \in \mathcal{S}\backslash\mathcal{S}_0$ and $\mathbf{z} \in \mathcal{S}^\perp$ are orthogonal. Clearly, $\mathbf{P}_{\mathcal{S}_0}\mathbf{y} = \mathbf{r}$ while $\mathbf{P}_{\mathcal{S}}\mathbf{y} = \mathbf{r} + \mathbf{s}$ and $\mathbf{P}_{\mathcal{S}_0}\mathbf{P}_{\mathcal{S}}\mathbf{y} = \mathbf{P}_{\mathcal{S}_0}(\mathbf{r}+\mathbf{s}) = \mathbf{r}$.

The remaining proofs are developed in Problem 1.9. ■

1.3.2 Implementation

For the linear regression model

$$\mathbf{Y}_{(T\times 1)} = \mathbf{X}_{(T\times k)}\boldsymbol{\beta}_{(k\times 1)} + \boldsymbol{\epsilon}_{(T\times 1)}, \tag{1.51}$$

with subscripts indicating the sizes and $\epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_T)$, we seek that $\widehat{\boldsymbol{\beta}}$ such that $\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2$ is minimized. From Theorem 1.5, $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is given by $\mathbf{P}_\mathbf{X}\mathbf{Y}$, where $\mathbf{P}_\mathbf{X} \equiv \mathbf{P}_{C(\mathbf{X})}$ is an abbreviated notation for the projection matrix onto the space spanned by the columns of $\mathbf{X}$. We will assume that $\mathbf{X}$ is of full rank k, though this assumption can be relaxed in a more general treatment; see, e.g., Section 1.4.2.

If $\mathbf{X}$ happens to consist of k orthonormal column vectors, then $\mathbf{T} = \mathbf{X}$, where $\mathbf{T}$ is the orthonormal matrix given in (1.42), so that $\mathbf{P}_\mathbf{X} = \mathbf{T}\mathbf{T}'$. If (as usual), $\mathbf{X}$ is not orthonormal, with columns, say, $\mathbf{v}_1, \ldots, \mathbf{v}_k$, then $\mathbf{T}$ could be constructed by applying the Gram–Schmidt procedure to $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Recall that, under our assumption that $\mathbf{X}$ is full rank, $\mathbf{v}_1, \ldots, \mathbf{v}_k$ forms a basis (albeit not orthonormal) for $C(\mathbf{X})$.

This can be more compactly expressed in the following way: From Theorem 1.1, vector $\mathbf{Y}$ can be decomposed as $\mathbf{Y} = \mathbf{P}_\mathbf{X}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}$, with $\mathbf{P}_\mathbf{X}\mathbf{Y} = \sum_{i=1}^k c_i \mathbf{v}_i$, where $\mathbf{c} = (c_1, \ldots, c_k)'$ is the unique coefficient vector corresponding to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $C(\mathbf{X})$. Also from Theorem 1.1, $(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}$ is perpendicular to $C(\mathbf{X})$, i.e., $\langle (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}, \mathbf{v}_i \rangle = 0$, $i = 1, \ldots, k$. Thus,

$$\langle \mathbf{Y}, \mathbf{v}_j \rangle = \langle \mathbf{P}_\mathbf{X}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}, \mathbf{v}_j \rangle = \langle \mathbf{P}_\mathbf{X}\mathbf{Y}, \mathbf{v}_j \rangle = \left\langle \sum_{i=1}^k c_i \mathbf{v}_i, \mathbf{v}_j \right\rangle = \sum_{i=1}^k c_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle,$$

$j = 1, \ldots, k$, which can be written in matrix terms as

$$\begin{bmatrix} \langle \mathbf{Y}, \mathbf{v}_1 \rangle \\ \langle \mathbf{Y}, \mathbf{v}_2 \rangle \\ \vdots \\ \langle \mathbf{Y}, \mathbf{v}_k \rangle \end{bmatrix} = \begin{bmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_1, \mathbf{v}_k \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_2, \mathbf{v}_k \rangle \\ \vdots & \vdots & & \vdots \\ \langle \mathbf{v}_k, \mathbf{v}_1 \rangle & \langle \mathbf{v}_k, \mathbf{v}_2 \rangle & & \langle \mathbf{v}_k, \mathbf{v}_k \rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix},$$

or, in terms of $\mathbf{X}$ and $\mathbf{c}$, as $\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\mathbf{c}$. As $\mathbf{X}$ is full rank, so is $\mathbf{X}'\mathbf{X}$, showing that $\mathbf{c} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the coefficient vector for expressing $\mathbf{P}_\mathbf{X}\mathbf{Y}$ using the basis matrix $\mathbf{X}$. Thus, $\mathbf{P}_\mathbf{X}\mathbf{Y} = \mathbf{X}\mathbf{c} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, i.e.,

$$\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{1.52}$$

As $\mathbf{P}_\mathbf{X}\mathbf{Y}$ is unique from Theorem 1.1 (and from the full rank assumption on $\mathbf{X}$), it follows that the least squares estimator $\widehat{\boldsymbol{\beta}} = \mathbf{c}$. This agrees with the direct approach used in Section 1.2. Notice also that, if $\mathbf{X}$ is orthonormal, then $\mathbf{X}'\mathbf{X} = \mathbf{I}$ and $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ reduces to $\mathbf{X}\mathbf{X}'$, as in (1.42).

It is easy to see that $\mathbf{P}_\mathbf{X}$ is symmetric and idempotent, so that from Theorem 1.4 and the uniqueness of projection matrices (Problem 1.4), it is the projection matrix onto $\mathcal{S}$, the space spanned by its columns. To see that $\mathcal{S} = C(\mathbf{X})$, we must show that, for all $\mathbf{Y} \in \mathbb{R}^T$, $\mathbf{P}_\mathbf{X}\mathbf{Y} \in C(\mathbf{X})$ and $(\mathbf{I}_T - \mathbf{P}_\mathbf{X})\mathbf{Y} \perp C(\mathbf{X})$. The former is easily verified by taking $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in (1.38). The latter is equivalent to the statement that $(\mathbf{I}_T - \mathbf{P}_\mathbf{X})\mathbf{Y}$ is perpendicular to every column of $\mathbf{X}$. For this, defining the projection matrix

$$\mathbf{M} := \mathbf{I} - \mathbf{P}_\mathbf{X} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \tag{1.53}$$

we have

$$\mathbf{X}'\mathbf{M}\mathbf{Y} = \mathbf{X}'(\mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y}) = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{0}, \tag{1.54}$$

and the result is shown. Result (1.54) implies $\mathbf{M}\mathbf{X} = \mathbf{0}$. This follows from direct multiplication, but can also be seen as follows: Note that (1.54) holds for any $\mathbf{Y} \in \mathbb{R}^T$, and taking transposes yields $\mathbf{Y}'\mathbf{M}'\mathbf{X} = \mathbf{0}$, or, as $\mathbf{M}$ is symmetric, $\mathbf{M}\mathbf{X} = \mathbf{0}$.

Example 1.7 The method of Gram–Schmidt orthogonalization is quite naturally expressed in terms of projection matrices. Let $\mathbf{X}$ be a $T \times k$ matrix not necessarily of full rank, with columns $\mathbf{z}_1, \ldots, \mathbf{z}_k$, $\mathbf{z}_1 \neq \mathbf{0}$. Define $\mathbf{w}_1 = \mathbf{z}_1/\|\mathbf{z}_1\|$ and

$$\mathbf{P}_1 = \mathbf{P}_{C(\mathbf{z}_1)} = \mathbf{P}_{C(\mathbf{w}_1)} = \mathbf{w}_1(\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1' = \mathbf{w}_1\mathbf{w}_1'.$$

Now let $\mathbf{r}_2 = (\mathbf{I} - \mathbf{P}_1)\mathbf{z}_2$, which is the component in $\mathbf{z}_2$ perpendicular to $\mathbf{z}_1$. If $\|\mathbf{r}_2\| > 0$, then set $\mathbf{w}_2 = \mathbf{r}_2/\|\mathbf{r}_2\|$ and $\mathbf{P}_2 = \mathbf{P}_{C(\mathbf{w}_1,\mathbf{w}_2)}$, otherwise set $\mathbf{w}_2 = \mathbf{0}$ and $\mathbf{P}_2 = \mathbf{P}_1$. This is then repeated for the remaining columns of $\mathbf{X}$. The matrix $\mathbf{W}$ with columns consisting of the j nonzero $\mathbf{w}_i$, $1 \leqslant j \leqslant k$, is then an orthonormal basis for $C(\mathbf{X})$. ■

Example 1.8 Let $\mathbf{P}_\mathbf{X}$ be given in (1.52) with $\mathbf{1} \in C(\mathbf{X})$ and $\mathbf{P}_\mathbf{1} = \mathbf{1}\mathbf{1}'/T$ be the projection matrix onto $\mathbf{1}$, i.e., the line $(1, 1, \ldots, 1)$ in $\mathbb{R}^T$. Then, from Theorem 1.6, $\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{1}$ is the projection matrix onto $C(\mathbf{X})\backslash C(\mathbf{1})$ and

$$\|(\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{1})\mathbf{Y}\|^2 = \|\mathbf{P}_\mathbf{X}\mathbf{Y}\|^2 - \|\mathbf{P}_\mathbf{1}\mathbf{Y}\|^2.$$

Also from Theorem 1.6, $\|\mathbf{P}_{\mathbf{X}\backslash\mathbf{1}}\mathbf{Y}\|^2 = \|\mathbf{P}_{\mathbf{1}^\perp\backslash\mathbf{X}^\perp}\mathbf{Y}\|^2 = \|\mathbf{P}_{\mathbf{1}^\perp}\mathbf{Y}\|^2 - \|\mathbf{P}_{\mathbf{X}^\perp}\mathbf{Y}\|^2$. As

$$\begin{aligned}
\|\mathbf{P}_{\mathbf{X}\backslash\mathbf{1}}\mathbf{Y}\|^2 &= \|(\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{1})\mathbf{Y}\|^2 = \sum (\hat{Y} - \bar{Y})^2, \\
\|\mathbf{P}_{\mathbf{1}^\perp}\mathbf{Y}\|^2 &= \|(\mathbf{I} - \mathbf{P}_\mathbf{1})\mathbf{Y}\|^2 = \sum (Y_t - \bar{Y})^2, \\
\|\mathbf{P}_{\mathbf{X}^\perp}\mathbf{Y}\|^2 &= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2 = \sum (Y_t - \hat{Y})^2,
\end{aligned}$$

we see that

$$\sum_{t=1}^{T} (Y_t - \bar{Y})^2 = \sum_{t=1}^{T} (Y_t - \hat{Y})^2 + \sum_{t=1}^{T} (\hat{Y} - \bar{Y})^2, \tag{1.55}$$

proving (1.12). ■

Often it will be of interest to work with the estimated residuals of the regression (1.51), namely

$$\hat{\boldsymbol{\epsilon}} := \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I}_T - \mathbf{P}_\mathbf{X})\mathbf{Y} = \mathbf{M}\mathbf{Y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{M}\boldsymbol{\epsilon}, \tag{1.56}$$

where $\mathbf{M}$ is the projection matrix onto the orthogonal complement of $\mathbf{X}$, given in (1.53), and the last equality in (1.56) follows because $\mathbf{MX} = \mathbf{0}$, confirmed by direct multiplication or as shown in (1.54). From (1.4) and (1.56), the RSS can be expressed as

$$\text{RSS} = \mathsf{S}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = (\mathbf{M}\mathbf{Y})'\mathbf{M}\mathbf{Y} = \mathbf{Y}'\mathbf{M}\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}. \tag{1.57}$$

Example 1.9 ***Example 1.1, the Frisch–Waugh–Lovell Theorem, cont.***
From the symmetry and idempotency of $\mathbf{M}_1$, the expression in (1.21) can also also be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_2 &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y} = (\mathbf{X}_2'\mathbf{M}_1'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1'\mathbf{M}_1\mathbf{Y} \\
&= (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Z},
\end{aligned}$$

where $\mathbf{Q} = \mathbf{M}_1\mathbf{X}_2$ and $\mathbf{Z} = \mathbf{M}_1\mathbf{Y}$. That is, $\hat{\boldsymbol{\beta}}_2$ can be computed *not* by regressing $\mathbf{Y}$ onto $\mathbf{X}_2$, but by regressing *the residuals of* $\mathbf{Y}$ onto *the residuals of* $\mathbf{X}_2$, where residuals refers to having removed the component spanned by $\mathbf{X}_1$. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, then

$$\mathbf{Q} = \mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2,$$

and, with $\mathbf{I} = \mathbf{M}_1 + \mathbf{P}_1$,

$$\begin{aligned}(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y} &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{M}_1 + \mathbf{P}_1)\mathbf{Y} \\ &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Z},\end{aligned}$$

so that, under orthogonality, $\widehat{\boldsymbol{\beta}}_2$ can indeed be obtained by regressing $\mathbf{Y}$ onto $\mathbf{X}_2$. ■

It is clear that $\mathbf{M}$ should have rank $T-k$, or $T-k$ eigenvalues equal to one and k equal to zero. We can thus express $\hat{\sigma}^2$ given in (1.11) as

$$\hat{\sigma}^2 = \frac{\mathsf{S}(\widehat{\boldsymbol{\beta}})}{T-k} = \frac{(\mathbf{MY})'\mathbf{MY}}{T-k} = \frac{\mathbf{Y}'\mathbf{MY}}{\operatorname{rank}(\mathbf{M})} = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{Y}}{\operatorname{rank}(\mathbf{I}-\mathbf{P}_\mathbf{X})}. \tag{1.58}$$

Observe also that $\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon} = \mathbf{Y}'\mathbf{MY}$.

It is now quite easy to show that $\hat{\sigma}^2$ is unbiased. Using properties of the trace operator and the fact $\mathbf{M}$ is a projection matrix (i.e., $\mathbf{M}'\mathbf{M} = \mathbf{MM} = \mathbf{M}$),

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}] &= \mathbb{E}[\boldsymbol{\epsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}] = \operatorname{tr}(\mathbb{E}[\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}]) = \mathbb{E}[\operatorname{tr}(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon})] \\ &= \mathbb{E}[\operatorname{tr}(\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}')] = \operatorname{tr}(\mathbf{M}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']) = \sigma^2\operatorname{tr}(\mathbf{M}) = \sigma^2\operatorname{rank}(\mathbf{M}) = \sigma^2(T-k),\end{aligned}$$

where the fact that $\operatorname{tr}(\mathbf{M}) = \operatorname{rank}(\mathbf{M})$ follows from Theorem 1.2. In fact, a similar derivation was used to obtain the general result (A.6), from which it directly follows that

$$\mathbb{E}[\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}] = \operatorname{tr}(\sigma^2\mathbf{M}) + \mathbf{0}'\mathbf{M}\mathbf{0} = \sigma^2(T-k). \tag{1.59}$$

Theorem A.3 shows that, if $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$, then the vector $\mathbf{CY}$ is independent of the quadratic form $\mathbf{Y}'\mathbf{AY}$ if $\mathbf{C}\boldsymbol{\Sigma}\mathbf{A} = 0$. Using this with $\boldsymbol{\Sigma} = \mathbf{I}$, $\mathbf{C} = \mathbf{P}$ and $\mathbf{A} = \mathbf{M} = \mathbf{I} - \mathbf{P}$, it follows that $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{PY}$ and $(T-k)\hat{\sigma}^2 = \mathbf{Y}'\mathbf{MY}$ are independent. That is:

> Under the usual regression model assumptions (including that $\mathbf{X}$ is not stochastic, or is such that the model is variation-free), point estimators $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

This generalizes the well-known result in the i.i.d. case: Specifically, if $\mathbf{X}$ is just a column of ones, then $\mathbf{PY} = T^{-1}\mathbf{11}'\mathbf{Y} = (\bar{Y}, \bar{Y}, \ldots, \bar{Y})'$ and $\mathbf{Y}'\mathbf{MY} = \mathbf{Y}'\mathbf{M}'\mathbf{MY} = \sum_{t=1}^{T}(Y_t - \bar{Y})^2 = (T-1)S^2$, so that $\bar{Y}$ and S^2 are independent.

As $\widehat{\boldsymbol{\epsilon}} = \mathbf{M}\boldsymbol{\epsilon}$ is a linear transformation of the normal random vector $\boldsymbol{\epsilon}$,

$$(\widehat{\boldsymbol{\epsilon}} \mid \sigma^2) \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{M}), \tag{1.60}$$

though note that $\mathbf{M}$ is rank deficient (i.e., is less than full rank), with rank $T-k$, so that this is a degenerate normal distribution. In particular, by definition, $\widehat{\boldsymbol{\epsilon}}$ is in the column space of $\mathbf{M}$, so that $\widehat{\boldsymbol{\epsilon}}$ must be perpendicular to the column space of $\mathbf{X}$, or

$$\widehat{\boldsymbol{\epsilon}}'\mathbf{X} = \mathbf{0}. \tag{1.61}$$

If, as usual, $\mathbf{X}$ contains a column of ones, denoted $\mathbf{1}_T$, or, more generally, $\mathbf{1}_T \in C(\mathbf{X})$, then (1.61) implies that $\sum_{t=1}^{T}\hat{\epsilon}_t = 0$.

We now turn to the generalized least squares case, with the model given by (1.3) and (1.24), and estimator (1.28). In this more general setting when $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, the residual vector is given by

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{Y}, \tag{1.62}$$

where $\mathbf{M}_{\boldsymbol{\Sigma}} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$. Although $\mathbf{M}_{\boldsymbol{\Sigma}}$ is idempotent, it is not symmetric, and cannot be referred to as a projection matrix. Observe also that the estimated residual vector is no longer orthogonal to the columns of $\mathbf{X}$. Instead we have

$$\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}) = \mathbf{0}, \tag{1.63}$$

so that the residuals do not necessarily sum to zero.

We now state a result from matrix algebra, and then use it to prove a theorem that will be useful for some hypothesis testing situations in Chapter 5.

Theorem 1.7 Let $\mathbf{V}$ be an $n \times n$ positive definite matrix, and let $\mathbf{U}$ and $\mathbf{T}$ be $n \times k$ and $n \times (n-k)$ matrices, respectively, such that, if $\mathbf{W} = [\mathbf{U}, \mathbf{T}]$, then $\mathbf{W}'\mathbf{W} = \mathbf{W}\mathbf{W}' = \mathbf{I}_n$. Then

$$\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{U}(\mathbf{U}'\mathbf{V}^{-1}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}^{-1} = \mathbf{T}(\mathbf{T}'\mathbf{V}\mathbf{T})^{-1}\mathbf{T}'. \tag{1.64}$$

Proof: See Rao (1973, p. 77). ■

Let $\mathbf{P} = \mathbf{P}_{\mathbf{X}}$ be the usual projection matrix on the column space of $\mathbf{X}$ from (1.52), let $\mathbf{M} = \mathbf{I}_T - \mathbf{P}$, and let $\mathbf{G}$ and $\mathbf{H}$ be matrices such that $\mathbf{M} = \mathbf{G}'\mathbf{G}$ and $\mathbf{P} = \mathbf{H}'\mathbf{H}$, in which case $\mathbf{W} = [\mathbf{H}', \mathbf{G}']$ satisfies $\mathbf{W}'\mathbf{W} = \mathbf{W}\mathbf{W}' = \mathbf{I}_T$.

Theorem 1.8 For the regression model given by (1.3) and (1.24), with $\widehat{\boldsymbol{\epsilon}} = \mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{Y}$ from (1.62),

$$\widehat{\boldsymbol{\epsilon}}'\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}'\mathbf{G}'(\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')^{-1}\mathbf{G}\boldsymbol{\epsilon}. \tag{1.65}$$

Proof: As in King (1980, p. 1268), using Theorem 1.7 with $\mathbf{T} = \mathbf{G}'$, $\mathbf{U} = \mathbf{H}'$, and $\mathbf{V} = \boldsymbol{\Sigma}$, and the fact that $\mathbf{H}'$ can be written as $\mathbf{X}\mathbf{K}$, where $\mathbf{K}$ is a $k \times k$ full rank transformation matrix, we have

$$\begin{aligned}
\boldsymbol{\epsilon}'\mathbf{G}'(\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')^{-1}\mathbf{G}\boldsymbol{\epsilon} &= \mathbf{U}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}\mathbf{H}\boldsymbol{\Sigma}^{-1})\mathbf{U} \\
&= \mathbf{U}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{K})^{-1}\mathbf{K}'\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{U} \\
&= \mathbf{U}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{U} = \widehat{\boldsymbol{\epsilon}}'\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\epsilon}},
\end{aligned}$$

which is (1.65). ■

1.4 Linear Parameter Restrictions

> [D]eleting a small unimportant parameter from the model is generally a good idea, because we will incur a small bias but may gain much precision. This is true even if the estimated parameter happens to be highly 'significant', that is, have a large t-ratio. Significance indicates that we have managed to estimate the parameter rather precisely, possibly because we have many observations. It does not mean that the parameter is important.
>
> (Jan R. Magnus, 2017, p. 30)

In much applied regression analysis, the analyst will wish to know the extent to which certain linear restrictions on $\boldsymbol{\beta}$ hold. As the quote above by Magnus (2017) suggests, we recommend doing so via

means more related to the purpose of the research, e.g., forecasting, and, particularly, in applications in the social sciences for which the notion of repeatability of the experiment does not apply, being aware of the pitfalls of the classic significance testing (use of p-values) and Neyman–Pearson hypothesis testing paradigm. This issue was discussed in some detail in Section III.2.8, where strong arguments were raised, and evidence presented, that significance and hypothesis testing might one day make it to the ash heap of statistical history. In addition to the numerous references provided in Section III.2.8, such as Ioannidis (2005), the interested reader is encouraged to read Ioannidis (2014), and a rebuttal to that paper in Leek and Jager (2017), as well as the very pertinent overview in Spiegelhalter (2017), addressing this issue and the more general theme of trustworthiness in statistical reports, amid concerns of reproducibility, fake news, and alternative facts.

1.4.1 Formulation and Estimation

A common goal in regression analysis is to test is whether an individual regression coefficient is "significantly" different than a given value, often zero. More general tests might involve testing whether the sum of certain coefficients is a particular value, or testing for the equality of two or more coefficients. These are all special cases of a general linear test that can be expressed as (regrettably with many Hs, but following standard terminology)

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}, \tag{1.66}$$

versus the alternative, H_1, corresponding to the unrestricted model. The matrix $\mathbf{H}$ is of dimension $J \times k$ and, without loss of generality, assumed to be of full rank J, so that $J \leqslant k$ and $\mathbf{h}$ is $J \times 1$. The null hypothesis can also be written

$$H_0 : \mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \mathbf{X}\boldsymbol{\gamma} \in \mathcal{S}_H, \tag{1.67}$$

where

$$\mathcal{S}_H = \{\mathbf{z} : \mathbf{z} = \mathbf{X}\boldsymbol{\beta},\ \mathbf{H}\boldsymbol{\beta} = \mathbf{h},\ \boldsymbol{\beta} \in \mathbb{R}^k\}. \tag{1.68}$$

If $\mathbf{h} \neq \mathbf{0}$, then $\mathcal{S}_H$ is an **affine subspace** because it does not contain the zero element (provided both $\mathbf{X}$ and $\mathbf{H}$ are full rank, as is assumed).

As an important illustration, for testing if the last J regressors are not significant, i.e., if $\beta_{k-J+1} = \cdots = \beta_k = 0$, set $\mathbf{h} = \mathbf{0}$ and $\mathbf{H} = [\mathbf{0}_{J\times k-J} \mid \mathbf{I}_J]$. For example, if $k = 6$ and $J = 2$, then

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We next consider how $\boldsymbol{\gamma}$ in (1.67) can be estimated, followed by the distribution theory associated with the formal frequentist testing framework of the null hypothesis for assessing whether or not the data are in agreement with the proposed set of restrictions.

In many cases of interest, the reduced column space is easily identified. For example, if a set of coefficients are taken to be zero, then the nonzero elements of $\widehat{\boldsymbol{\gamma}}$ are found by computing the o.l.s. estimator using an $\mathbf{X}$ matrix with the appropriate columns removed. In general, however, it will not always be clear how to identify the reduced column space, so that a more general method will be required. Theorem 1.9 gives a nonconstructive proof, i.e., we state the result and confirm it satisfies the requirements. We subsequently show two constructive proofs.

Theorem 1.9 Assuming $\mathbf{H}$ and $\mathbf{X}$ are full rank, the least squares estimator of $\boldsymbol{\gamma}$ in (1.67) is given by

$$\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\beta}} + \mathbf{A}\mathbf{H}'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}), \tag{1.69}$$

where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$.

Proof: By definition, we require that $\widehat{\boldsymbol{\gamma}}$ is the least squares estimator subject to the linear constraint. Thus, the proof entails showing that (1.69) satisfies the following two conditions:

1) $\mathbf{H}\widehat{\boldsymbol{\gamma}} = \mathbf{h}$ and
2) $\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 \leqslant \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$ for all $\mathbf{b} \in \mathbb{R}^k$ such that $\mathbf{H}\mathbf{b} = \mathbf{h}$.

This is straightforward and detailed in Problem 1.6. ■

We will refer to $\widehat{\boldsymbol{\gamma}}$ in (1.69) as the **restricted least squares**, or r.l.s., estimator. It can be derived in several ways, two important ones of which are now shown. A third way, using projection, is also straightforward and instructive; see, e.g., Ravishanker and Dey (2002, Sec. 4.6.2) or Seber and Lee (2003, p. 61).

Derivation of (1.69) Method I: This method makes use of the results for the generalized least squares estimator and does not explicitly require the use of calculus. We will need the following well-known matrix result: If matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{D}$ are such that $\mathbf{A} + \mathbf{B}\mathbf{D}\mathbf{B}'$ is a square matrix of full rank, then

$$(\mathbf{A} + \mathbf{B}\mathbf{D}\mathbf{B}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B} + \mathbf{D}^{-1})^{-1}\mathbf{B}'\mathbf{A}^{-1}. \tag{1.70}$$

See, e.g., Abadir and Magnus (2005, p. 107) for proof of the more general case of $(\mathbf{A} + \mathbf{B}\mathbf{D}\mathbf{C}')^{-1}$.

Let (uncharacteristically, using a lower case letter) $\mathbf{v}$ be a vector random variable with mean $\mathbf{0}$ and finite covariance matrix $\sigma_{\mathbf{v}}^2\mathbf{V}$, denoted $\mathbf{v} \sim (\mathbf{0}, \sigma_{\mathbf{v}}^2\mathbf{V})$. The constraint in (1.66) can be understood as the limiting case, as $\sigma_{\mathbf{v}}^2 \to 0$, of the *stochastic* set of extraneous information equations on $\boldsymbol{\beta}$,

$$\mathbf{H}\boldsymbol{\beta} + \mathbf{v} = \mathbf{h}. \tag{1.71}$$

The regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_T$, can be combined with (1.71) via the so-called **mixed model** of Theil and Goldberger (1961) to give

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{H} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{v} \end{pmatrix}.$$

This can be expressed more compactly as

$$\mathbf{Y}_m = \mathbf{X}_m\boldsymbol{\beta}_m + \boldsymbol{\epsilon}_m, \quad \boldsymbol{\epsilon}_m \sim (\mathbf{0}, \boldsymbol{\Sigma}_m), \quad \boldsymbol{\Sigma}_m = \begin{pmatrix} \sigma^2\mathbf{I}_T & \mathbf{0} \\ \mathbf{0} & \sigma_{\mathbf{v}}^2\mathbf{V} \end{pmatrix},$$

where the subscript m denotes "mixed". Using generalized least squares,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_m &= (\mathbf{X}_m'\boldsymbol{\Sigma}_m^{-1}\mathbf{X}_m)^{-1}\mathbf{X}_m'\boldsymbol{\Sigma}_m^{-1}\mathbf{Y}_m \\ &= (\sigma^{-2}\mathbf{X}'\mathbf{X} + \sigma_{\mathbf{v}}^{-2}\mathbf{H}'\mathbf{V}^{-1}\mathbf{H})^{-1}(\sigma^{-2}\mathbf{X}'\mathbf{Y} + \sigma_{\mathbf{v}}^{-2}\mathbf{H}'\mathbf{V}^{-1}\mathbf{h}) \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{H}'\mathbf{V}^{-1}\mathbf{H})^{-1}(\mathbf{X}'\mathbf{Y} + \lambda\mathbf{H}'\mathbf{V}^{-1}\mathbf{h}), \end{aligned}$$

where $\lambda := \sigma^2/\sigma_{\mathbf{v}}^2$. Next, following Alvarez and Dolado (1994), use (1.70) with

$$\mathbf{A} := (\mathbf{X}'\mathbf{X})^{-1} \quad \text{and} \quad \mathbf{C}_\lambda := \mathbf{A}\mathbf{H}'(\mathbf{H}\mathbf{A}\mathbf{H}' + \lambda^{-1}\mathbf{V})^{-1}$$

to get

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_m &= [\mathbf{A} - \mathbf{C}_\lambda\mathbf{H}\mathbf{A}](\mathbf{X}'\mathbf{Y} + \mathbf{H}'(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h}) \\
&= \mathbf{A}\mathbf{X}'\mathbf{Y} + \mathbf{A}\mathbf{H}'(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h} - \mathbf{C}_\lambda\mathbf{H}\mathbf{A}\mathbf{X}'\mathbf{Y} - \mathbf{C}_\lambda\mathbf{H}\mathbf{A}\mathbf{H}'(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h} \\
&= \widehat{\boldsymbol{\beta}} + \mathbf{C}_\lambda(\mathbf{H}\mathbf{A}\mathbf{H}' + \lambda^{-1}\mathbf{V})(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h} - \mathbf{C}_\lambda\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{C}_\lambda\mathbf{H}\mathbf{A}\mathbf{H}'(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h} \\
&= \widehat{\boldsymbol{\beta}} + \mathbf{C}_\lambda[\mathbf{H}\mathbf{A}\mathbf{H}'(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h} + \mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{H}\mathbf{A}\mathbf{H}'(\lambda^{-1}\mathbf{V})^{-1}\mathbf{h}] \\
&= \widehat{\boldsymbol{\beta}} + \mathbf{C}_\lambda(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}),
\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}$ is the unrestricted least squares estimator. Letting $\sigma_{\mathbf{v}}^2 \to 0$ gives (1.69). Note that the inverse of $\mathbf{H}\mathbf{A}\mathbf{H}'$ exists because both $\mathbf{H}$ and $\mathbf{X}$ (and thus $\mathbf{A}$) are full rank. ■

Remark The mixed model structure is useful in several regression modeling contexts, and is related to formal Bayesian methods, whereby model parameters are treated as random variables, though not requiring Bayesian methodology. For example, as stated by Lee and Griffiths (1979, pp. 4–5), "Thus, for stochastic prior information of the form given in [(1.71)], the mixed estimation procedure is more efficient, is distribution free, and does not involve a Bayesian argument."

It also provides the most straightforward derivation of the so-called Black–Litterman model for incorporating viewpoints into a statistical model for financial portfolio allocation; see, e.g., Kolm et al. (2008, p. 362), as well as Black and Litterman (1992), Meucci (2006), Giacometti et al. (2007), Brandt (2010, p. 313), and the references therein. ■

Derivation of (1.69) Method II: The calculus technique of Lagrange multipliers is applicable in this setting.[6] Besides being of interest in itself for deriving $\widehat{\boldsymbol{\gamma}}$, we will subsequently need equation (1.72) derived along the way, in Section 1.4.2.

The method implies that the $k + J$ constraints

$$\begin{aligned}
\frac{\partial}{\partial\hat{\gamma}_i}\{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 + \lambda'(\mathbf{H}\widehat{\boldsymbol{\gamma}} - \mathbf{h})\} = 0, \quad i = 1, \ldots, k, \\
\mathbf{H}\widehat{\boldsymbol{\gamma}} - \mathbf{h} = \mathbf{0},
\end{aligned}$$

must be satisfied, where $\lambda = (\lambda_1, \ldots, \lambda_J)'$. The ith equation, $i = 1, \ldots, k$, is easily seen to be

$$2\sum_{t=1}^{T}(Y_t - \mathbf{x}_t'\widehat{\boldsymbol{\gamma}})(-x_{it}) + (\text{the } i\text{th component of } \mathbf{H}'\lambda) = 0,$$

so that the first k equations can be written together as $-2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}) + \mathbf{H}'\lambda = \mathbf{0}$. These, in turn, can be expressed together with constraint $\mathbf{H}\widehat{\boldsymbol{\gamma}} = \mathbf{h}$ as

$$\begin{bmatrix} 2\mathbf{X}'\mathbf{X} & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \widehat{\boldsymbol{\gamma}} \\ \lambda \end{bmatrix} = \begin{bmatrix} 2\mathbf{X}'\mathbf{Y} \\ \mathbf{h} \end{bmatrix}, \tag{1.72}$$

6 A particularly lucid discussion of Lagrange multipliers is provided by Hubbard and Hubbard (2002, Sec. 3.7).

from which an expression for $\widehat{\boldsymbol{\gamma}}$ could be derived using the formula for the inverse of a partitioned matrix. More directly, with $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$, the first set of constraints gives

$$\widehat{\boldsymbol{\gamma}} = \mathbf{A}\left(\mathbf{X}'\mathbf{Y} - \frac{1}{2}\mathbf{H}'\boldsymbol{\lambda}\right). \tag{1.73}$$

Inserting (1.73) into constraint $\mathbf{H}\widehat{\boldsymbol{\gamma}} = \mathbf{h}$ gives $\mathbf{HAX}'\mathbf{Y} - \frac{1}{2}\mathbf{HAH}'\boldsymbol{\lambda} = \mathbf{h}$ or (as we assume that $\mathbf{X}$ and $\mathbf{H}$ are full rank)

$$\boldsymbol{\lambda} = 2[\mathbf{HAH}']^{-1}[\mathbf{HAX}'\mathbf{Y} - \mathbf{h}] = 2[\mathbf{HAH}']^{-1}[\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h}],$$

where $\widehat{\boldsymbol{\beta}} = \mathbf{AX}'\mathbf{Y}$ is the unconstrained least squares estimator. Thus, from (1.73),

$$\begin{aligned}\widehat{\boldsymbol{\gamma}} &= \mathbf{A}\left(\mathbf{X}'\mathbf{Y} - \frac{1}{2}\mathbf{H}'\boldsymbol{\lambda}\right)\\ &= \mathbf{A}(\mathbf{X}'\mathbf{Y} - \mathbf{H}'[\mathbf{HAH}']^{-1}[\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h}])\\ &= \widehat{\boldsymbol{\beta}} - \mathbf{AH}'[\mathbf{HAH}']^{-1}[\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h}],\end{aligned}$$

which is the same as (1.69). ■

Remark Up to this point, we have considered the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ from (1.3). This is an example of what we refer to as a **static** model, as opposed to the important class of models involving time-varying coefficients $\boldsymbol{\beta}_t$, which we refer to as a type of **dynamic** model. Section 5.6 is dedicated to some dynamic model classes with time-varying $\boldsymbol{\beta}_t$. The most flexible way of dealing with estimation and inference of the linear model with time-varying parameters is via use of the so-called **state space representation** and **Kalman filtering** techniques; see the remarks at the end of Section 5.6.1.

In some contexts, one is interested in the dynamic regression model $Y_t = \mathbf{x}_t'\boldsymbol{\beta}_t + \epsilon_t$ subject to **time-varying linear constraints** $\mathbf{H}_t\boldsymbol{\beta}_t = \mathbf{h}_t$, generalizing (1.66). Examples of econometric models that use such structures, as well as the augmentation of the Kalman filter required for its estimation are detailed in Doran (1992) and Doran and Rambaldi (1997); see also Durbin and Koopman (2012). ■

1.4.2 Estimability and Identifiability

Expression (1.69) uses $\widehat{\boldsymbol{\beta}}$, which may not be well-defined, as occurs when $\mathbf{X}$ is rank deficient. In our presentation of the linear model for regression analysis, we always assume that $\mathbf{X}$ is of full rank (or can be transformed to be), so that (1.69) is computable. However, contexts exist for which it is natural and convenient to work with a rank deficient $\mathbf{X}$, such as the ANOVA models in Chapters 2 and 3. Use of such $\mathbf{X}$ matrices are common in these and other designed experiments; see, e.g., Graybill (1976) and Christensen (2011).

As a simple, unrealistic example to help illustrate the point, let the true data-generating process be given by $Y_t = \mu + \epsilon_t$, and consider using the model $Y_t = \mu_1 + \mu_2 + \epsilon_t$. Clearly, unique estimators of μ_1 and μ_2 do not exist, though $\mu_1 + \mu_2$ can be estimated. More generally, μ_1 and μ_2 can also be estimated, provided one imposes an additional linear constraint, e.g., $\mu_1 - \mu_2 = 0$. With this latter constraint, one would choose $\mathbf{H}$ and $\mathbf{h}$ in (1.66) such that μ_1 and μ_2 are equal, i.e., $\mathbf{H} = [1, -1]$ and $h = 0$. Of course, in this simple setting, $\widehat{\boldsymbol{\gamma}}$ is trivially obtained by fitting the regression with $\mathbf{X} = \mathbf{1}$, but observe that (1.69) cannot be used for computing it. A straightforward resolution, as proposed

in Greene and Seaks (1991), is to *define the restricted least squares estimator as the solution to* (1.72), written, say, as $\mathbf{Wd} = \mathbf{v}$, which will be unique if rank($\mathbf{W}$) = $k + J$.

In our example, $\mathbf{X}$ is a $T \times 2$ matrix of all ones, and

$$\mathbf{W} = \begin{bmatrix} 2\mathbf{X}'\mathbf{X} & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} 2T & 2T & 1 \\ 2T & 2T & -1 \\ 1 & -1 & 0 \end{bmatrix},$$

which is full rank, with rank $k + J = 3$, for any sample size T. Let $Y_\bullet = \sum_{t=1}^{T} Y_t$, so that $\mathbf{v}$ in (1.72) when expressed as $\mathbf{Wd} = \mathbf{v}$ is $[2Y_\bullet,\ 2Y_\bullet,\ 0]'$. The solution to

$$\mathbf{Wd} = \begin{bmatrix} 2T & 2T & 1 \\ 2T & 2T & -1 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \lambda \end{bmatrix} = \mathbf{v} = \begin{bmatrix} 2Y_\bullet \\ 2Y_\bullet \\ 0 \end{bmatrix}$$

is $\hat{\gamma}_i = Y_\bullet/(2T) = \bar{Y}/2$, $i = 1, 2$, (and $\lambda = 0$), as was obvious from the simple structure of the setup. An equivalent condition was derived in Bittner (1974): Estimator $\widehat{\boldsymbol{\gamma}}$ is unique if

$$\text{rank}\left(\begin{bmatrix} \mathbf{H} \\ \mathbf{X} \end{bmatrix}\right) = k, \tag{1.74}$$

which is clearly the case in this simple example.

We now briefly discuss the concept of **estimability**, which is related to **identifiability**, as defined in Section III.5.1.1. In the previous simple example, μ_1 and μ_2 are not identifiable, though $\mu_1 + \mu_2$ is estimable. For vector $\boldsymbol{\ell}$ of size $1 \times k$, the linear combination $\boldsymbol{\ell}\boldsymbol{\beta}$ is said to be **estimable** if it possesses a linear, unbiased estimator, say $\boldsymbol{\kappa}\mathbf{Y}$, where $\boldsymbol{\kappa}$ is a $1 \times T$ vector. If $\boldsymbol{\ell}\boldsymbol{\beta}$ is estimable, then $\boldsymbol{\ell}\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{\kappa}\mathbf{Y}] = \boldsymbol{\kappa}\mathbb{E}[\mathbf{Y}] = \boldsymbol{\kappa}\mathbf{X}\boldsymbol{\beta}$, so that $\boldsymbol{\ell} = \boldsymbol{\kappa}\mathbf{X}$, or $\boldsymbol{\ell}' = \mathbf{X}'\boldsymbol{\kappa}'$. This implies that $\boldsymbol{\ell}\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{\ell}' \in C(\mathbf{X}')$, recalling definition (1.38). In the simple example above, it is easy to see that, for $\boldsymbol{\ell} = (1,\ 1)$, $\boldsymbol{\ell}\boldsymbol{\beta}$ is estimable, i.e., $\mu_1 + \mu_2$ can be estimated, as we stated above. However, for $\boldsymbol{\ell} = (0,\ 1)$ and $\boldsymbol{\ell} = (1,\ 0)$, $\boldsymbol{\ell}\boldsymbol{\beta}$ is not estimable, as, obviously, $\nexists \boldsymbol{\kappa}$ such that $\boldsymbol{\ell}' = \mathbf{X}'\boldsymbol{\kappa}'$, which agrees with our intuition that neither μ_1 nor μ_2 is identifiable.

Turning to a slightly less trivial example, consider the regression model with sample size $T = 2n$ and

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & \mathbf{1}_n & \mathbf{0}_n \\ \mathbf{1}_n & \mathbf{0}_n & \mathbf{1}_n \end{bmatrix}. \tag{1.75}$$

The baseline (or null hypothesis) model is that all the observations have the same mean, which corresponds to use of only the first column in $\mathbf{X}$ in (1.75), whereas interest centers on knowing if the two populations, represented with samples $Y_1, \ldots, Y_n$ and $Y_{n+1}, \ldots, Y_T$, respectively, have different means, in which case the alternative model takes $\mathbf{X}$ in (1.75) to be the latter two columns. This is an example of a (balanced) one-way ANOVA model with $a = 2$ groups, studied in more detail in Chapter 2. The first regressor corresponds to the mean of all the data, while the other two correspond to the means specific to each of the two populations. It should be clear from the simple structure that the regression coefficients β_1, β_2, and β_3 are not simultaneously identified. However, it might be of interest to use the model in this form, such that β_1 refers to the overall mean, and β_2 (β_3) is the *deviation* of the mean in group one (two) from the overall mean β_1, in which case we want the constraint that $\beta_2 + \beta_3 = 0$. This is achieved by taking $\mathbf{H} = (0,\ 1,\ 1)$ and $h = 0$.

```
X= [1 1 0 ; 1 1 0; 1 0 1; 1 0 1]; ell = [1 0 1];
kappaPRIME = pinv(X') * ell' % try to solve
% now check:
disc = ell' - X' * kappaPRIME; check = sum(abs(disc)) % should be zero if estimable
```

Program Listing 1.3: Attempts to solve $\boldsymbol{\ell}' = \mathbf{X}'\boldsymbol{\kappa}'$ for $\boldsymbol{\kappa}$ via use of the generalized inverse.

Clearly, $\mathbf{X}$ in (1.75) is rank deficient, with $\text{rank}(\mathbf{X}) = 2$, also seen by deleting all redundant rows, to give

$$\mathbf{X}^* = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

which is (full) rank 2. From (1.74),

$$\text{rank}\left(\begin{bmatrix} \mathbf{H} \\ \mathbf{X} \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} \mathbf{H} \\ \mathbf{X}^* \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}\right) = 3 = k,$$

so that estimator $\widehat{\boldsymbol{\gamma}}$ is unique, also seen from

$$\mathbf{W} = \begin{bmatrix} 2n & n & n & 0 \\ n & n & 0 & 1 \\ n & 0 & n & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

which is (full) rank $k + J = 4$.

Without constraints on $\boldsymbol{\beta}$, for $\boldsymbol{\ell} = (1,\ 1,\ 1)$ and $\boldsymbol{\ell} = (0,\ 1,\ 1)$, $\boldsymbol{\ell}\boldsymbol{\beta}$ is not estimable because $\nexists \boldsymbol{\kappa}$ such that $\boldsymbol{\ell}' = \mathbf{X}'\boldsymbol{\kappa}'$, which the reader should confirm, and also should make intuitive sense. Likewise, $\boldsymbol{\ell}\boldsymbol{\beta}$ is estimable for $\boldsymbol{\ell} = (1,\ 0,\ 1)$ and $\boldsymbol{\ell} = (1,\ 1,\ 0)$ (both of which form the two unique rows of $\mathbf{X}$). These results can be checked using Matlab with the code given in Listing 1.3, taking $n = 2$. For example, running it with $\boldsymbol{\ell} = (1,\ 0,\ 1)$ yields solution $\boldsymbol{\kappa} = (0,\ 0,\ 1/2,\ 1/2)$. Inspection shows another solution to be $(1/2, -1/2,\ 1/2,\ 1/2)$, emphasizing that $\boldsymbol{\kappa}$ need not be unique, only that $\boldsymbol{\ell}' \in C(\mathbf{X}')$.

A good discussion of estimability (and also its connection to their software) is provided in SAS/STAT 9.2 User's Guide (2008, Ch. 15), from which our notation was inspired (they use $\mathbf{L}$ and $\mathbf{K}$ in place of our $\boldsymbol{\ell}$ and $\boldsymbol{\kappa}$).

1.4.3 Moments and the Restricted GLS Estimator

Derivation of the first two moments of $\widehat{\boldsymbol{\gamma}}$ is straightforward: As $\widehat{\boldsymbol{\beta}}$ is unbiased, (1.69) implies

$$\mathbb{E}[\widehat{\boldsymbol{\gamma}}] = \boldsymbol{\beta} + \mathbf{A}\mathbf{H}'(\mathbf{H}\mathbf{A}\mathbf{H}')^{-1}(\mathbf{h} - \mathbf{H}\boldsymbol{\beta}), \tag{1.76}$$

where, as usual, $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$. It is then easy to verify that $\widehat{\boldsymbol{\gamma}} - \mathbb{E}[\widehat{\boldsymbol{\gamma}}] = (\mathbf{I} - \mathbf{B})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, where $\mathbf{B} = \mathbf{A}\mathbf{H}'(\mathbf{H}\mathbf{A}\mathbf{H}')^{-1}\mathbf{H}$, and

$$(\mathbf{I} - \mathbf{B})\mathbf{A}(\mathbf{I} - \mathbf{B}') = \mathbf{A} - \mathbf{B}\mathbf{A} - \mathbf{A}\mathbf{B}' + \mathbf{B}\mathbf{A}\mathbf{B}' = \mathbf{A} - \mathbf{B}\mathbf{A},$$

so that

$$\mathbb{V}(\widehat{\boldsymbol{\gamma}} \mid \sigma^2) = \mathbb{E}[(\widehat{\boldsymbol{\gamma}} - \mathbb{E}[\widehat{\boldsymbol{\gamma}}])(\widehat{\boldsymbol{\gamma}} - \mathbb{E}[\widehat{\boldsymbol{\gamma}}])' \mid \sigma^2] = (\mathbf{I} - \mathbf{B})\mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2)(\mathbf{I} - \mathbf{B})'$$
$$= \sigma^2(\mathbf{I} - \mathbf{B})\mathbf{A}(\mathbf{I} - \mathbf{B})' = \sigma^2(\mathbf{I} - \mathbf{B})\mathbf{A} = \mathbb{V}(\widehat{\boldsymbol{\beta}}) - \mathbf{K}, \tag{1.77}$$

where $\mathbf{K} = \sigma^2\mathbf{B}\mathbf{A} = \sigma^2\mathbf{A}\mathbf{H}'(\mathbf{H}\mathbf{A}\mathbf{H}')^{-1}\mathbf{H}\mathbf{A}$ is positive semi-definite for $J < k$ (Problem 1.12), so that $\widehat{\boldsymbol{\gamma}}$ has a lower variance than $\widehat{\boldsymbol{\beta}}$, assuming that the same estimate of σ^2 is used. Observe, however, that if the null hypothesis is wrong, then, via the bias evident in (1.76) with $\mathbf{h} \neq \mathbf{H}\boldsymbol{\beta}$, the mean squared error (hereafter m.s.e.) of $\widehat{\boldsymbol{\gamma}}$ could be higher than that of $\widehat{\boldsymbol{\beta}}$. A good discussion of this and related issues is provided in Judge et al. (1985, pp. 52–62).

So far, the derivation of $\widehat{\boldsymbol{\gamma}}$ pertained to the linear regression model with i.i.d. normal errors. If the errors instead are of the form $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$ for known positive definite matrix $\boldsymbol{\Sigma}$, then we can combine the methods of g.l.s. and r.l.s. In particular, just use (1.69) with $\boldsymbol{\Sigma}^{-1/2}\mathbf{Y}$ in place of $\mathbf{Y}$ and $\boldsymbol{\Sigma}^{-1/2}\mathbf{X}$ in place of $\mathbf{X}$. We will denote this estimator as $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\Sigma}}$ and refer to it as the **restricted generalized least squares**, or r.g.l.s., estimator.

Example 1.10 We wish to compute by simulation the m.s.e. of $\widehat{\boldsymbol{\beta}}$ based on the four estimators o.l.s., g.l.s., r.l.s. and r.g.l.s., using, for convenience, the scalar measure $M = \sum_{i=1}^{k}(\hat{\beta}_i - \beta)^2$. Let the model be

$$Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 X_{t,3} + \beta_4 X_{t,4} + \epsilon_t, \quad t = 1, \ldots, T = 20,$$

for $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_T)' \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a known, full rank covariance matrix, and the regression parameters are constrained as $\beta_2 + \beta_3 + \beta_4 = 1$, for which we take $\beta_1 = 10$, $\beta_2 = 0.4$, $\beta_3 = -0.2$ and $\beta_4 = 1 - \beta_2 - \beta_3 = 0.8$. The choice of $\mathbf{X}$ matrix will determine the m.s.e., and so, for each of the 50,000 replications, we let $X_{t,i} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$, $i = 2, 3, 4$, $t = 1, \ldots, T$. Measure M is then approximated by its sample average.

Five models are used. The first takes $\epsilon \sim \mathrm{N}(0, \sigma^2 w_t)$, $w_t = \sqrt{t}$; the second is with $w_t = t$. The third and fourth models assume an AR(1) structure for ϵ_t (recall Example 1.3), with parameters $a = 0.25$ and

```
function compareRGLS
T=20; beta=[10 0.4 -0.2 0.8]'; H=[0 1 1 1]; h=1;
Sigma = diag( [(1:T)'].^(0.5)); Sigmainv=inv(Sigma);
[V,D]=eig(0.5*(Sigma+Sigma')); W=sqrt(D);
Sighalf = V*W*V'; Sighalfinv=inv(Sighalf);
sim=500; emat=zeros(sim,4);
for s=1:sim
  X=[ones(T,1),randn(T,3)]; y=X*beta+Sighalf*randn(T,1);
  OLS = inv(X'*X)*X'*y; GLS = inv(X'*Sigmainv*X)*X'*Sigmainv*y;
  RLS = OLSrestrict(y,X,H,h);
  RGLS = OLSrestrict(Sighalfinv*y,Sighalfinv*X,H,h);
  emat(s,:) = [sum((OLS-beta).^2) sum((GLS-beta).^2) ...
               sum((RLS-beta).^2) sum((RGLS-beta).^2)];
end
M=mean(emat)

function gamma = OLSrestrict(y,X,H,h)
  [J,k]=size(H); if nargin<4, h=zeros(J,1); end
  b=regress(y,X); A=inv(X'*X); gamma = b+A*H'*inv(H*A*H')*(h-H*b);
```

Program Listing 1.4: Compares performance of o.l.s., g.l.s., r.l.s., and r.g.l.s. for a specific model.

Table 1.1 Empirical mean squared error over the four regression parameters, based on 50,000 replications.

	Model				
Method	**1**	**2**	**3**	**4**	**5**
o.l.s.	0.80	2.73	0.30	0.44	0.36
g.l.s.	0.72	1.85	0.28	0.36	0.28
r.l.s.	0.56	1.90	0.22	0.35	0.27
r.g.l.s.	0.50	1.23	0.21	0.29	0.22

$a = 0.5$, respectively. The fifth model assumes an MA(1) structure for ϵ_t with $b = 0.5$. The program to compute M is given in Listing 1.4. The results are shown in Table 1.1.

We see that, for all the models, o.l.s. is the worst and r.g.l.s. is the best estimator. Model 2 stands out because the covariance matrix differs markedly from the identity matrix. As such, the difference between o.l.s. and g.l.s., and the difference between r.l.s. and r.g.l.s. is quite large. For the other models, these differences are less pronounced, particularly for model 3 (the AR(1) with $a = 0.25$). ■

1.4.4 Testing With $\mathbf{h} = \mathbf{0}$

> The source of all great mathematics is the special case, the concrete example. It is frequent in mathematics that every instance of a concept of seemingly great generality is in essence the same as a small and concrete special case.
>
> (Paul R. Halmos, 1985, p. 324)

The above quote from Halmos is not fully applicable here because the general case of $\mathbf{h} \neq \mathbf{0}$ *is* important. It is straightforward and subsequently detailed, but the derivation for the special case $\mathbf{h} = \mathbf{0}$ is both easier and more intuitive because it turns out that we can explicitly express the projection matrix corresponding to $\mathcal{S}_H$.

With $\mathcal{S} = \mathcal{C}(\mathbf{X})$ and $\mathcal{S}_H \subset \mathcal{S}$ as defined in (1.68), consider the hypothesis as given in (1.67), but with the additional normality assumption:

$$H_0 : \mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}),\ \mathbf{X}\boldsymbol{\gamma} \in \mathcal{S}_H$$
$$H_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}),\ \mathbf{X}\boldsymbol{\beta} \in \mathcal{S}.$$

For notational convenience, denote the projection matrix onto $\mathcal{C}(\mathbf{X})$ as simply $\mathbf{P}$ instead of $\mathbf{P}_{\mathcal{S}}$, let $\mathbf{M} = \mathbf{I} - \mathbf{P}$ and let $\mathbf{P}_{\mathcal{H}} = \mathbf{P}_{\mathcal{S}_H}$. With $\mathbf{h} = \mathbf{0}$, $\mathbf{X}\widehat{\boldsymbol{\gamma}}$ from (1.69) can be expressed as

$$\begin{aligned}\mathbf{X}\widehat{\boldsymbol{\gamma}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{A}\mathbf{H}'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}\mathbf{H}\widehat{\boldsymbol{\beta}} \\ &= (\mathbf{X}\mathbf{A}\mathbf{X}' - \mathbf{X}\mathbf{A}\mathbf{H}'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}\mathbf{H}\mathbf{A}\mathbf{X}')\mathbf{Y} \\ &= (\mathbf{P} - \mathbf{X}\mathbf{A}\mathbf{H}'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}\mathbf{H}\mathbf{A}\mathbf{X}')\mathbf{Y} =: (\mathbf{P} - \mathbf{N})\mathbf{Y}, \end{aligned} \tag{1.78}$$

where $\mathbf{N}$ is so defined. Straightforward algebra verifies that $\mathbf{P} - \mathbf{N}$ is symmetric and idempotent, so that, from Theorem 1.4, it is the unique projection matrix onto the subspace

$$\{\mathbf{z} : \mathbf{z} = \mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^k,\ \mathbf{H}\boldsymbol{\beta} = \mathbf{0}\}.$$

Thus, for $\mathbf{h} = \mathbf{0}$, we can express $\mathbf{P}_{\mathcal{H}}$ explicitly as

$$\mathbf{P}_{\mathcal{H}} = \mathbf{P} - \mathbf{N} = \mathbf{I} - \mathbf{M} - \mathbf{N}, \quad \text{where} \quad \mathbf{P} - \mathbf{P}_{\mathcal{H}} = \mathbf{N} \tag{1.79}$$

is symmetric and idempotent. Then, from Theorem 1.2, $\text{rank}(\mathbf{N}) = \text{tr}(\mathbf{N})$, where

$$\text{tr}(\mathbf{N}) = \text{tr}(\mathbf{XAH}'[\mathbf{HAH}']^{-1}\mathbf{HAX}') = \text{tr}([\mathbf{HAH}']^{-1}\mathbf{HAX}'\ \mathbf{XAH}') = \text{tr}(\mathbf{I}_J) = J.$$

The constrained residual vector is then $\widehat{\boldsymbol{\epsilon}}_H = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}$, or

$$(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}) = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\gamma}} = (\mathbf{I} - \mathbf{P}_{\mathcal{H}})\mathbf{Y} = (\mathbf{M} + \mathbf{N})\mathbf{Y},$$

so that $(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}) = \mathbf{NY}$. The following result is (in light of previous results) simple, and very important:

From Theorem 1.6,

$$\mathbf{PP}_{\mathcal{H}} = \mathbf{P}_{\mathcal{H}}\mathbf{P} = \mathbf{P}_{\mathcal{H}}, \quad \text{and} \quad \mathbf{N} = \mathbf{P} - \mathbf{P}_{\mathcal{H}} = \mathbf{P}_{\mathcal{S}\backslash\mathcal{H}} \text{ is a projection matrix.} \tag{1.80}$$

In particular, note that $\mathbf{X}\widehat{\boldsymbol{\gamma}} = \mathbf{P}_{\mathcal{H}}\mathbf{Y} = \mathbf{P}_{\mathcal{H}}\mathbf{PY} = \mathbf{P}_{\mathcal{H}}\mathbf{X}\widehat{\boldsymbol{\beta}}$, so that $\mathbf{X}\widehat{\boldsymbol{\gamma}}$ is the projection of $\mathbf{X}\widehat{\boldsymbol{\beta}}$ onto $\mathcal{S}_H$.

If H_0 is true, then $\mathbf{PY}$ and $\mathbf{P}_{\mathcal{H}}\mathbf{Y}$ should be close, with the discrepancy arising only from sampling error. A natural measure[7] of the magnitude of the difference is the norm, $\|(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y}\|$, or its square, given by

$$[(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y}]'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y} = \mathbf{Y}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y}.$$

From (A.6),

$$\mathbb{E}[\mathbf{Y}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y}] = \sigma^2\text{rank}(\mathbf{P} - \mathbf{P}_{\mathcal{H}}) + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{X}\boldsymbol{\beta}, \tag{1.81}$$

where the latter term is, from (1.79), given by

$$\boldsymbol{\beta}'\mathbf{X}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{NX}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{H}'[\mathbf{HAH}']^{-1}\mathbf{H}\boldsymbol{\beta}. \tag{1.82}$$

Under H_0, $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\gamma}$ so that

$$(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{X}\boldsymbol{\beta} = (\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{X}\boldsymbol{\gamma} = \mathbf{0}, \tag{1.83}$$

and (1.81) reduces to

$$\mathbb{E}[\mathbf{Y}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y}] = \sigma^2\text{rank}(\mathbf{P} - \mathbf{P}_{\mathcal{H}}) = \sigma^2\text{rank}(\mathbf{N}) = J\sigma^2. \tag{1.84}$$

By using $\hat{\sigma}^2$ from the unrestricted model as an estimate for σ^2, as given in (1.58), and dividing $\mathbf{Y}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y}$ by $J\hat{\sigma}^2 = \text{rank}(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\hat{\sigma}^2$, we expect the value

$$F = \frac{\mathbf{Y}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y} \ / \ \text{rank}(\mathbf{P} - \mathbf{P}_{\mathcal{H}})}{\hat{\sigma}^2} = \frac{\mathbf{Y}'(\mathbf{P} - \mathbf{P}_{\mathcal{H}})\mathbf{Y} \ / \ \text{rank}(\mathbf{P} - \mathbf{P}_{\mathcal{H}})}{\mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} \ / \ \text{rank}(\mathbf{I} - \mathbf{P})} \tag{1.85}$$

7 Other measures, such as the sum or maximum of the vector of absolute values might also seem "natural". However, the sampling distribution of the chosen measure is tractable, and also leads to a UMPI test.

to be "close to" one under H_0 and larger than one under H_1. The choice of variable name F alludes to its distribution, which will be shown shortly. Before doing so, we first note that

$$\mathbf{Y}'(\mathbf{P}-\mathbf{P}_{\mathcal{H}})\mathbf{Y} = \mathbf{Y}'\mathbf{P}'\mathbf{P}\mathbf{Y} - \mathbf{Y}'\mathbf{P}_{\mathcal{H}}'\mathbf{P}_{\mathcal{H}}\mathbf{Y} = \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 - \|\mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2, \tag{1.86}$$

or, in terms of sums of squares quantities already defined,

$$\begin{aligned}\mathbf{Y}'(\mathbf{P}-\mathbf{P}_{\mathcal{H}})\mathbf{Y} &= \mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathcal{H}})\mathbf{Y} - \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I}-\mathbf{P}_{\mathcal{H}})'(\mathbf{I}-\mathbf{P}_{\mathcal{H}})\mathbf{Y} - \mathbf{Y}'(\mathbf{I}-\mathbf{P})'(\mathbf{I}-\mathbf{P})\mathbf{Y} \\ &= \mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}}).\end{aligned} \tag{1.87}$$

(These also follow from Theorem 1.6.) Thus, from (1.84) and (1.87), F in (1.85) can also be expressed in the attractively simple form

$$F = \frac{[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})]/J}{\mathsf{S}(\widehat{\boldsymbol{\beta}})/(T-k)} = \frac{\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})}{J\,\hat{\sigma}^2}. \tag{1.88}$$

Direct calculation shows $(\mathbf{I}-\mathbf{P})(\mathbf{P}-\mathbf{P}_{\mathcal{H}}) = \mathbf{0}$, so that

$$(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\epsilon}} = (\mathbf{I}-\mathbf{P})\mathbf{Y} \perp (\mathbf{P}-\mathbf{P}_{\mathcal{H}})\mathbf{Y} = (\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}),$$

and computing the squared length of both sides of $\widehat{\boldsymbol{\epsilon}}_H = (\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}})$ yields

$$\mathsf{S}(\widehat{\boldsymbol{\gamma}}) = \mathsf{S}(\widehat{\boldsymbol{\beta}}) + \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2. \tag{1.89}$$

Thus, $\widehat{\boldsymbol{\epsilon}}_H$ can be decomposed into two orthogonal parts, $\widehat{\boldsymbol{\epsilon}} = \mathbf{M}\mathbf{Y}$ and $\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}$. In fact, substituting $\widehat{\boldsymbol{\gamma}}$ from (1.69) into $\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2$ and simplifying shows that (for any $\mathbf{h}$, not just $\mathbf{0}$), from (1.89),

$$\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}}) = (\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}})'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}), \tag{1.90}$$

so that $\widehat{\boldsymbol{\gamma}}$ and $\mathsf{S}(\widehat{\boldsymbol{\gamma}})$ need not be explicitly calculated. Also, (1.81), (1.82) and (1.87) imply that

$$\mathbb{E}[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})] = \sigma^2 J + \boldsymbol{\beta}'\mathbf{H}'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}\mathbf{H}\boldsymbol{\beta}. \tag{1.91}$$

As an aside, from (1.86), (1.87) and (1.89), $\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 = \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 - \|\mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2$. By direct expansion, $\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 = \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 - 2\mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\gamma}}$, implying $\mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\gamma}} = \|\mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2$, i.e., that $\widehat{\boldsymbol{\gamma}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}'\mathbf{X}'\mathbf{Y}$. It is *not* true, however, that $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\gamma}} = \mathbf{X}'\mathbf{Y}$, which obviously holds for $\widehat{\boldsymbol{\beta}}$, i.e., $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$ from (1.6).

To obtain the distribution of F, recall Theorems A.1 and A.2. With $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we see that the product $\mathbf{N}\boldsymbol{\Sigma} = (\mathbf{P}-\mathbf{P}_{\mathcal{H}})\sigma^2\mathbf{I}$ is not idempotent, but it is only a scale factor that gets in the way. So, using Theorem A.1 and the fact that $(\mathbf{Y}/\sigma) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}/\sigma, \mathbf{I})$,

$$(\mathbf{Y}/\sigma)'(\mathbf{P}-\mathbf{P}_{\mathcal{H}})(\mathbf{Y}/\sigma) \sim \chi^2(J, \boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\mathbf{P}_{\mathcal{H}})\mathbf{X}\boldsymbol{\beta}/\sigma^2), \tag{1.92}$$

and, as $(\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{0}$,

$$(\mathbf{Y}/\sigma)'(\mathbf{I}-\mathbf{P})(\mathbf{Y}/\sigma) \sim \chi^2(T-k, 0). \tag{1.93}$$

As $(\mathbf{P}-\mathbf{P}_{\mathcal{H}})(\mathbf{I}-\mathbf{P}) = \mathbf{0}$, Theorem A.2 implies that the numerator and denominator of F are independent. By dividing both the numerator and denominator by σ^2, it follows that F follows a (singly) noncentral F distribution,

$$F \sim \mathrm{F}(J, T-k, \theta), \qquad \theta = \boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\mathbf{P}_{\mathcal{H}})\mathbf{X}\boldsymbol{\beta} \;/\; \sigma^2. \tag{1.94}$$

Recalling (1.83), the noncentrality parameter θ is zero under the null H_0. Thus, a test with size α of $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$ against the unrestricted alternative H_1 is to reject when $F > c$, where c is the quantile for which $\Pr(F(J, T-k) \geqslant c) = \alpha$.

The test with $H_0 : \beta_i = 0$, $1 \leqslant i \leqslant k$, is a very important special case in multiple regression, as it tests whether the contribution of the ith regressor is "significant". Then $J = 1$, $\mathbf{H}$ is a row vector of zeros with a one in the ith place, $\mathbf{h} = 0$, and the test $F > c$ is equivalent to a two-sided t-test, recalling the relation between the F and t distributions (see, e.g., page II.374).

1.4.5 Testing With Nonzero h

If $\mathbf{h} \neq \mathbf{0}$, then S_H is not a subspace, in which case $\mathbf{P}_{\mathcal{H}}$ should be viewed as an "operator" and not as a matrix. In particular, it is easy to see that an expression such as (1.78) in which $\mathbf{Y}$ can be factored out onto the right-hand side is no longer possible. However, we discovered that (1.90) (stated here again)

$$\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}}) = (\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}})'[\mathbf{HAH}']^{-1}(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}), \qquad (*1.90*)$$

also holds for $\mathbf{h} \neq \mathbf{0}$. As such, we might postulate that a similar expression as in (1.91) holds for $\mathbf{h} \neq \mathbf{0}$, i.e.,

$$\mathbb{E}[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})] \stackrel{?}{=} \sigma^2 J + (\mathbf{h} - \mathbf{H}\boldsymbol{\beta})'[\mathbf{HAH}]^{-1}(\mathbf{h} - \mathbf{H}\boldsymbol{\beta}). \qquad (1.95)$$

This is indeed true: Using (1.90), define vector random variable $\mathbf{Z}$ such that

$$\sigma\mathbf{Z} = \mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h} = \mathbf{HAX}'\mathbf{Y} - \mathbf{h} = \mathbf{HAX}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \mathbf{h} = \mathbf{H}\boldsymbol{\beta} - \mathbf{h} + \mathbf{HAX}'\boldsymbol{\epsilon},$$

so that $\mathbf{Z} \sim \mathrm{N}(\sigma^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{h}),\ \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \sigma^{-2}\mathbf{HAX}'\ \sigma^2\mathbf{I}\ \mathbf{XAH} = \mathbf{HAH}' > 0$, and

$$\sigma^{-2}[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})] = \mathbf{Z}'[\mathbf{HAH}']^{-1}\mathbf{Z}.$$

Then, from Theorem A.1, as $[\mathbf{HAH}']^{-1}\boldsymbol{\Omega} = \mathbf{I}_J$ is idempotent,

$$\sigma^{-2}[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})] \sim \chi^2(J, \eta), \qquad \eta = \sigma^{-2}(\mathbf{H}\boldsymbol{\beta} - \mathbf{h})'[\mathbf{HAH}']^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{h}). \qquad (1.96)$$

Using the fact that $\mathbb{E}[\chi^2(J, \eta)] = J + \eta$, (1.95) follows. Also, under the null hypothesis $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$, $\sigma^{-2}[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})] \sim \chi^2(J, 0)$.

From (1.90), the only stochastic element in $\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})$ is $\widehat{\boldsymbol{\beta}}$, which implies that $\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})$ is independent of $\hat{\sigma}^2$. Thus, the F statistic defined above in (1.88), i.e.,

$$F = \frac{[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})]/J}{\mathsf{S}(\widehat{\boldsymbol{\beta}})/(T-k)} = \frac{\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})}{J\ \hat{\sigma}^2}, \qquad (1.88)$$

follows the noncentral F distribution, $F \sim \mathrm{F}(J, (T-k), \eta)$.

1.4.6 Examples

Example 1.11 A company claims that its new method of coaching for a particular college entrance exam is superior to the old, standard method. In particular, they say that, initially, the student's improvement is slower than that using the old method, but as the student "gets the hang of it", they improve faster than they would training with the old method. For both methods, customers have the choice of how many full-day sessions they wish to take, with one, two, three, or four being typical.

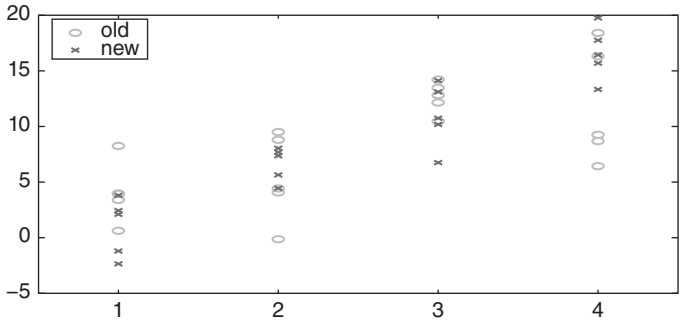

Figure 1.3 Percentage improvement for the two test groups as a function of number of sessions.

To test the claim, a study was conducted (by an independent researcher) as follows. From a total of $T = 40$ people interested in taking lessons (and who have never previously taken the exam or such a study course), 20 were randomly assigned to the standard method, say A, and the other 20 to the new method, say B. For each group of 20, 5 received one session, 5 two sessions, 5 three and 5 four sessions. Each person took a practice exam before, and a practice exam after "treatment" and Y_i, the percent improvement of each person, was recorded. The resulting (fictitious) data are shown in Figure 1.3. The claim is that, when using a simple linear regression to model the data as a function of s, the number of sessions, the intercept under teaching method B will be lower than that of A, while the slope (the coefficient of s) will be higher.

One way of modeling this is to let $\mathbf{Y}$ be the stack of observations Y_i such that the first 20 belong to group A, the second 20 to group B, and within a group, the first five correspond to $s = 1$, the next five to $s = 2$, etc. The 40×4 design matrix $\mathbf{X}$ for the unrestricted model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is then given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{20} & \mathbf{0}_{20} & \mathbf{v} & \mathbf{0}_{20} \\ \mathbf{0}_{20} & \mathbf{1}_{20} & \mathbf{0}_{20} & \mathbf{v} \end{pmatrix},$$

where $\mathbf{v} = (1 \quad 2 \quad 3 \quad 4)' \otimes \mathbf{1}_5 = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad \cdots \quad 5)'$. The o.l.s. estimates are $\hat{\beta}_1 = 0.794(1.73)$, $\hat{\beta}_2 = -4.02(1.73)$, $\hat{\beta}_3 = 3.06(0.631)$, $\hat{\beta}_4 = 5.13(0.631)$, and $\hat{\sigma} = 3.15$, where the approximate standard errors based on (1.8) are given in parentheses, and $\mathsf{S}(\widehat{\boldsymbol{\beta}}) = 358.1$. Note that $\hat{\beta}_1 > \hat{\beta}_2$ and $\hat{\beta}_3 < \hat{\beta}_4$ as claimed. To test this, take

$$\mathbf{H} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad \mathbf{h} = \mathbf{0}, \tag{1.97}$$

and use (1.69) to get $\widehat{\boldsymbol{\gamma}} = (-1.61, -1.61, \ 4.09, \ 4.09)'$ and $\mathsf{S}(\widehat{\boldsymbol{\gamma}}) = 412.4$, so that $F = 2.7310$ from (1.88), with p-value 0.0787. Value $\mathsf{S}(\widehat{\boldsymbol{\gamma}})$ could also be obtained by noting that the reduced column space is given by $\mathbf{Z} = \begin{pmatrix} \mathbf{1}_{20} & \mathbf{v} \\ \mathbf{1}_{20} & \mathbf{v} \end{pmatrix}$.

The data used in the illustration were simulated using $\boldsymbol{\beta} = (0, -5, \ 3, \ 5)'$ and $\sigma = 3$, using the code in Listing 1.5. With these values, the noncentrality parameter in (1.94) is $\theta = \boldsymbol{\beta}'\mathbf{H}'[\mathbf{HAH}']^{-1}\mathbf{H}\boldsymbol{\beta}/\sigma^2 = 50/9$ from (1.82). Thus, with $c = F^{-1}_{J,T-k}(1-\alpha) = 3.26$ for $J = 2$, $T - k = 36$ and $\alpha = 0.05$, the power of the F test is 0.513, or not much better than flipping a fair coin. The reader is encouraged to construct a program to confirm this power via simulation. Observe this is trivially done based on the code in Listing 1.5, omitting the superfluous graphics commands and calculation of `num2` and `num3`. Based on

```
randn('state',2); % this is now deprecated in Matlab, but still works in version R2010a
cc=5; T=2*4*cc; % cc is cell count. So T is a multiple of 2*4
beta=[0 -5 3 5]';
dum1=[ones(T/2,1); zeros(T/2,1)]; dum2=1-dum1;
time=kron((1:4)',ones(cc,1)); c3=kron([1,0]',time); c4=kron([0,1]',time);
X=[dum1 dum2 c3 c4]; y=X*beta+3*randn(T,1);

figure
for i=1:T
  if X(i,1)==1, h1=plot(X(i,3),y(i),'go','linewidth',2); set(h1,'markersize',8)
  else h2=plot(X(i,4),y(i),'rx','linewidth',2); set(h2,'markersize',8), end
  hold on
end
hold off, set(gca,'XTick',1:4), set(gca,'fontsize',16)
ax=axis; axis([0.5 4.5 ax(3) ax(4)]), legend([h1,h2],'old','new',2)

A=inv(X'*X); betahat=A*X'*y; %#ok<*MINV?
yhat=X*betahat; res=y-yhat; Sbeta=sum(res.^2);
sig2hat=Sbeta/(T-4); sigma_hat = sqrt(sig2hat); H=[1 -1 0 0; 0 0 1 -1];
num1 = (H*betahat)'*inv(H*A*H')*(H*betahat) %#ok<*NOPTS>
F = num1 / 2 / sig2hat, pvalue = 1-fcdf(F,2,T-4)
gammahat = OLSrestrict(y,X,H); yhat=X*gammahat; res=y-yhat;
Sgamma=sum(res.^2); num2 = Sgamma - Sbeta
Z = [dum1 + dum2, c3 + c4]; A=inv(Z'*Z); bhat=A*Z'*y; yhat=Z*bhat;
res=y-yhat; Sb=sum(res.^2); num3 = Sb - Sbeta
```

Program Listing 1.5: Computes F statistic (1.88) and the corresponding p-value. Three ways of obtaining the numerator in (1.88) are computed: `num1` uses (1.90), `num2` computes $\widehat{\boldsymbol{\gamma}}$ and its associated residual sum of squares $\mathsf{S}(\widehat{\boldsymbol{\gamma}})$, and `num3` is computed based on the reduced column space given by matrix `Z` in the program. Function `OLSrestrict` is given in Listing 1.6 below.

```
function gamma = OLSrestrict(y,X,H,h)
[J,k]=size(H); if nargin<4, h=zeros(J,1); end
b=regress(y,X); A=inv(X'*X); gamma = b+A*H'*inv(H*A*H')*(h-H*b);
```

Program Listing 1.6: Called by the code in Listing 1.5 to compute $\widehat{\boldsymbol{\gamma}}$ from (1.69).

(a total overindulgence of) `sim` $= 10{,}000{,}000$ replications, the empirical power is, to three significant digits, the same, 0.513 (and, for $\alpha = 0.01$, is 0.265).

Problem 1.13 asks the reader to construct a simple program to calculate the minimum necessary sample size, T, to obtain a specified test size and power. For example, to get a power of 0.90 with $\alpha = 0.05$, T needs to be at least 96. Simulation with $T = 96$ confirms this, giving an (empirical) power of 0.906, as the reader should verify, and is 0.752 for $\alpha = 0.01$. ■

Example 1.12 ***Example 1.11 cont.***

We now wish to see how this regression would be conducted using the SAS system (with details of its basic use given in Appendix D). The first issue concerns getting the data into SAS. The simple Matlab code in Listing 1.7 outputs variables **y** and **X**, as were generated in Listing 1.5, to a text file, so that they can be, for example, read in by other programs, as we require here. In general, a bit of trial and error might be required with the `fprintf` command to get the desired format.

```
YX=[y,X]; fileID = fopen('coachingdata.txt','w');
fprintf(fileID,'%8.5g %1u %1u %1u %1u \r\n',YX'); fclose(fileID);
```

Program Listing 1.7: Outputs variables **y** and **X** generated in Listing 1.5 as a text file.

```
ods pdf file='Coaching Regression Output.pdf';
data coach;
  infile 'coachingdata.txt';
  input y X1-X4;
run;
proc reg data=coach;
  RestrictedModel: model y = X1-X4 / NOINT;
    restrict X1=X2, X3=X4;
  UnRestricted: model y = X1-X4 / NOINT;
    SameInterceptAndSlope: test X1=X2, X3=X4;
run;
ods _all_ close;
ods html;
```

SAS Program Listing 1.1: SAS statements for (i) reading the text data set produced from the Matlab output generated by the code in Listing 1.7, and (ii) performing a regression analysis of the restricted model and the unrestricted model, and, for the latter, conducting the F test for the restrictions in (1.97). The output is a report, as an Adobe portable document format (pdf), including several useful graphics.

Next, the code in SAS Listing 1.1 performs two regression analyses. The first is of the restricted model, where the `restrict` statement is used to indicate (in terms of the variable names associated with the **X** matrix, and not the β coefficients). The second is unrestricted, and performs the F test associated with the restriction we wish to test. Observe how the `NOINT` option is necessary to tell SAS not to include an intercept term (a column of ones) in the regression, which it otherwise does by default. The SAS output (not shown) for the test in (1.97) yields $F = 2.73$ with a p-value of 0.0787, agreeing with the values obtained above using manual calculations in Matlab. ■

A **time-series regression** is such that Y_t and $\mathbf{x}_t'$ correspond to time point t. For simplicity, assume that the time points for which observations are observed are equally spaced, so that $t = 1, \ldots, T$. A simple case is the model $Y_t = \beta_1 + \beta_2 t + \epsilon_t$. Examples of dependent variables that could be modelled as a time-series regression include:

1) Quarterly sales of a certain product, using regressors such as quarterly "dummy" variables, price, and amount of advertising, as well as prices and amounts of advertising for similar products offered from various market competitors.
2) Monthly rate of:
   a) fatalities caused by car accidents, using as regressors monthly dummies and/or dummies for particular days, such as weekend days or holidays
   b) alcohol-related car accidents
   c) homicides caused by guns.

3) Blood pressure of a patient, measured at weekly intervals, with regressors such as weight, number of cigarettes smoked, etc.

There are occasions in which the (linear) relationship describing a variable over time undergoes a pronounced change, due perhaps to the occurrence of a relevant and major event at some time point t_0, $1 \leqslant t_0 \leqslant T$.[8] In this case, the model is said to undergo a **structural break** at time t_0. Referring to the above dependent variables, examples of events that might cause a structural break include:

1) Discovery of a significant positive (or negative) side-effect from consuming the product.
2) Introduction of a new law for:
   a) the mandatory wearing of seat belts,
   b) the legal threshold of blood alcohol levels deemed acceptable to drive,
   c) gun control.
3) Change in diet, medication, etc.

If a structural break occurs, then two coefficient vectors need to be estimated: the first, say $\boldsymbol{\beta}_{[1]}$, for the sample of data corresponding to time points $1, \dots, t_0$, and the second, say $\boldsymbol{\beta}_{[2]}$, corresponding to $t_0 + 1, \dots, T$. We assume that σ^2 in both segments of time is constant. Such a model is said to be a **piecewise (linear) regression** if we constrain the two regression lines to touch at t_0, i.e., if $\mathbf{x}_{t_0}'\boldsymbol{\beta}_{[1]} = \mathbf{x}_{t_0}'\boldsymbol{\beta}_{[2]}$ is imposed. Point t_0 is said to be a **knot** or **join point**. The extension to more than one knot should be clear.[9]

Example 1.13 Let $Y_t = a_1 + a_2 t + e_t,\ t = 1, \dots, t_0,$ and $Y_t = b_1 + b_2 t + e_t,\ t = t_0 + 1, \dots, T,$ with $e_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), t = 1, \dots, T$.[10] Then, for the regression function to be continuous over the whole range, it must be the case that $a_1 + a_2 t_0 = b_1 + b_2 t_0$, or

$$a_1 - b_1 + a_2 t_0 - b_2 t_0 = 0. \tag{1.98}$$

Another way of stating this model is

$$\mathbf{Y} = a_1\mathbf{x}_1 + b_1\mathbf{x}_2 + a_2\mathbf{x}_3 + b_2\mathbf{x}_4 + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \mathbf{x}_4]$, with

$$\mathbf{x}_1 = (\mathbf{1}_{t_0}'\ \mathbf{0}_{T-t_0}')', \quad \mathbf{x}_2 = (\mathbf{0}_{t_0}'\ \mathbf{1}_{T-t_0}')',$$
$$\mathbf{x}_3 = (1, 2, \dots, t_0, 0, \dots, 0)', \quad \mathbf{x}_4 = (0, \dots, 0, t_0 + 1, \dots, T)',$$

and parameter vector $\boldsymbol{\beta} = (a_1, b_1, a_2, b_2)'$ is subject to the constraint $\mathbf{H}\boldsymbol{\beta} = 0$ from (1.98), where $\mathbf{H} = [1 \quad -1 \quad t_0 \quad -t_0]$. From (1.69), the restricted parameter vector is

$$\widehat{\boldsymbol{\gamma}} = \left(\mathbf{I}_4 - \frac{\mathbf{A}\mathbf{H}'\mathbf{H}}{\mathbf{H}\mathbf{A}\mathbf{H}'}\right)\widehat{\boldsymbol{\beta}},$$

---

8 In fact, such a phenomenon can occur in any type of data for which the order of the observations is relevant. Another example would be for spatial data, e.g., weather measurements taken simultaneously at different locations.

9 Less obvious, however, is how to proceed if the locations of the knots are not known. See, for example, Judge et al. (1985, pp. 800-814) for discussion of this and other related issues.

10 If for $t = t_0 + 1, \dots, T$, we take $Y_t = b_1 + b_2(t - t_0) + e_t$, which is sometimes referred to as a **locally disjoint broken trend model**, its first usage being from Perron and Zhu (2005); see also Deng and Perron (2006), Sobreira and Nunes (2016), Chang and Perron (2016), and the references therein.

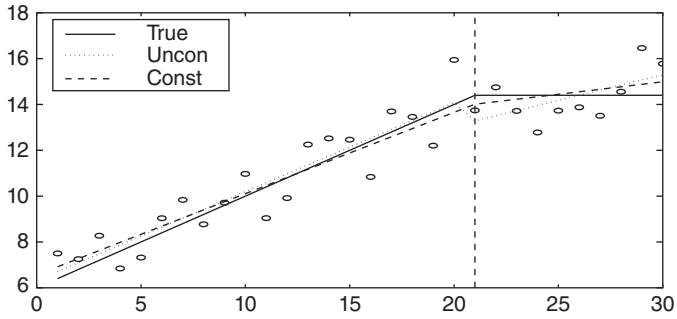

Figure 1.4 True and fitted piecewise regression.

where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$ and $\widehat{\boldsymbol{\beta}}$ is the unrestricted estimated parameter vector. It is worth emphasizing that the value of the F test (1.88), and, hence, its p-value, depends only on $\mathbf{H}\boldsymbol{\beta}$ and is otherwise invariant to the choice of $\boldsymbol{\beta}$.

Figure 1.4 shows a simulated sample using $T = 30$, $t_0 = 21$, $\sigma^2 = 1$ and parameter values $a_1 = 6$, $a_2 = 0.4$, $b_2 = 0$ and $b_1 = a_1 + t_0 a_2 - t_0 b_2 = 14.4$, so that (1.98) is satisfied.[11] The p-value of the F test for constraint (1.98) is 0.130, so that the null hypothesis of a knot would not be rejected at conventional testing levels. In addition, the hypothesis that only one regression line is needed, i.e., that $a_1 = a_2$ and $b_1 = b_2$, was tested and resulted in a p-value of 0.0318. The data and plot were generated with the code in Listing 1.8.

Finally, to test whether the slope changes at the knot, let the unrestricted model be $Y_t = \alpha_1 + \alpha_2 t + \alpha_3 (t - t_0) B_t + \epsilon_t$, $t = 1, \ldots, T$, where $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$ and B_t is a boolean (or dummy) variable that is one if $t \geqslant t_0$ and zero otherwise, i.e., $B_t = \mathbb{I}_{\{t_0, t_0+1, \ldots\}}(t)$. The null hypothesis is that $\alpha_3 = 0$, for which the reduced column space is easy to express. For the data used, the p-value was 0.0310. As the true model is piecewise, it comes as no surprise that this p-value is quite close to the p-value given above for testing $a_1 = a_2$ and $b_1 = b_2$. ■

1.4.7 Confidence Intervals

Recall from (1.88) and (1.90) that, under the null hypothesis that $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$,

$$\frac{(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})' \mathbf{V}^{-1} (\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{J \, \hat{\sigma}^2} \sim F_{J, T-k},$$

where $\mathbf{V} = \mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'$. This implies that

$$Q = \frac{(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta})}{J \, \hat{\sigma}^2} \sim F_{J, T-k} \tag{1.99}$$

11 The parameter values were chosen so that the data somewhat resemble actual data for rates of homicide in the USA, measured quarterly from 1985 to 1994, as shown in the Morbidity and Mortality Weekly Report from the Centers for Disease Control and Prevention (CDC), June 7, 1996, Vol. 45, No. 22, pp. 460–464. In their study, a piecewise linear regression was used to model the data.

```
function [pvalF1, pvalF2] = piecewise(seed,b2,doplot);
if nargin<2, b2=0.1; end, if nargin<3, doplot=1; end
t0=21; T=30; n=T-t0+1; x1=[ones(t0-1,1); zeros(n,1)]; x2=1-x1;
x3=[(1:t0-1)'; zeros(n,1)]; x4=[zeros(t0-1,1); (t0:T)']; X=[x1 x2 x3 x4];
a1=6; a2=0.4; b1=a1+a2*t0-b2*t0, beta=[a1 b1 a2 b2]'; sigma=1;
randn('state',seed); y=X*beta+sigma*randn(T,1); betahat=regress(y,X);
yfit=X*betahat; SSbeta=sum((y-yfit).^2); sigsqr_hat = SSbeta / (T-4);
% test the piecewise regression
H=[1 -1 t0 -t0]; J=1; gamma=OLSrestrict(y,X,H); yfitH=X*gamma;
SSgam=sum((y-yfitH).^2); F1 = (SSgam-SSbeta) / J / sigsqr_hat;
pvalF1 = 1-fcdf(F1,J,T-4);
if doplot==1
  true=X*beta;
  plot(1:T,true,'k-', 1:T,yfit,'g:', 1:T,yfitH,'r--', 1:T,y,'bo')
  set(gca,'fontsize',16), legend('True','Uncon','Const',2)
  ax=axis; h=line([t0 t0],[ax(3) ax(4)]); set(h,'linestyle','--')
end
% now test if both intercepts are equal and both slopes are equal
H=[1 -1 0 0; 0 0 1 -1]; J=2; gamma=OLSrestrict(y,X,H); yfitH=X*gamma;
SSgam=sum((y-yfitH).^2); F2 = (SSgam-SSbeta) / J / sigsqr_hat;
pvalF2=1-fcdf(F2,J,T-4);
```

Program Listing 1.8: Simulates and estimates a piecewise simple regression.

is a pivotal quantity for $\mathbf{H}\boldsymbol{\beta}$. In particular, letting $q = F^{-1}_{J,T-k}(1-\alpha)$ be the quantile such that $\Pr(Q \leqslant q) = 1-\alpha$, the ellipsoid $\{\mathbf{H}\boldsymbol{\beta} : Q \leqslant q\}$ is a $100(1-\alpha)\%$ confidence region for $\mathbf{H}\boldsymbol{\beta}$. If $J = 1$, then the region is just an interval.

Take, for example, the i.i.d. model: Let $Y_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, i.e., $\mathbf{X} = \mathbf{1}_n$ and $\boldsymbol{\beta} = \mu$, so that $\hat{\mu} = \bar{Y}$ and $Q = n(\hat{\mu}-\mu)^2/\hat{\sigma}^2 = (\hat{\mu}-\mu)^2/(S^2/n) \sim F_{1,n-1}$. Then, as $\sqrt{F^{-1}_{1,n-1}(1-\alpha)} = t^{-1}_{n-1}(1-\alpha/2)$, and from the symmetry of the Student's t distribution,

$$\{\mu : Q \leqslant q\} = \{\mu : |\hat{\mu}-\mu| \leqslant \sqrt{q}S/\sqrt{n}\} = (\hat{\mu} - \sqrt{q}S/\sqrt{n},\ \hat{\mu} + \sqrt{q}S/\sqrt{n})$$

is the usual confidence interval for μ. Similarly, for the general linear model with $J = 1$, $\mathbf{H}\boldsymbol{\beta}$ is a single linear combination of the elements in $\boldsymbol{\beta}$, which we denote $\boldsymbol{\ell}'\boldsymbol{\beta}$ for clarity, i.e., $\boldsymbol{\ell}' = \mathbf{H}'$. Then $\mathbf{V} = \boldsymbol{\ell}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell}$ is a scalar and, with $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$,

$$\left\{ \boldsymbol{\ell}'\boldsymbol{\beta} : \frac{(\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\ell}'\boldsymbol{\beta})^2}{\hat{\sigma}^2\, \boldsymbol{\ell}'\mathbf{A}\boldsymbol{\ell}} \leqslant q \right\} = \{\boldsymbol{\ell}'\boldsymbol{\beta} : |\boldsymbol{\ell}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\ell}'\boldsymbol{\beta}| \leqslant q^{1/2}\sqrt{\hat{\sigma}^2\, \boldsymbol{\ell}'\mathbf{A}\boldsymbol{\ell}}\} = \boldsymbol{\ell}'\widehat{\boldsymbol{\beta}} \pm c\sqrt{\hat{\sigma}^2\, \boldsymbol{\ell}'\mathbf{A}\boldsymbol{\ell}}, \tag{1.100}$$

where $c = t^{-1}_{T-k}(1-\alpha/2)$. For $J \geqslant 2$, $\{\mathbf{H}\boldsymbol{\beta} : Q \leqslant q\}$ cannot be so easily "pivoted" to get intervals for the rows of $\mathbf{H}\boldsymbol{\beta}$, but, if $J = 2$ or $J = 3$, the region can be plotted.

Example 1.14 Let $Y_t = \beta_1 + \beta_2 t + e_t$, $t = 1, \ldots, T$, $e_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$ and take $\mathbf{H} = \mathbf{I}_2$, so that the ellipsoid provides a confidence region for β_1 and β_2. For a simulated vector $\mathbf{Y}$ with $T = 10$, $\beta_1 = 1$, $\beta_2 = 2$, and $\sigma^2 = 1$, the region was computed with the program in Listing 1.9 and is shown in Figure 1.5 for the three common levels of significance $\alpha = 0.01$, 0.05, and 0.1. The relative size increase in going from $\alpha = 0.05$ to 0.01 is much larger than that from 0.1 to 0.05. ■

```
T=10; k=2; J=2; Y=1+2*(1:T)' + randn(T,1);
X=[ones(10,1),(1:10)'];
if 1==1, O=X; else O=orth(X); end
[betahat,BINT,R,RINT,STATS] = regress(Y,O,0.0001);
s2= sum(R.^2)/(T-k);
q90=finv(0.90,J,T-k); q95=finv(0.95,J,T-k); q99=finv(0.99,J,T-k);
Vi=(O'*O); % H is the 2X2 identity matrix
figure, h=plot(betahat(1),betahat(2),'k.'), set(h,'MarkerSize',30), hold on
inc=0.05;
for b1=BINT(1,1):inc:BINT(1,2)
  for b2=BINT(2,1):inc:BINT(2,2)
    beta=[b1 b2]'; Q=(betahat-beta)' * Vi * (betahat-beta) / (J*s2);
    if (Q <= q90), plot(b1,b2,'ro'), elseif (Q <= q95), plot(b1,b2,'gx')
    elseif (Q <= q99), plot(b1,b2,'b+'), end
  end
end, hold off
```

Program Listing 1.9: Generates ellipsoid for parameters of time-trend linear model. (Takes a relatively long to run; adjust `inc` accordingly.)

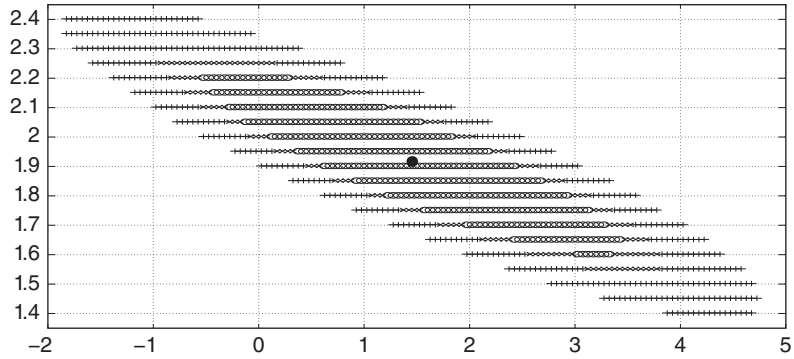

Figure 1.5 Ellipsoid for intercept β_1 (horizontal axis) and slope β_2 (vertical axis) for the model in Example 1.14, for $\alpha = 0.01$ (plus signs), $\alpha = 0.05$ (crosses) and $\alpha = 0.10$ (circles). The black dot is $\widehat{\boldsymbol{\beta}}$.

For $J = 3$, a three-dimensional plot of the region will be of limited use, while for $J \geqslant 4$, the whole region cannot be visualized as such, although one could plot it for two (or three) rows of $\mathbf{H}\boldsymbol{\beta}$ for fixed values of the remaining rows. This is clearly quite cumbersome and is essentially never done in practice. Instead, methods are used that yield simultaneous confidence intervals for each row of $\mathbf{H}\boldsymbol{\beta}$. One obvious way is to use Bonferroni's inequality as follows. Let $\hbar_i$ denote the ith row of $\mathbf{H}$, $i = 1, \ldots, J$. Then the confidence region for $\hbar_i\boldsymbol{\beta}$ is precisely that in (1.100) with $\hbar_i$ instead of $\boldsymbol{\ell}'$. For simultaneous confidence intervals on the J values of $\hbar_i\boldsymbol{\beta}$, the **Bonferroni method** just takes $c = t^{-1}_{T-k}(1 - \alpha/(2J))$. The obvious disadvantage of this method is the inevitable large size of the intervals when J is large. An approach that makes explicit use of the normality assumption (and results in shorter confidence intervals) is based on the multivariate t distribution and referred to as **maximum modulus t intervals**; see Graybill (1976, Sec. 6.6) for further details.

We now consider another alternative to the Bonferroni intervals known as the S-method or **Scheffé's method**, from Scheffé (1953). We first need the following result: If $\mathbf{V} > 0$ (i.e., positive definite), and $\boldsymbol{\ell}$ and $\mathbf{b}$ are conformable vectors such that $\boldsymbol{\ell}'\mathbf{b}$ is a scalar, then

$$\max_{\boldsymbol{\ell}\neq\mathbf{0}} \frac{(\boldsymbol{\ell}'\mathbf{b})^2}{\boldsymbol{\ell}'\mathbf{V}\boldsymbol{\ell}} = \mathbf{b}'\mathbf{V}^{-1}\mathbf{b}. \tag{1.101}$$

Proof: First observe that, as matrix $\mathbf{V}$ enters only via a quadratic form, it can be assumed symmetric without loss of generality, and thus it makes sense to state that $\mathbf{V} > 0$, as all its eigenvalues are real. Take symmetric $\mathbf{V}^{1/2} > 0$ such that $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{V}$ and define $\mathbf{u} = \mathbf{V}^{1/2}\boldsymbol{\ell}$ and $\mathbf{w} = \mathbf{V}^{-1/2}\mathbf{b}$, so that

$$\frac{(\boldsymbol{\ell}'\mathbf{b})^2}{\boldsymbol{\ell}'\mathbf{V}\boldsymbol{\ell}} = \frac{(\mathbf{u}'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{w})^2}{\mathbf{u}'\mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2}\mathbf{u}} = \frac{(\mathbf{u}'\mathbf{w})^2}{\mathbf{u}'\mathbf{u}} = \frac{\langle\mathbf{u},\mathbf{w}\rangle^2}{\|\mathbf{u}\|^2}.$$

From the Cauchy–Schwarz inequality (see Problem 1.7), $\langle\mathbf{u},\mathbf{w}\rangle^2 \leqslant \|\mathbf{u}\|^2\|\mathbf{w}\|^2$, with equality when $\mathbf{u} = \mathbf{w}$, i.e., $\mathbf{V}^{1/2}\boldsymbol{\ell} = \mathbf{V}^{-1/2}\mathbf{b}$ or $\boldsymbol{\ell} = \mathbf{V}^{-1}\mathbf{b}$. Thus, with $\boldsymbol{\ell} = \mathbf{V}^{-1}\mathbf{b}$,

$$\frac{\langle\mathbf{u},\mathbf{w}\rangle^2}{\|\mathbf{u}\|^2} = \|\mathbf{w}\|^2 = \|\mathbf{V}^{-1/2}\mathbf{b}\|^2 = \mathbf{b}'\mathbf{V}^{-1}\mathbf{b},$$

which is (1.101). See Graybill (1976, pp. 224–225) for an alternative proof. ■

Now, with $\mathbf{V} = \mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'$, $\boldsymbol{\theta} = \mathbf{H}\boldsymbol{\beta}$ and $\mathbf{b} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, (1.99) and (1.101) imply

$$\begin{aligned} 1-\alpha &= \Pr(Q \leqslant q) = \Pr((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{V}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leqslant Jq\hat{\sigma}^2) \\ &= \Pr\left(\max_{\boldsymbol{\ell}\neq\mathbf{0}} \frac{(\boldsymbol{\ell}'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))^2}{\boldsymbol{\ell}'\mathbf{V}\boldsymbol{\ell}} \leqslant Jq\hat{\sigma}^2\right) = \Pr(|\boldsymbol{\ell}'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})| \leqslant \sqrt{Jq\hat{\sigma}^2\boldsymbol{\ell}'\mathbf{V}\boldsymbol{\ell}},\ \forall\boldsymbol{\ell} \neq \mathbf{0}), \end{aligned}$$

where, as before, $q = F^{-1}_{J,T-k}(1-\alpha)$. That is, $\boldsymbol{\ell}'\widehat{\boldsymbol{\theta}} \pm \sqrt{Jq\hat{\sigma}^2\boldsymbol{\ell}'\mathbf{V}\boldsymbol{\ell}}$ simultaneously covers $\boldsymbol{\ell}'\boldsymbol{\theta}$ for an infinite set of vectors $\boldsymbol{\ell} \neq \mathbf{0}$ with level of significance $1-\alpha$. An alternative proof of this result using only basic calculus is given in Klotz (1969) and Roussas (1997, Sec. 17.4).

As only a finite number of such intervals will ever be constructed for a particular data set, the actual level exceeds $1-\alpha$. In particular, with $\boldsymbol{\ell}_i = (0,\ldots,0,1,0\ldots,0)'$ with the one in the ith position, $i = 1,\ldots,J$, $\boldsymbol{\ell}_i'\widehat{\boldsymbol{\theta}} = \boldsymbol{\ell}_i'\mathbf{H}\widehat{\boldsymbol{\beta}} = \hbar_i\widehat{\boldsymbol{\beta}}$, so that the J intervals $\hbar_i\widehat{\boldsymbol{\beta}} \pm \sqrt{Jq\hat{\sigma}^2\boldsymbol{\ell}_i'\mathbf{V}\boldsymbol{\ell}_i}$ have simultaneous level of significance at least $1-\alpha$. As

$$\widehat{\mathbb{V}}(\hbar_i\widehat{\boldsymbol{\beta}}) = \hat{\sigma}^2\hbar_i\mathbf{A}\hbar_i' = \hat{\sigma}^2\boldsymbol{\ell}_i'\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\boldsymbol{\ell}_i = \hat{\sigma}^2\boldsymbol{\ell}_i'\mathbf{V}\boldsymbol{\ell}_i,$$

these intervals are often written as $\hbar_i\widehat{\boldsymbol{\beta}} \pm \sqrt{Jq\widehat{\mathbb{V}}(\hbar_i\widehat{\boldsymbol{\beta}})}$, $i = 1,\ldots,J$.

Example 1.15 Consider the same setup as in Example 1.14, with

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{15}\begin{bmatrix} 7 & -1 \\ -1 & 2/11 \end{bmatrix}$$

and $\mathbf{H} = \mathbf{I}_2$. Let $\boldsymbol{\ell}_1 = (1, 0)'$, $\boldsymbol{\ell}_2 = (0, 1)'$, $a_1 = \boldsymbol{\ell}_1' \mathbf{A} \boldsymbol{\ell}_1 = 7/15$ and $a_2 = \boldsymbol{\ell}_2' \mathbf{A} \boldsymbol{\ell}_2 = 2/165$. Then, with $J = 2$, $c = t_8^{-1}(1 - 0.05/4) \approx 2.7515$, the simultaneous 95% Bonferroni confidence intervals for β_1 and β_2 are $\beta_i \pm c\hat{\sigma}\sqrt{a_i}$, $i = 1, 2$, with lengths $3.759\hat{\sigma}$ and $0.6059\hat{\sigma}$, respectively. With $J = k = 2$ and $q = F_{2,8}^{-1}(0.95) \approx 4.459$, the S-method confidence intervals are $\beta_i \pm \hat{\sigma}\sqrt{2qa_i}$, $i = 1, 2$, with respective lengths $4.080\hat{\sigma}$ and $0.6576\hat{\sigma}$. The latter are about 8.5% longer than Bonferroni confidence intervals. ■

Remark In the previous example, the S-method intervals were longer than those from Bonferroni. To compare the lengths for other parameters, the top panel of Figure 1.6 plots the ratio of $t_{T-k}^{-1}(1 - \alpha/2J)$ to $\sqrt{JF_{J,T-k}^{-1}(1 - \alpha)}$ as a function of J, using $T - k = 40$ and three values of α. It would appear that the S-method is virtually useless compared to Bonferroni. This picture is misleading, however, because k or, more generally, the rank of $\mathbf{H}$ was not specified. In particular, with $\hbar_i$ the ith row of $\mathbf{H}$, assume $\hbar_1, \ldots, \hbar_R$ are independent, $R \leqslant k$, and the remaining rows, $\hbar_{R+1}, \ldots, \hbar_J$, are linear combinations of $\hbar_1, \ldots, \hbar_R$. Let $\mathbf{H}^* = (\hbar_1', \ldots, \hbar_R')'$ be the upper $R \times k$ portion of $\mathbf{H}$, so that

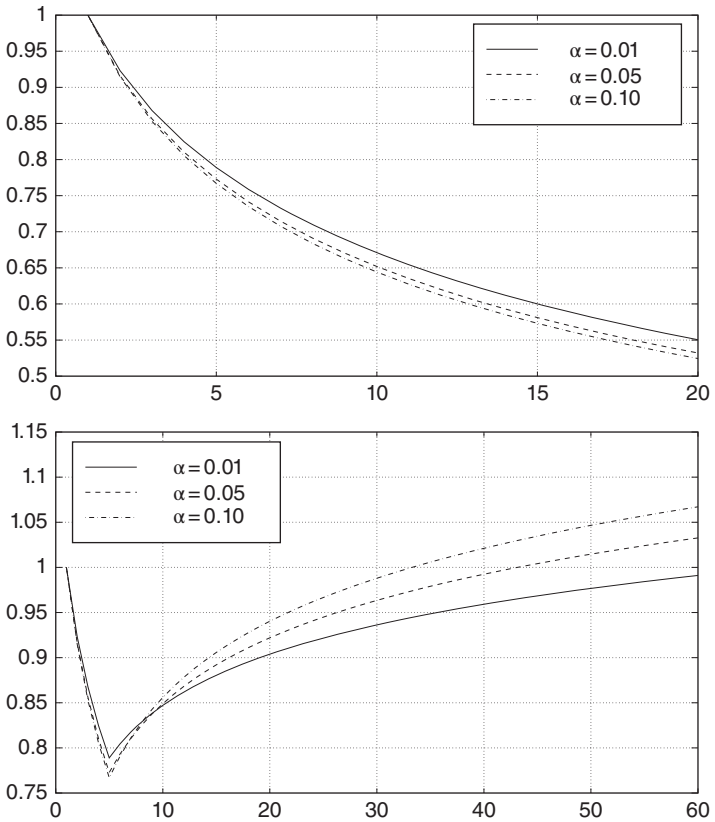

Figure 1.6 Ratio of lengths of Bonferroni to Scheffé confidence intervals. The top panel does not adjust for rank of **H**, while the bottom panel does adjust.

$\text{rank}(\mathbf{H}) = \text{rank}(\mathbf{H}^*) = R$. Then, with $\theta^* = (\theta_1^*, \dots, \theta_R^*)' = \mathbf{H}^* \boldsymbol{\beta}$, the S-method implies that

$$1 - \alpha = \Pr(|\boldsymbol{\ell}'(\hat{\theta}^* - \theta^*)| \leqslant \sqrt{Rq\hat{\sigma}^2 \boldsymbol{\ell}' \mathbf{V}^* \boldsymbol{\ell}},\ \forall \boldsymbol{\ell} \in \mathbb{R}^R \setminus \mathbf{0}), \tag{1.102}$$

where $q = F_{R,T-k}^{-1}(1-\alpha)$ and $\mathbf{V}^* = \mathbf{H}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}^{*\prime}$. But, by construction, each row $\hbar_i$ can be written as $\boldsymbol{\ell}_i'\mathbf{H}^*$ for some $\boldsymbol{\ell}_i \in \mathbb{R}^R \setminus \mathbf{0}$, so that (1.102) also includes the intervals for $\theta_{R+1}, \dots \theta_J$.[12] To see the effect this has, the right side of Figure 1.6 plots the ratio $t_{T-k}^{-1}(1-\alpha/2J)$ to $\sqrt{mF_{m,T-k}^{-1}(1-\alpha)}$ versus J, where $m = \min(J, k)$, $k = 5$ and, as before, $T - k = 40$. In this case, $\mathbf{H}^* = \mathbf{I}_k$. Indeed, if a relatively large number of intervals are to be computed, the S-method can be superior. ■

In most realistic cases, the S-method gives rise to the longest intervals. Their additional length is the price to pay to be able to simultaneously construct infinitely many of them. In practice, their use allows a certain extent of "data mining", i.e., the researcher can keep computing intervals of interest until something "significant" is found, and still claim validity of the procedure. Preferably, however, one has a particular set of intervals in mind before the data are collected, to which the Bonferroni method (or others) can be applied.

Further details on confidence intervals can be found in numerous books on regression, including Ravishanker and Dey (2002, Sec. 7.3), Seber and Lee (2003, Ch. 5), and Khuri (2010, Ch. 7).

1.5 Alternative Residual Calculation

Recall from (1.60) that $\widehat{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{M})$. Not only is $\mathbf{M}$ rank deficient, but the fact that the regression residuals are dependent on the $\mathbf{X}$ matrix implies that the distribution of common test statistics based on $\widehat{\epsilon}$, often ratios of quadratic forms, cannot be tabulated. This has historically been quite an inconvenience, though it should not be an issue now with modern computing power and the computational methods discussed in Section A.3. Perhaps the most popular example of a statistic whose use had been hampered by this fact (in the 1950s and 1960s) is the Durbin–Watson test D for detecting serial autocorrelation in the residuals; see Section 5.3.4. This was among the motivations for research on regression residuals that are independent of the regressor matrix.

Before proceeding, a comment on the relevance of this material is perhaps in order. In addition to being of historical importance for the reason just mentioned, we will also remark below that the recursive residuals are a special case of the ubiquitous and highly important Kalman filter. Next, as a theoretical curiosity, the derivation of the (below defined) BLUS and recursive residuals is instructive and, while arguably straightforward (especially after one sees the answer), is a great example of statistical mathematical ingenuity. Their practical relevance in some 21st century applications is admittedly less, such as in a machine-learning context and/or where large dimensional models are used, with mean terms being simply "regressed off" as part of a larger paradigm (see Section 11.2.2 for one such

12 Linear combinations of vectors are usually expressed in column form when using matrices. In this case,

$$\begin{pmatrix} | \\ \hbar_i' \\ | \end{pmatrix} = \ell_{i1} \begin{pmatrix} | \\ \hbar_1' \\ | \end{pmatrix} + \cdots + \ell_{iR} \begin{pmatrix} | \\ \hbar_R' \\ | \end{pmatrix} = \mathbf{H}^{*\prime} \begin{pmatrix} \ell_{i1} \\ \vdots \\ \ell_{iR} \end{pmatrix} = \mathbf{H}^{*\prime} \boldsymbol{\ell}_i, \quad i = 1, \dots, J,$$

or, taking transposes, $\hbar_i = \boldsymbol{\ell}_i'\mathbf{H}^*$.

example). As such, we illustrate the main concepts here, and place further details in Appendices 1.A and 1.B as optional reading for those interested in the proverbial "full Monty".

Several estimators of the regression residuals have been proposed, each sharing the three properties of linearity, unbiasedness, and a scalar covariance matrix; these are typically abbreviated with the acronym LUS. We denote such residuals by $\widehat{\boldsymbol{\epsilon}}_{\text{LUS}} = \mathbf{CY}$, where $\mathbf{C}$ is a nonstochastic matrix (it can depend on $\mathbf{X}$, but not on $\mathbf{Y}$) satisfying $\mathbf{CX} = \mathbf{0}$ and $\mathbf{CC}' = \mathbf{I}$. Clearly, $\widehat{\boldsymbol{\epsilon}}_{\text{LUS}} = \mathbf{CY}$ is linear in $\mathbf{Y}$, and as

$$\mathbb{E}[\widehat{\boldsymbol{\epsilon}}_{\text{LUS}}] = \mathbb{E}[\mathbf{CY}] = \mathbb{E}[\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\epsilon}] = \mathbf{CX}\boldsymbol{\beta},$$

we see that the requirement $\mathbf{CX} = \mathbf{0}$ is necessary for unbiasedness. If $\mathbf{CC}' = \mathbf{I}$, then

$$\mathbb{E}[\widehat{\boldsymbol{\epsilon}}_{\text{LUS}}\widehat{\boldsymbol{\epsilon}}'_{\text{LUS}}] = \mathbb{E}[\mathbf{C}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{C}'] = \sigma^2\mathbf{CC}' = \sigma^2\mathbf{I},$$

so that $\widehat{\boldsymbol{\epsilon}}_{\text{LUS}}$ has a scalar covariance matrix.

Observe that the requirements $\mathbf{CX} = \mathbf{0}$ and $\mathbf{CC}' = \mathbf{I}$ (which is full rank) together imply that $\mathbf{C}$ cannot be $T \times T$, but rather $(T-k) \times T$, so that $\mathbf{CC}' = \mathbf{I}_{T-k}$ and $\widehat{\boldsymbol{\epsilon}}_{\text{LUS}} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I}_{T-k})$. In particular, the rows of $\mathbf{C}$ are orthogonal to the columns of $\mathbf{X}$, i.e., they are contained in $C(\mathbf{X})^{\perp}$, which has dimension $T-k$. Thus $\mathbf{CC}' = \mathbf{I} \iff$ the rows of $\mathbf{C}$ are orthogonal to one another $\iff$ there are at most $T-k$ rows in $\mathbf{C}$. Thus, only $T-k$ LUS residuals can be identified.

There are numerous matrices $\mathbf{C}$ that satisfy the LUS properties, and a "best" criteria was desired. This was pursued by Theil (1965, 1968) and Koerts (1967), and detailed in the books from Theil (1971) and Koerts and Abrahamse (1969). Consider the partition of the model

$$\begin{bmatrix} \mathbf{Y}_0 \\ \mathbf{Y}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_0 \\ \boldsymbol{\epsilon}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \end{bmatrix} \widehat{\boldsymbol{\beta}}_{\text{LS}} + \begin{bmatrix} \mathbf{e}_0 \\ \mathbf{e}_1 \end{bmatrix}, \tag{1.103}$$

where the quantities indexed with 0 have k rows and the quantities indexed with 1 contain the remaining $T-k$ rows. The vector $\boldsymbol{\epsilon}_0$ contains the k errors not represented in the LUS estimator. Given this partitioning, the best LUS, or **BLUS residuals**, denoted by $\widehat{\boldsymbol{\epsilon}}_{\text{BLUS}}$, are defined as the vector of residuals among the class of LUS residuals that has the minimum expected sum of squared errors, i.e., the vector that minimizes

$$\mathbb{E}[(\widehat{\boldsymbol{\epsilon}}_{\text{LUS}} - \boldsymbol{\epsilon}_1)'(\widehat{\boldsymbol{\epsilon}}_{\text{LUS}} - \boldsymbol{\epsilon}_1)].$$

Some work is required to show that the vector of BLUS residuals can be expressed in the computationally attractive form

$$\widehat{\boldsymbol{\epsilon}}_{\text{BLUS}} = \mathbf{e}_1 - \mathbf{X}_1\mathbf{X}_0^{-1}\left[\sum_{h=1}^{H} \frac{d_h}{1+d_h}\mathbf{q}_h\mathbf{q}_h'\right]\mathbf{e}_0, \tag{1.104}$$

where $d_1^2, \ldots, d_H^2$ are the eigenvalues of the matrix $\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'$ that are less than one, $H \leqslant T-k$, and $\mathbf{q}_1, \ldots, \mathbf{q}_H$ are the corresponding eigenvectors. A detailed derivation is given in Appendix 1.A.

Furthermore, the $(T-k) \times T$ matrix $\mathbf{C}$ in this case is given by the partitioned matrix $\mathbf{C} = [\mathbf{C}_0 \ \mathbf{C}_1]$ where the $(T-k) \times k$ matrix $\mathbf{C}_0$ and the $(T-k) \times (T-k)$ matrix $\mathbf{C}_1$ are derived by the following relationships:

$$\mathbf{C}_0 = -\mathbf{C}_1\mathbf{Z}, \quad \mathbf{C}_1 = \mathbf{PDP}',$$

where $\mathbf{Z} = \mathbf{X}_1\mathbf{X}_0^{-1}$, $\mathbf{D}$ is the $(T-k) \times (T-k)$ diagonal matrix whose first H successive diagonal elements are $d_1 \leqslant d_2 \leqslant \ldots \leqslant d_H < 1$ (the ds being the positive square roots of the d_k^2 defined in (1.104)),

```
function C = blusmat(X)
[T,k]=size(X); X_0 = X(1:k,:); X_1 = X(k+1:end,:);
Z = X_1*inv(X_0); D = eig(X_0*inv(X'*X)*X_0');
index1 = find(D<1 & D>0); H = size(index1,1);
D = [D(index1);ones(T-k-H,1)]; D = sort(D); D = diag(D);
[P tempD] = eig(eye(T-k) + Z*Z');
tempD = diag(tempD); [tempD index2] = sortrows(tempD);
P = P(:,index2(end:-1:1)); C_1 = P*D*P'; C = [-C_1*Z C_1];
```

Program Listing 1.10: Constructs the BLUS residual matrix **C**.

and **P** is the $(T-k)\times(T-k)$ orthogonal matrix with columns given by the eigenvectors of $\mathbf{I}+\mathbf{ZZ}'$ corresponding to the eigenvalues $1/d_1^2,\ldots,1/d_H^2,1,\ldots,1$; see Appendix 1.A. The code in Listing 1.10 computes matrix **C**.

One particular LUS residual estimator, the so-called **recursive residuals**, introduced by Hedayat and Robson (1970), Harvey and Phillips (1974), and Brown et al. (1975), is noteworthy. (Their use can be traced back all the way to Gauss; see Plackett, 1950; Stigler, 1981; and Young, 2011.) The procedure is computationally simple and turns out to be a special case of the Kalman filter; see the remarks in Section 5.6.

Phillips and Harvey (1974) show that the corresponding **C** matrix such that $\mathbf{V}=\mathbf{CY}$ and $\mathbf{V}\sim \mathrm{N}(\mathbf{0},\sigma^2\mathbf{I}_{T-k})$ can be expressed as

$$\mathbf{C}=\begin{pmatrix} \mathbf{a}_{k+1} & d_{k+1}^{-1/2} & 0 & \cdots & 0 \\ \mathbf{a}_{k+2} & & d_{k+2}^{-1/2} & & \vdots \\ \vdots & & & \ddots & \\ & & & & 0 \\ \mathbf{a}_T & & & & d_T^{-1/2} \end{pmatrix}, \tag{1.105}$$

of size $(T-k)\times T$, where, for $j=k+1,\ldots,T$,

$$\mathbf{a}_j=-d_j^{-1/2}\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\quad \mathbf{X}_{j-1}',\quad d_j=1+\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j, \tag{1.106}$$

and $\mathbf{x}_j'$ is the jth row of **X**. Note that $\mathbf{a}_j$ is a row vector with length $j-1$.

Direct multiplication verifies that $\mathbf{CX}=\mathbf{0}$ and $\mathbf{CC}'=\mathbf{I}_{T-k}$, and one may show (Theil, 1971, p. 209) that $\mathbf{C}'\mathbf{C}=\mathbf{M}$. Thus, in Theorem 1.3 above, one could take **G** to be **C**. The program in Listing 1.11 computes (1.105). Appendix 1.B provides details on the derivation of the recursive residuals.

```
function C = recmat(X)
[T,k]=size(X); C=zeros(T,T);
for j=(k+1):T
  mid=inv (X(1:(j-1),:)' * X(1:(j-1),:));
  d=sqrt (1+X(j,:) * mid * X(j,:)');
  p2=mid * X(1:(j-1),:)'; v=-(X(j,:) * p2)/d;
  C(j,1:(j-1))=v; C(j,j)=1/d;
end
C=C((k+1):T,:);
```

Program Listing 1.11: Constructs the recursive residual matrix **C**.

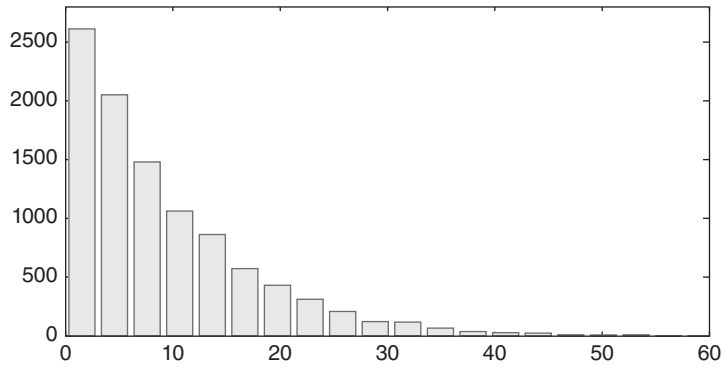

Figure 1.7 Simulated relative percentage change between the recursive and BLUS residuals for a model with intercept and time trend, and 20 observations.

Example 1.16 We wish to compare the magnitudes of the sum of squared BLUS and recursive residuals. Take the model to be $Y_j = 1 + 2j + e_j, j = 1, \ldots, 20$, with $e_j \overset{\text{i.i.d.}}{\sim} \text{N}(0, 1)$, so that the $\mathbf{X}$ matrix consists of a constant and a time vector. By using the code in Listings 1.10 and 1.11, it is a very simple Matlab exercise to simulate the model a large number of times and, for each, compute the relative percentage change between the recursive and BLUS residuals (i.e., $100 * (r - b)/r$, where r and b denote the sum of squares of the recursive and BLUS residuals, respectively).

Doing this for 10,000 replications and plotting the resulting histogram results in Figure 1.7. Note that, in every case, the sum of squared BLUS residuals is smaller than that for the recursive, as the theory dictates. Based on the simulation, there is more than a 35% chance that the relative percentage change will be more than 10%. ■

Remarks

a) Statistical tests common with the linear model using the BLUS residuals do not necessarily possess greater power than those using the "usual" o.l.s. residuals, or some other $\mathbf{C}$. The use of BLUS residuals has faded considerably since the 1970s, although more recently Magnus and Sinha (2005) conducted studies comparing the power of BLUS against the recursive residuals when testing against heteroskedasticity (one of the original motivations for BLUS) and structural breaks (for which the recursive residuals are intuitively appealing). The reported simulation results lend mild support for the use of BLUS residuals over recursive residuals.

b) We will see later that the recursive residuals (or any LUS estimate) have other desirable properties that make their use valuable. In particular, in the context of time-series analysis, Chapter 8 will show that, for any $\mathbf{X}$ matrix, the coefficients of the sample autocorrelation function (SACF) based on the recursive residuals always have zero expectation and are symmetric, a property not shared by the SACF based on the usual o.l.s. residuals, even when $\mathbf{X}$ is only a column of ones. This is important because, in practice, the SACF coefficients are compared to their limiting distribution, which is normal (i.e., symmetric) with zero mean. For small samples and $\mathbf{X}$ matrices common in econometric applications, this can be an important factor. ■

1.6 Further Topics

> As it happens, the econometric modeling was done in the basement of the building and the econometric theory courses were taught on the top floor (the third). I was perplexed by the fact that the same language was used in both places. Even more amazing was the transmogrification of particular individuals who wantonly sinned in the basement and metamorphosed into the highest of high priests as they ascended to the third floor.
>
> (Edward Leamer, 1978, p. vi)

> With increasing interest in the stable distributions and their domains of attraction, the Cauchy distribution is found to occupy a less isolated position; indeed the normal distribution is extremal and rather special among stable distributions.
>
> (E. J. Pitman and E. J. Williams, 1967, p. 916)

> What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a corresponding incomplete picture for a set of distributions.
>
> (Frederick Mosteller and John W. Tukey, 1977, p. 266)

An important special case of the linear model is the so-called analysis of variance, or ANOVA, for fixed and random effects, as introduced in Chapters 2 and 3, respectively. However, as these chapters are aimed at the underlying distribution theory of the core linear regression model and the ANOVA setting, numerous important topics associated with regression are regretfully not discussed. Two obvious ones are its extension to a multivariate framework, such as MANOVA and discriminant analysis (see, e.g., Huberty and Olejnik, 2006) and the use of Bayesian inferential methods (see, e.g., Christensen et al., 2011 and Gelman et al., 2013). Here, we mention several other omitted topics associated with regression analysis, albeit without much detail, so that the reader is at least aware of them, and provide useful references for further reading.

1) **Forecasting**.
Based on regression model (1.3), interest might center on predicting the random variable Y_{T+1} for a given $\mathbf{x}_{T+1} = (x_{T+1,1}, \ldots, x_{T+1,k})'$, so that $Y_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta} + \epsilon_{T+1}$, where $\epsilon_{T+1} \sim \mathrm{N}(0, \sigma^2)$. As $\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators for $\boldsymbol{\beta}$, the minimum variance unbiased point estimator is $\widehat{Y}_{T+1} = \mathbf{x}'_{T+1}\widehat{\boldsymbol{\beta}}$, and, from (1.8),

$$\mathbb{V}(\widehat{Y}_{T+1} - Y_{T+1}) = \mathbb{V}(\widehat{Y}_{T+1}) + \mathbb{V}(Y_{T+1}) = \sigma^2\mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{T+1} + \sigma^2.$$

Thus, an exact $100(1-\alpha)\%$ confidence interval for Y_{T+1} is

$$\widehat{Y}_{T+1} \pm c\hat{\sigma}\sqrt{1 + \mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{T+1}}, \tag{1.107}$$

where $\hat{\sigma}^2$ is given in (1.11), and c is the $\alpha/2$ quantile of a Student's t random variable with $T-k$ degrees of freedom.

The reader is encouraged to set up the parametric and nonparametric bootstrap to generate confidence intervals for Y_{T+1} for both the Gaussian and non-Gaussian cases. Under the normality assumption, simulation can be used to confirm that the bootstrap results are comparable to the analytic method in (1.107). For a non-Gaussian, leptokurtic, and asymmetric distributional assumption, confidence intervals (hereafter c.i.s) based on (1.107) (i) will almost surely be such that the actual and nominal coverage probabilities are not equal, and (ii) restricted to being incorrectly symmetric. Bootstrap c.i.s are expected to be be more accurate, particularly as the level of non-Gaussianity increases.

Further details on multiple prediction intervals making use of the methods in Section 1.4.7 can be found in, e.g., Seber and Lee (2003, Sec. 5.3) and Rao et al. (2008, Ch. 6).

2) **Multicollinearity**.
Particularly in the social sciences, some regressors can be highly correlated with one another, and give rise to what is called multicollinearity. With very high correlation, the resulting standard errors on the coefficients are large, and thus the point estimates are rather imprecise. Several ways of dealing with this issue exist, including use of shrinkage (recall Section III.5.4), empirical Bayes estimators, **ridge regression** (which is related to the former two methods), and use of (generalized) cross validation.

Further methods that also relate more generally to model specification and estimation are the so-called garrote and LASSO estimators. The LASSO and ridge regression are generalized by the so-called **elastic net**. These tools are important for dimension reduction, variable selection, and improved predictive performance when modeling high-dimensional (big) data. Their respective Wikipedia entries are a good starting point and include original references, while further information can be found in textbook presentations such as Seber and Lee (2003, Sec. 12.5), Murphy (2012), Fahrmeir et al. (2013, Sec. 4.2), and Efron and Hastie (2016, Ch. 7, 12, 16). See also Lansangan and Barrios (2017) and the references therein for an introduction, further methods, and comparisons among them.

3) The choice of regressors, or, more generally, **model specification**.
Recall the reference to Leamer (1983) in Section 1.1, indicating the potentially severe implications resulting from the choice of variables to include in a regression. The tidy, impressive analytic results and distribution theory throughout this chapter are child's play (and arguably of secondary relevance) compared to the much thornier issue of model specification with real data, particularly from the social sciences. The quote by Magnus (2017) at the beginning of Section 1.4 serves to remind us that inspection of the "t-statistics" is not a viable method for model selection (in general agreement with the diatribe in Section III.2.8), and Magnus (2017, Sec. 2.14, 2.15) provides a very readable presentation of the bias/variance tradeoff associated with including a particular regressor into the model. The amusing quote by Leamer (1978) at the beginning of this section might be a reflection of the state of affairs during what might now appear to be a primordial age of econometrics, though it still contains more than just a grain of truth on the discrepancies between theory and practice.

As mentioned, model selection is related to multicollinearity—it might be preferred to simply omit regressors that are highly correlated with others. The inherent difficulty in establishing the "best" model is nicely stated in Seber and Lee (2003, p. 424): "The relative merits of ridge regression versus least squares and subset selection have been endlessly debated." Textbooks

on regression analysis present many of the numerous ways that have been devised to select an optimal set (in some sense) from an available pool of regressors. See, e.g., the relevant chapters in Graybill and Iyer (1994), Ravishanker and Dey (2002), Seber and Lee (2003), Christensen (2011), Montgomery et al. (2012), Chatterjee and Hadi (2012), and Harrell, Jr. (2015).

Those books also cover numerous additional topics associated with applied regression analysis, and make use of real-data examples.

Particularly in econometrics, an influential body of work and methodology centers around the influential David F. Hendry, sometimes referred to general-to-specific (GETS) modeling, or the "LSE (London School of Economics) approach (to econometrics)" (see the same-titled Wikipedia entry). Good starting points include Hendry (1995, 2009), Castle et al. (2011), Hendry and Doornik (2014), and Castle et al. (2017).

4) **Missing values**.
   It is not uncommon that one or more entries of the desired regressor matrix $\mathbf{X}$ are missing. A good starting point for methods of dealing with this important issue in the context of regression is Rao et al. (2008, Ch. 8). In a more general setting, analysis of data with missing values is addressed by so-called **multiple imputation**, often using simulation and, when applicable, an expectation-maximization (hereafter EM) algorithm. An internet search for books along the lines of "multiple imputation of missing data" will reveal numerous possible resources for addressing this common and pernicious issue when dealing with real data.
5) **Time-varying parameters**, such that one or more of the regression coefficients varies through time.
   We deal with some aspects of this in Section 5.6. Consideration of such models leads naturally to the more general class of so-called state space models; see the references in Section 5.6.
6) One or more of the regression coefficients undergoes a **structural break**, i.e., a change in its value at some unknown point in time.
   Estimation and testing in this case has been considered by numerous authors; see, e.g., Bai and Perron (1998, 2003), Qu and Perron (2007), Yamamoto and Perron (2013).[13] Another method is via **impulse indicator saturation**, as first investigated by Hendry (1999). It provides a general test for an unknown number of breaks, at unknown times, and is applicable in many model situations besides the linear regression model, such as vector autoregressions; see, e.g., Ericsson (2012), Castle et al. (2015), and the references therein for further development and application. It also has applications to testing for parameter constancy; see, e.g., Johansen and Nielsen (2009), Hendry and Doornik (2014), and the references therein. A package for R is available from Sucarrat et al. (2017) for automated GETS modeling of the mean and variance of a regression, and indicator saturation methods for detecting and testing for structural breaks in the mean.
7) Use of **robust estimators**.
   In the presence of outliers, the least squares estimator is not optimal. Alternative estimation procedures have been developed to address this, e.g., Seber and Lee (2003, Sec. 3.13), Andersen (2008), and Huber and Ronchetti (2009, Ch. 7), as well as the note below on quantile regression.
8) **Partially adaptive estimation** for regression amid non-Gaussian disturbances.
   This is related to the previous issue of robustness, but in that setting the assumption is that the disturbances are Gaussian, but such that one or more observations deviates substantially from

13 The authors conveniently provide Matlab codes for this last test, and others; see Perron's web page: http://people.bu.edu/perron/code.html.

the main group. Here, the assumption is not the presence of outliers *per se*, but rather that the underlying error distribution is non-Gaussian (and usually leptokurtic or heavy tailed, and possibly asymmetric), thus also giving rise to observations more extreme than the main cluster.

While general nonparametric methods are applicable in this setting, the method of partially adaptive estimation is very straightforward and still within the paradigm of parametric inference. It involves replacing the normality assumption with a flexible non-Gaussian distribution that embodies asymmetry and (semi-)heavy tails, and usually such that normality is a special or limiting case. General optimization routines will be required for computing the m.l.e., and bootstrap methods can be used for computing confidence intervals and other aspects of inference, such as forecasting.

The use of the Student's t distribution and its generalizations in regression analysis has been considered by McDonald and Newey (1988), Lange et al. (1989), and Butler et al. (1990). A less popular candidate, due to its historical complication regarding the evaluation of the p.d.f. (and thus the likelihood) is the (asymmetric) stable Paretian, as discussed in detail in Chapter II.8, and Sections III.9.4, III.9.5, and III.A.16. It also was the motivation for including the quote above by Pitman and Williams (1967).

The reason for its appeal, as compared to, say, use of (asymmetric) Student t variations, is the applicability of the generalized central limit theorem: One presumes that the standardized sum of all the neglected factors in the model (yielding the error term) converges to a stable distribution, of which normality is a special case. Note, however, that the non-Gaussian stable distribution does not possess a variance, and (as with any non-Gaussian distribution), the use of the bootstrap is recommended for inference on parameter and forecast uncertainty.

9) Use of **threshold regression**.

This is a type of **sample splitting model**, leading to far more general structures, such as cluster analysis and various multivariate methods in machine learning. As in Hansen (1999, 2000), under the assumption of two groups (referred to as classes, or regimes, in Hansen, 2000),

$$Y_t = \begin{cases} \mathbf{x}_t'\boldsymbol{\theta}_1 + \epsilon_t, \text{ if } & q_t \leqslant \gamma, \\ \mathbf{x}_t'\boldsymbol{\theta}_2 + \epsilon_t, \text{ if } & q_t > \gamma, \end{cases} \tag{1.108}$$

$t = 1, \ldots, T$, where $\mathbf{x}_t$ is a known $k \times 1$ vector; q_t is exogenous (not involving any Y_t) and is referred to as the threshold variable; and $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. It can be an element of $\mathbf{x}_t$ and, for the asymptotic theory developed by Hansen (2000), is assumed to be continuous. Finally, γ is the **threshold parameter**. Let, as usual, the regressor matrix be $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]'$, let $\mathbf{q} = [q_1, \ldots, q_T]'$ and $\mathbf{b} = \mathbb{I}\{\mathbf{q} \leqslant \gamma\}$, both $T \times 1$. Then, with $\mathbf{1}_k' = [1, 1, \ldots, 1]$ and selection matrix $\mathbf{S} = \mathbf{1}_k' \otimes \mathbf{b}$, define $\mathbf{X}_\gamma = \mathbf{S} \odot \mathbf{X}$, so that model (1.108) can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{X}_\gamma\boldsymbol{\delta} + \boldsymbol{\epsilon} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.109}$$

where $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are defined in the usual way, $\boldsymbol{\theta} = \boldsymbol{\theta}_2$, $\mathbf{Z} = [\mathbf{X},\ \mathbf{X}_\gamma]$ and $\boldsymbol{\beta} = [\boldsymbol{\theta}',\ \boldsymbol{\delta}']'$. Sample Matlab code to generate $\mathbf{X}_\gamma$ is given in Listing 1.12. For a given threshold γ, the usual least squares estimator (1.5) for $\boldsymbol{\beta}$ is used, and is also the m.l.e. under the usual Gaussian assumption on $\boldsymbol{\epsilon}$.

If γ were known, then the model reduces to the usual linear regression model, and the "significance" of $\boldsymbol{\delta}$ is assessed in the usual way, from Section 1.4. Matters are less clear when γ is to be elicited from the data. Let the **concentrated sum of squares** be given by (1.4), but as a function of γ, i.e.,

$$\mathsf{S}(\gamma) = \mathsf{S}(\gamma; \widehat{\boldsymbol{\beta}}; \mathbf{Y}, \mathbf{Z}) = \mathbf{Y}'\mathbf{M}_\gamma\mathbf{Y}, \quad \mathbf{M}_\gamma = \mathbf{I}_T - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'.$$

```
1    T=10; k=2; X=[ones(T,1), (1:T)']; b=rand(T,1)<0.5;
2    S = kron(ones(1,k),b); Xg = S.* X;
```

Program Listing 1.12: Example code for generating $\mathbf{X}_\gamma$ in (1.109).

(This is similar in concept to the **concentrated likelihood**, as will be used later in Section 5.6.3.1.) Assume $\gamma \in [\underline{\gamma}, \bar{\gamma}]$, and let

$$\hat{\gamma} = \underset{\gamma \in G}{\operatorname{argmin}}\ \mathsf{S}(\gamma)$$

be the least squares estimator of γ, where $G = [\underline{\gamma}, \bar{\gamma}] \cap \{q_1, \ldots, q_T\}$, noting that $\mathsf{S}(\gamma)$ takes on less than T distinct values. Hansen (2000) derives the asymptotic theory associated with estimator $\hat{\gamma}$, and approximate confidence intervals for γ based on a likelihood ratio statistic.

The case of model (1.108) with more than two groups is a straightforward generalization of this two-group setup. Examples of its use in macroeconomics include Rousseau and Wachtel (2002), Jude (2010), Stolbov (2013), Perri (2014), Pan et al. (2016), and the references therein.

10) **Quantile regression.**

The above quote by Mosteller and Tukey (1977) serves as a clear reminder of the limits of standard regression analysis and as one (of several) motivating factors for using quantile regression (QR). In particular, some contemplation reveals that, perhaps more often than not, it is not the mean that is of interest, but rather a particular quantile. For example, in income studies, interest might center on how the various exogenous factors influence not the mean income, but rather the lower 1, 5, and 10% quantiles, or their right-tail counterparts. Another benefit of QR compared to standard linear regression is that the median could be used instead of the mean as a type of robustified estimator, and/or its resulting implications (such as forecasts) compared to those based on the traditional use of the mean. Furthermore, QR allows for heteroskedasticity of the response function (recall the simple example in Figure 1.1) in a natural way, without requiring an explicit model for the error term that allows the exogenous variables to influence the estimates of σ_t (see, e.g., Fahrmeir et al., 2013, Ch. 10, for such an example and comparison to the use of QR).

A—clearly no longer relevant—disadvantage of QR is that closed-form solutions of the estimator no longer exist, and either linear programming techniques, or just general optimization algorithms, are required. One of the earliest survey articles on the topic is Koenker and Hallock (2001), while more detailed accounts can be found in the highly readable initial books of Koenker (2005) and Hao and Naiman (2007), as well as the newer Davino et al. (2014), which also provides code in R, SAS, and Stata.

11) **Generalized Linear Models.**

Above, we mentioned the use of robust estimators, or partially adaptive estimation, when the Gaussianity assumption is not applicable. However, these techniques are suitable when the unknown error distribution is "approximately Gaussian" in the sense of being unimodal, roughly bell-shaped, and having support over the whole real line. If the dependent variable is strictly positive and thus right-skewed, as occurs, for example, with lifetimes, waiting times, incomes, dividend payments, insurance claims, etc., then these aforementioned techniques are less applicable. Instead, one could model the expected value of a positive continuous random variable, such as the gamma, Pareto, (generalized) inverse Gaussian, etc., and the fitted regression coefficients would somehow need to be constrained such that $\mathbf{x}_t'\boldsymbol{\beta}$ is positive for all relevant $\mathbf{x}_t$.

Yet more complicated situations arise if the dependent variable is discrete, say, Bernoulli, binomial, multinomial, negative binomial, or Poisson. The above situation, as well as the discrete case, can all be elegantly handled by the use of what is referred to as the generalized linear model, or GLIM, whereby a transformation of the dependent variable is applied such that a regression can be used for modeling its mean. The assumed distribution of the dependent variable is usually taken to be a member of the exponential family, one example of which is the Gaussian, as studied in this chapter, in which case no transformation is required.

We briefly illustrate the mechanics assuming a Bernoulli distribution (with support zero and one) for the dependent variable Y. An example of this could be in so-called **credit scoring**, or **probability of default** models, whereby the credit-worthiness of a bank client (no or yes, i.e., 0 or 1) for receiving a loan is to be assessed, based on several exogenous factors (there are numerous books on this topic, e.g., Baesens et al. (2016) and Bluhm et al. (2010)). Let $\pi_i = \Pr(Y_i = 1) = \mathbb{E}[Y_i]$, and denote by η_i the linear predictor $\eta_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} = \mathbf{x}_i'\boldsymbol{\beta}$, $i = 1, 2, \ldots, n$, as in (1.2). They are related via a **response function** h such that $\pi_i = h(\eta_i)$, where h is a strictly monotone increasing function that maps to the interval $(0, 1)$, such as the standard normal c.d.f. Φ, and inverse function $\eta_i = g(\pi_i)$, where function $g = h^{-1}$ is referred to as the **link function**. The so-called **logit** model takes

$$\pi_i = h(\eta_i) = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}}, \quad g(\pi_i) = h^{-1}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta},$$

while the **probit** model takes $\pi_i = h(\eta_i) = \Phi(\eta_i)$.

Good introductory accounts of GLIM (with the benefit of having books that cover numerous other aspects of linear and other models) can be found in Rao et al. (2008, Ch. 10), Khuri (2010, Ch. 13), Fahrmeir et al. (2013, Ch. 5), and Greene (2017), while several highly detailed books dedicated to the subject exist, such as Fahrmeir and Tutz (2001), Winkelmann (2008), and Agresti (2015).

1.7 Problems

Problem 1.1 Consider the simple linear regression model $Y_t = \beta_1 + \beta_2 X_t + \epsilon_t$, $t = 1, \ldots, T$.

a) By setting $\partial \mathrm{S}(\boldsymbol{\beta})/\partial\beta_1$ to zero, show that $\widehat{\beta}_1 = \bar{Y} - \widehat{\beta}_2 \bar{X}$. Using this with $0 = \partial \mathrm{S}(\boldsymbol{\beta})/\partial\beta_2$, show that $\widehat{\beta}_2 = \hat{\sigma}_{X,Y}/\hat{\sigma}_X^2$, where $\hat{\sigma}_{X,Y}$ denotes the sample covariance between X and Y,

$$\hat{\sigma}_{X,Y} := \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y}),$$

and $\hat{\sigma}_X^2 := \hat{\sigma}_{X,X}$.

b) Show that $\widehat{Y}_t - \bar{Y} = \widehat{\beta}_2(X_t - \bar{X})$.

c) Define the standardized variables $x_t = (X_t - \bar{X})/\hat{\sigma}_X$ and $y_t = (Y_t - \bar{Y})/\hat{\sigma}_Y$, and consider the regression $y_t = \alpha_1 + \alpha_2 x_t + \varepsilon_t$. Show that $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \hat{\rho}$, where $\hat{\rho} = \hat{\rho}_{X,Y}$ is the sample correlation between X and Y, with $|\hat{\rho}| \leqslant 1$. Thus, we can write

$$\widehat{Y}_t = \hat{\alpha}_1 + \hat{\alpha}_2 x_t = \hat{\rho} x_t,$$

and squaring and summing both sides yields $\hat{\rho}^2 = \sum \widehat{Y}_t^2 / \sum x_t^2$. Show that the R^2 statistics for the two regression models are the same, namely $\hat{\rho}^2$.

Problem 1.2 Show (1.12) directly (without use of Theorem 1.6) for the simple linear regression model $Y_t = \beta_1 + \beta_2 X_t + \epsilon_t$.

Problem 1.3 For nonsingular matrix $\mathbf{A}$, its **partitioned inverse** $\mathbf{A}^{-1}$ is

$$\begin{aligned}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} \mathbf{W}^{-1} & -\mathbf{W}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{W}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{W}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{Z}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{Z}^{-1} \\ -\mathbf{Z}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{Z} \end{bmatrix},\end{aligned} \tag{1.110}$$

where $\mathbf{W} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ and $\mathbf{Z} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. This is a well-known result that can be found in numerous books on matrix algebra, and confirmed by computing $\mathbf{A}\mathbf{A}^{-1}$. Derive the Frisch–Waugh–Lovell theorem by applying the partitioned inverse (1.110) expression to (1.5).

Problem 1.4 Prove that the projection matrix $\mathbf{P}_{\mathcal{S}}$ in (1.42) is unique.
Hint: Let $\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2\ \dots\ \mathbf{h}_k]$ be a different basis for $\mathcal{S}$. Justify that we can write $\mathbf{H} = \mathbf{T}\mathbf{A}$ for some $\mathbf{A}$.

Problem 1.5 This is a less direct, but instructive, method for proving Theorem 1.3. Let $\mathbf{M} = \mathbf{I}_T - \mathbf{P}_{\mathcal{S}}$ with $\dim(\mathcal{S}) = k$, $k \in \{1, 2, \dots, T-1\}$. Via the spectral decomposition, let $\mathbf{H}$ be an orthogonal matrix whose rows consist of the eigenvectors of $\mathbf{M}$. Partition $\mathbf{H}$ as

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

with "correct" sizes, and use Theorem 1.2 to write $\mathbf{H}\mathbf{M}\mathbf{H}'$ as a block matrix. Show that $\mathbf{M}\mathbf{H}_2' = \mathbf{0}$ and $\mathbf{H}_2' = \mathbf{P}_{\mathcal{S}}\mathbf{H}_2'$. This implies the rows of $\mathbf{H}_2$ are in $\mathcal{S}$. Use this to show $\mathbf{H}_1\mathbf{H}_2' = \mathbf{0} \iff \mathbf{H}_1\mathbf{M} = \mathbf{H}_1$. Postmultiply $\mathbf{H}'\mathbf{H} = \mathbf{I}_T$ by $\mathbf{M}$ to show $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{M}$. Finally, show that $\mathbf{H}_1\mathbf{H}_1' = \mathbf{I}_{T-k}$.

Problem 1.6 Prove that the restricted least squares estimator $\widehat{\boldsymbol{\gamma}}$ given in (1.69) satisfies

1. $\mathbf{H}\widehat{\boldsymbol{\gamma}} = \mathbf{h}$ and
2. $\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 \leqslant \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$ for all $\mathbf{b} \in \mathbb{R}^k$ such that $\mathbf{H}\mathbf{b} = \mathbf{h}$.

Hint: For 2, first show that, for every $\mathbf{b} \in \mathbb{R}^k$ such that $\mathbf{H}\mathbf{b} = \mathbf{h}$,

$$\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}\|^2,$$

and then argue it suffices to show that $\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 \leqslant \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}\|^2$. Add and subtract $\widehat{\boldsymbol{\gamma}}$ to the latter term, expand, and show the cross term is zero.

Problem 1.7 Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Prove the Cauchy–Schwarz inequality $\langle \mathbf{u}, \mathbf{v} \rangle \leqslant \|\mathbf{u}\|\ \|\mathbf{v}\|$ as follows.

1. Show that $0 \leqslant \langle \mathbf{u} - a\mathbf{v}, \mathbf{u} - a\mathbf{v} \rangle$ for all $a \in \mathbb{R}$.
2. Expand $\langle \mathbf{u} - a\mathbf{v}, \mathbf{u} - a\mathbf{v} \rangle$ and let $a = \langle \mathbf{u}, \mathbf{v} \rangle / \langle \mathbf{v}, \mathbf{v} \rangle$.

Problem 1.8 Prove Theorem 1.2, i.e., if $\mathbf{P}$ is symmetric and idempotent with $\text{rank}(\mathbf{P}) = k$, then (i) k of the eigenvalues of $\mathbf{P}$ are unity and the remaining $T - k$ are zero, and (ii) $\text{tr}(\mathbf{P}) = k$.
Hint: For (i), continue with the relation $\lambda\mathbf{x} = \mathbf{P}\mathbf{x} = \mathbf{P}\mathbf{P}\mathbf{x}$, and for (ii), let $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}'$ and continue with the relation $k = \text{rank}(\mathbf{P}) = \text{tr}(\mathbf{D})$.

The converse of the result in Theorem 1.2 is, however, not true. For example, with $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A}) = 2$ and a standard computation shows that the eigenvalues of $\mathbf{A}$ are both one. But $\mathbf{A}$ is neither symmetric nor idempotent.

Finally, there are related results without requiring symmetry. For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1/4 & -1/6 \\ 18 & -7/2 & -3 \\ -15 & 15/4 & 7/2 \end{bmatrix}$$

is not symmetric, but it is idempotent, with rank two, eigenvalues 0, 1 and 1, and $\text{tr}(\mathbf{A}) = 2$. In general, if $\mathbf{A}$ is idempotent with k eigenvalues equal to one (and the rest zero), then $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A}) = k$; see, e.g., Magnus and Neudecker (2007, p. 22).

Problem 1.9 Prove Theorem 1.6.

Problem 1.10 Partition the linear regression model (1.3) as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

For convenience, let $\mathbf{M}_1 = \mathbf{M}_{\mathbf{X}_1} = \mathbf{I} - \mathbf{P}_{\mathbf{X}_1}$. Part (b) of Theorem 1.6 implies that $\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}$. Show this directly by using the projection and perpendicularity conditions (1.48) and (1.49).

Hint: Recall from the definition of column space (1.38) that, for an $\mathbf{x} \in C(\mathbf{X})$, there exists a $\boldsymbol{\gamma}$ such that $\mathbf{x} = \mathbf{X}\boldsymbol{\gamma} = \mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2$, where $\boldsymbol{\gamma}$ is appropriately partitioned into $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$.

Problem 1.11 Because $\mathbf{M}$ in (1.53) is a projection matrix onto $C(\mathbf{X})^{\perp}$, it follows from Theorem 1.2 that $\text{rank}(\mathbf{M}) = T - k$. Show this result using (B.67) and (B.68), i.e., if $\mathbf{A}$ and $\mathbf{B}$ are two matrices of the same size, then

$$\text{rank}(\mathbf{A} + \mathbf{B}) \leqslant \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B}),$$

and if $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ and $n \times k$ matrices, respectively, $k \geqslant 1$, then

$$\text{rank}(\mathbf{AB}) \geqslant \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B}) - n.$$

Problem 1.12 As in (1.66), let matrix $\mathbf{H}$ be of dimension $J \times k$ and full rank, with $J \leqslant k$. Show that $\mathbf{K} = \sigma^2\mathbf{AH}'(\mathbf{HAH}')^{-1}\mathbf{HA}$ is positive semi-definite for $J < k$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$.

Hint: If you are not convinced of the following fact, then prove it first: If $\mathbf{A}$ is a real symmetric matrix of size n with full rank n, then so is $\mathbf{A}^{-1}$.

What happens when $J = k$?

Problem 1.13 Numerically find the minimum number of observations T required in Example 1.11 to achieve a given power, using $\alpha = 0.05$.

Problem 1.14 We had derived the restricted least squares estimator $\widehat{\boldsymbol{\gamma}}$ for the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ when the restriction $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ holds, where $\mathbf{H}$ is $J \times k$ of full rank $J \leqslant k$. There is another way of doing

this. It begins by expressing $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ as $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\eta} + \mathbf{s}$, where the parameter vector $\boldsymbol{\eta}$ is of dimension $k - J$. That is, $\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where

$$\mathbf{X}\boldsymbol{\gamma} \in \mathcal{S}_H = \{\mathbf{y} : \mathbf{y} = \mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\beta} = \mathbf{S}\boldsymbol{\eta} + \mathbf{s},\ \boldsymbol{\eta} \in \mathbb{R}^{k-J}\}.$$

An extensive treatment of the relation between these parameterizations is provided by Hirschberg and Slottje (1999).

For example, let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)'$ and consider the constraint $\beta_2 = 2\beta_3$. Then we would take $\mathbf{H} = [0 \quad 1 \quad -2 \quad 0]$ and $\mathbf{h} = 0$. Alternatively, this can be expressed by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_2/2 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_4 \end{pmatrix} + \mathbf{0}, \text{ i.e., } \mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$\mathbf{s} = \mathbf{0}$ and $\eta = [\beta_1 \quad \beta_2 \quad \beta_4]'$.

a) Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)'$ but with the constraint that $\sum_{i=2}^{4} \beta_i = 1$. Give the appropriate values of $\mathbf{H}$, $\mathbf{h}$, $\mathbf{S}$, $\boldsymbol{\eta}$ and $\mathbf{s}$.
b) For some given values of $\mathbf{S}$, $\boldsymbol{\eta}$ and $\mathbf{s}$, derive $\widehat{\boldsymbol{\gamma}}$.
Hint: Plug in $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\eta} + \mathbf{s}$ into the regression model.
(Ruud, 2000, pp. 79–80)
c) Express $\mathbf{X}\widehat{\boldsymbol{\gamma}}$ as $\mathbf{P}_\mathbf{Z}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_\mathbf{Z})\mathbf{X}\mathbf{s}$, where $\mathbf{P}_\mathbf{Z}$ is a projection matrix.
d) Show that the constraint $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$, where $\mathbf{H}$ is $J \times k$ and $\text{rank}(\mathbf{H}) = J \leqslant k$, can always be expressed as $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\eta} + \mathbf{s}$.
(Ruud, 2000, p. 94(4.14a))

Problem 1.15 Recall the form of the generalized likelihood ratio statistic. For testing $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ in the linear model, it is given by

$$\text{LR} = \text{LR}(\mathbf{Y}, \mathbf{X}, \mathbf{H}, \mathbf{h}) = \frac{\max\limits_{\sigma^2, \boldsymbol{\beta}: \mathbf{H}\boldsymbol{\beta} = \mathbf{h}} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})}{\max\limits_{\sigma^2, \boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})} = \frac{\mathcal{L}(\widehat{\boldsymbol{\gamma}}, \tilde{\sigma}^2_\gamma; \mathbf{Y})}{\mathcal{L}(\widehat{\boldsymbol{\beta}}, \tilde{\sigma}^2; \mathbf{Y})},$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\tilde{\sigma}^2 = T^{-1}\mathsf{S}(\widehat{\boldsymbol{\beta}})$ refer to the unrestricted m.l.e. and $\widehat{\boldsymbol{\gamma}}$ and $\tilde{\sigma}^2_\gamma = T^{-1}\mathsf{S}(\widehat{\boldsymbol{\gamma}})$ refer to the restricted ones, where $\widehat{\boldsymbol{\gamma}}$ is given in (1.69). Show that a test of H_0 involving LR is equivalent to the F test given in (1.88).

Problem 1.16 This exercise will be of value in Section 2.5.2. Recall that, if $G \sim \text{Gam}(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$, its p.d.f. is

$$f_G(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathbb{I}(x > 0),$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1} \mathrm{e}^{-x}\, \mathrm{d}x \quad \text{and} \quad \int_0^\infty x^{\alpha-1} \exp(-\beta x)\, \mathrm{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha}. \tag{1.111}$$

Let $G_i \overset{\text{ind}}{\sim} \text{Gam}(\alpha_i, 1)$ and let $R_1 = G_1/G_3$ and $R_2 = G_2/G_3$. It is clear that, conditional on G_3, R_1 and R_2 are independent. Show that without conditioning they are not, by confirming (omitting the

obvious indicator functions)

$$f_{R_1,R_2}(r_1,r_2) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \frac{r_1^{\alpha_1-1} r_2^{\alpha_2-1}}{(1+r_1+r_2)^{\alpha_1+\alpha_2+\alpha_3}},$$

which does not factor as $f_{R_1}(r_1) \times f_{R_2}(r_2)$. Further confirm that $f_{R_1,R_2}(r_1,r_2)$ integrates to one by using the function `dblquad` in Matlab.

1.A Appendix: Derivation of the BLUS Residual Vector

This appendix derives the BLUS residual vector (1.104). It is a detailed amalgam of the various proofs given in Theil (1965, 1968, 1971), Chow (1976), and Magnus and Sinha (2005), with the hope that the development shown here (that becomes visible and straightforward once atop the proverbial shoulders of giants, notably Henri Theil and Jan Magnus) serves as a clear, complete, and perhaps definitive derivation.[14]

Recall that we wish a residual estimator of the form $\widehat{\boldsymbol{\epsilon}}_{\text{LUS}} = \mathbf{CY}$, where $\mathbf{C}$ is $(T-k) \times T$, and that the relevant minimization problem for the BLUS estimator is (writing just $\widehat{\boldsymbol{\epsilon}}$ for $\widehat{\boldsymbol{\epsilon}}_{\text{LUS}}$)

$$\widehat{\boldsymbol{\epsilon}}_{\text{BLUS}} = \arg\min_{\widehat{\boldsymbol{\epsilon}}} \mathbb{E}[(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)'(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)] \quad \text{subject to} \quad \mathbf{CX} = \mathbf{0},\ \mathbf{CC}' = \mathbf{I}, \tag{1.112}$$

where $\boldsymbol{\epsilon}_1$ is defined via the partition of the model in (1.103), repeated here as

$$\begin{bmatrix} \mathbf{Y}_0 \\ \mathbf{Y}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_0 \\ \boldsymbol{\epsilon}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \end{bmatrix} \widehat{\boldsymbol{\beta}}_{\text{LS}} + \begin{bmatrix} \mathbf{e}_0 \\ \mathbf{e}_1 \end{bmatrix}, \tag{1.113}$$

with $\boldsymbol{\epsilon}_0$ and $\mathbf{e}_0$ of size $k \times 1$, and $\boldsymbol{\epsilon}_1$ and $\mathbf{e}_1$ of size $(T-k) \times 1$.

We divide the derivation into several small parts.

Reduce the Two Constraints to One

The first part of the derivation consists in reducing the number of (matrix) constraints to one. The partition $\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{C}_1 \end{bmatrix}$ with $\mathbf{e} = \begin{bmatrix} \mathbf{e}_0 & \mathbf{e}_1 \end{bmatrix}'$, where $\mathbf{e}$ is of size $T \times 1$, yields

$$\mathbf{Ce} = \mathbf{C}_0\mathbf{e}_0 + \mathbf{C}_1\mathbf{e}_1, \tag{1.114}$$

where $\mathbf{C}_0$ is $(T-k) \times k$ and $\mathbf{C}_1$ is $(T-k) \times (T-k)$. Observe that the symmetry of $\mathbf{C}$ implies that of $\mathbf{C}_1$.

Using $\mathbf{CX} = \mathbf{0}$ and $\mathbf{X}'\mathbf{e} = \mathbf{0}$, we have

$$\mathbf{C}_0\mathbf{X}_0 + \mathbf{C}_1\mathbf{X}_1 = \mathbf{0}, \qquad \mathbf{X}'_0\mathbf{e}_0 + \mathbf{X}'_1\mathbf{e}_1 = \mathbf{0},$$

so that with

$$\mathbf{Z} = \mathbf{X}_1\mathbf{X}_0^{-1}, \tag{1.115}$$

we can write

$$\mathbf{e}_0 = -(\mathbf{X}_1\mathbf{X}_0^{-1})'\mathbf{e}_1 = -\mathbf{Z}'\mathbf{e}_1, \qquad \mathbf{C}_0 = -\mathbf{C}_1(\mathbf{X}_1\mathbf{X}_0^{-1}) = -\mathbf{C}_1\mathbf{Z}. \tag{1.116}$$

14 The author is grateful to my brilliant master's student Christian Frey for assembling this meticulous and detailed derivation from the original papers.

Further, using $\mathbf{CC}' = \mathbf{I}$, (1.116) yields

$$\mathbf{CC}' = \mathbf{C}_0\mathbf{C}_0' + \mathbf{C}_1\mathbf{C}_1' = \mathbf{C}_1\mathbf{Z}\mathbf{Z}'\mathbf{C}_1' + \mathbf{C}_1\mathbf{C}_1' = \mathbf{C}_1[\mathbf{I} + \mathbf{Z}\mathbf{Z}']\mathbf{C}_1' = \mathbf{I}, \tag{1.117}$$

so that both constraints $\mathbf{CX} = \mathbf{0}$ and $\mathbf{CC}' = \mathbf{I}$ are equivalent to (1.117). Moreover, by assumption $\mathbf{CX} = \mathbf{0}$, it follows that $\mathbf{CY} = \mathbf{C}\boldsymbol{\epsilon} = \mathbf{Ce}$. As $\mathbf{CY} = (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{C}\boldsymbol{\epsilon}$ and $\mathbf{Ce} = \mathbf{C}(\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}) = \mathbf{CY}$,

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{CY} = \mathbf{C}\boldsymbol{\epsilon} = \mathbf{C}_0\boldsymbol{\epsilon}_0 + \mathbf{C}_1\boldsymbol{\epsilon}_1 = -\mathbf{C}_1\mathbf{Z}\boldsymbol{\epsilon}_0 + \mathbf{C}_1\boldsymbol{\epsilon}_1,$$

and therefore

$$\begin{aligned}
&\mathrm{Cov}[(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1), (\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)] \\
&= \mathrm{Cov}[(-\mathbf{C}_1\mathbf{Z}\boldsymbol{\epsilon}_0 + (\mathbf{C}_1 - \mathbf{I})\boldsymbol{\epsilon}_1), (-\mathbf{C}_1\mathbf{Z}\boldsymbol{\epsilon}_0 + (\mathbf{C}_1 - \mathbf{I})\boldsymbol{\epsilon}_1)] \\
&= \sigma^2[\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}')\mathbf{C}_1' + \mathbf{I} - \mathbf{C}_1 - \mathbf{C}_1'].
\end{aligned} \tag{1.118}$$

The minimization problem for the BLUS estimator is then reduced to

$$\widehat{\boldsymbol{\epsilon}}_{\mathrm{BLUS}} = \arg\ \min_{\widehat{\boldsymbol{\epsilon}}} \mathbb{E}[(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)'(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)] \quad \text{subject to} \quad (1.117).$$

Solve with a Lagrangean Approach

Note that $\widehat{\boldsymbol{\epsilon}} = \mathbf{CY} = \mathbf{Ce}$, so that, with (1.118) and (1.117), the constrained minimization problem is equivalent to the Lagrangean

$$\begin{aligned}
L(\mathbf{C}_1, \lambda) = &\ \mathrm{tr}([\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}')\mathbf{C}_1' + \mathbf{I} - \mathbf{C}_1 - \mathbf{C}_1']) \\
&- \mathrm{tr}(\lambda[\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}')\mathbf{C}_1' - \mathbf{I}]),
\end{aligned} \tag{1.119}$$

where λ denotes the Lagrange multiplier matrix of dimension $(T-k)\times(T-k)$.

As $\partial\mathrm{tr}(\mathbf{AB})/\partial\mathbf{A} = \partial\mathrm{tr}(\mathbf{BA})/\partial\mathbf{A} = \mathbf{B}'$, the first-order condition with respect to $\mathbf{C}_1$ is

$$\frac{\partial L}{\partial \mathbf{C}_1} = 2\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}') - 2\mathbf{I} - 2\lambda\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}') = \mathbf{0}. \tag{1.120}$$

Symmetry of $\mathbf{C}_1$ Gives a Spectral Decomposition

To solve (1.120) for the two unknowns $\mathbf{C}_1$ and λ, postmultiply (1.120) by $\mathbf{C}_1'$ and use (1.117) to get

$$\lambda = \mathbf{I} - \mathbf{C}_1' = \mathbf{I} - \mathbf{C}_1, \tag{1.121}$$

which is obviously symmetric from the symmetry of $\mathbf{C}_1$. Substituting (1.121) in (1.120) yields

$$\mathbf{C}_1'\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}') = \mathbf{I}. \tag{1.122}$$

Thus, (1.122) and a spectral decomposition yield

$$\mathbf{C}_1^2 = (\mathbf{I} + \mathbf{Z}\mathbf{Z}')^{-1} = \mathbf{P}\mathbf{D}^2\mathbf{P}', \tag{1.123}$$

where, from the symmetry of $\mathbf{C}_1$, $\mathbf{D}^2$ is the $(T-k)\times(T-k)$ diagonal matrix with entries d_k^2 and $\mathbf{P}$ is the $(T-k)\times(T-k)$ orthogonal matrix ($\mathbf{PP}' = \mathbf{I}$) with columns given by the eigenvectors of $(\mathbf{I} + \mathbf{Z}\mathbf{Z}')^{-1}$ corresponding to the eigenvalues $d_1^2, \ldots, d_{T-k}^2$. It is worth emphasizing that the symmetry of $\mathbf{C}_1$ ensures that the d_i are real.

Note that the notation $\mathbf{D}^2$ stands for the d_k^2 entries of matrix $\mathbf{D}^2$, just to avoid usage of the root symbol, while $\mathbf{D}$ is the diagonal matrix with entries d_k restricted to the positive square roots. The solution

for (1.123) is then, say, $\mathbf{C}_1^* = (\mathbf{I} + \mathbf{Z}\mathbf{Z}')^{-1/2} = \mathbf{P}\mathbf{D}\mathbf{P}'$. To simplify notation, we subsequently take $\mathbf{C}_1 \equiv \mathbf{C}_1^*$.

It is useful to introduce the partition

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} \mathbf{M}_{00} & \mathbf{M}_{01} \\ \mathbf{M}_{10} & \mathbf{M}_{11} \end{bmatrix},$$

where $\mathbf{M}_{00} = \mathbf{I} - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'$, $\mathbf{M}_{01} = -\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'$, $\mathbf{M}_{10} = \mathbf{M}_{01}'$, and $\mathbf{M}_{11} = \mathbf{I} - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'$, though we will make use only of $\mathbf{M}_{11}$. Direct multiplication shows that $\mathbf{M}_{11}^{-1} = \mathbf{I} + \mathbf{X}_1(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_1'$, i.e., using this latter claim, $\mathbf{M}_{11}\mathbf{M}_{11}^{-1}$ is

$$\begin{aligned}
&[\mathbf{I} - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'][\mathbf{I} + \mathbf{X}_1(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_1'] \\
&= \mathbf{I} - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1' + \mathbf{X}_1(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_1' - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_1' \\
&= \mathbf{I} - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1' + \mathbf{X}_1(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_1' - \mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X} - \mathbf{X}_0'\mathbf{X}_0)(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_1' \\
&= \mathbf{I}.
\end{aligned}$$

Thus, with $\mathbf{Z} = \mathbf{X}_1\mathbf{X}_0^{-1}$ from (1.115),

$$\mathbf{M}_{11}^{-1} = \mathbf{I} + \mathbf{Z}\mathbf{Z}', \tag{1.124}$$

from which it follows that $\mathbf{M}_{11} = (\mathbf{I} + \mathbf{Z}\mathbf{Z}')^{-1}$. From (1.123) and (1.124), $\mathbf{M}_{11}^{-1} = (\mathbf{I} + \mathbf{Z}\mathbf{Z}') = (\mathbf{C}_1^2)^{-1} = \mathbf{C}_1^{-2}$ so that, from (1.116),

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_{\text{BLUS}} &= \mathbf{C}\mathbf{Y} = \mathbf{C}\mathbf{e} = \mathbf{C}_0\mathbf{e}_0 + \mathbf{C}_1\mathbf{e}_1 = (-\mathbf{C}_1\mathbf{Z})(-\mathbf{Z}'\mathbf{e}_1) + \mathbf{C}_1\mathbf{e}_1 \\
&= \mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}')\mathbf{e}_1 = \mathbf{C}_1\mathbf{M}_{11}^{-1}\mathbf{e}_1 = \mathbf{C}_1^{-1}\mathbf{e}_1 \\
&= \mathbf{e}_1 + (\mathbf{C}_1^{-1} - \mathbf{I})\mathbf{e}_1 \\
&= \mathbf{e}_1 + \sum_{k=1}^{T-k}(d_k^{-1} - 1)\mathbf{p}_k\mathbf{p}_k'\mathbf{e}_1,
\end{aligned} \tag{1.125}$$

where $\mathbf{p}_k$ are the eigenvectors and d_k^2 the eigenvalues of $\mathbf{M}_{11}$. The last equality follows by the existence of a spectral decomposition of $\mathbf{M}_{11} = \mathbf{C}_1^2 = \mathbf{P}\mathbf{D}^2\mathbf{P}'$, so that

$$\mathbf{M}_{11}\mathbf{p}_k = [\mathbf{I} - \mathbf{X}_1(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}_1']\mathbf{p}_k = d_k^2\mathbf{p}_k, \quad k = 1, \ldots, T-k. \tag{1.126}$$

Premultiplying both sides of (1.126) by $\mathbf{X}_1'$ and using $\mathbf{X}_1'\mathbf{X}_1 = \mathbf{X}'\mathbf{X} - \mathbf{X}_0'\mathbf{X}_0$,

$$\begin{aligned}
\mathbf{X}_1'\mathbf{p}_k - (\mathbf{X}'\mathbf{X} - \mathbf{X}_0'\mathbf{X}_0)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{p}_k &= d_k^2\mathbf{X}_1'\mathbf{p}_k \\
\mathbf{X}_0'\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{p}_k &= d_k^2\mathbf{X}_1'\mathbf{p}_k, \quad k = 1, \ldots, T-k.
\end{aligned} \tag{1.127}$$

Now premultiplying both sides of (1.127) by $(\mathbf{X}_0')^{-1}$, using $\mathbf{Z} = \mathbf{X}_1\mathbf{X}_0^{-1}$, and rearranging,

$$[\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0' - d_k^2\mathbf{I}]\mathbf{Z}'\mathbf{p}_k = \mathbf{0}, \quad k = 1, \ldots, T-k.$$

Use the Spectral Decomposition to Express the BLUS Estimator in terms of $\mathbf{e}_0$ and $\mathbf{e}_1$

Observe that d_k^2 is an eigenvalue of $\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'$. As the eigenvectors $\mathbf{Z}'\mathbf{p}_k$ do not have unit length, we normalize by a scalar to get, for $d_k < 1$,

$$\mathbf{q}_k = \frac{d_k}{\sqrt{1 - d_k^2}}\mathbf{Z}'\mathbf{p}_k, \quad k = 1, \ldots, T-k, \tag{1.128}$$

so that $\mathbf{q}_1, \ldots, \mathbf{q}_{T-k}$ have unit length and are pairwise orthogonal. As $\mathbf{P}$ is orthogonal, $\mathbf{P}^{-1} = \mathbf{P}'$, so that

$$\mathbf{Z}\mathbf{Z}' = \mathbf{M}_{11}^{-1} - \mathbf{I} = (\mathbf{P}\mathbf{D}^2\mathbf{P}')^{-1} - \mathbf{I} = (\mathbf{P}\mathbf{D}^{-2}\mathbf{P}') - \mathbf{I},$$

and observe that

$$\mathbf{Z}\mathbf{Z}'\mathbf{p}_k = \frac{1-d_k^2}{d_k^2}\mathbf{p}_k, \quad k = 1, \ldots, T-k.$$

Thus, $\mathbf{q}_l'\mathbf{q}_k = 1$ if $l = k$ and zero otherwise for $k, l = 1, \ldots, T-k$. From

$$\frac{1-d_k^2}{d_k^2}\mathbf{p}_k = \mathbf{Z}(\mathbf{Z}'\mathbf{p}_k) = \frac{\sqrt{1-d_k^2}}{d_k}\mathbf{Z}\mathbf{q}_k,$$

it follows that, if $d_k < 1$, $\mathbf{p}_k = \frac{d_k}{\sqrt{1-d_k^2}}\mathbf{Z}\mathbf{q}_k$, $k = 1, \ldots T-k$, so that, with $\mathbf{e}_0 = -\mathbf{Z}'\mathbf{e}_1$ and $\mathbf{Z} = \mathbf{X}_1\mathbf{X}_0^{-1}$, the last line of (1.125) can be written as

$$\hat{\boldsymbol{\epsilon}}_{\text{BLUS}} = \mathbf{e}_1 + \sum_{k=1}^{T-k}\left(\frac{1}{d_k} - 1\right)\mathbf{p}_k\mathbf{p}_k'\mathbf{e}_1 \tag{1.129}$$

$$= \mathbf{e}_1 + \mathbf{Z}\sum_{k=1}^{T-k}\left(\frac{1}{d_k} - 1\right)\frac{d_k^2}{1-d_k^2}\mathbf{q}_k\mathbf{q}_k'\mathbf{Z}'\mathbf{e}_1 \tag{1.130}$$

$$= \mathbf{e}_1 + \mathbf{X}_1\mathbf{X}_0^{-1}\sum_{k=1}^{T-k}\frac{d_k}{1+d_k}\mathbf{q}_k\mathbf{q}_k'\mathbf{e}_0, \tag{1.131}$$

where in (1.129), the kth term in the sum is zero if $d_k = 1$. Thus, we can restrict the summation in (1.130) and (1.131) to $k = 1, \ldots, H$, where $d_k < 1$, for all $k = 1, \ldots, H$, with $H \leqslant T-k$. The result is sometimes expressed as a permutation of the elements d_h, $h = 1, \ldots, H$, say $d_1 \leqslant d_2 \leqslant \ldots \leqslant d_H < 1$, such that the d_h are nondecreasing. This yields (1.104), i.e.,

$$\hat{\boldsymbol{\epsilon}}_{\text{BLUS}} = \mathbf{e}_1 + \mathbf{X}_1\mathbf{X}_0^{-1}\sum_{h=1}^{H}\frac{d_h}{1+d_h}\mathbf{q}_h\mathbf{q}_h'\mathbf{e}_0.$$

Observe that the BLUS estimator is represented as a deviation from the corresponding least squares errors.

Verification of Second-order Condition

As in Theil (1965), to verify that $\mathbf{C}^*$ or, equivalently, $\mathbf{C}_1^*$ is indeed a minimum of (1.123), consider an alternative estimator $\bar{\mathbf{C}}\mathbf{Y} = (\mathbf{C}+\mathbf{R})\mathbf{Y} = \begin{bmatrix}\mathbf{C}_0' + \mathbf{R}_0' & \mathbf{C}_1' + \mathbf{R}_1'\end{bmatrix}\mathbf{Y}$, where $\mathbf{C}_1 = \mathbf{P}\mathbf{D}\mathbf{P}'$ is the optimal symmetric matrix $\mathbf{C}_1$ from the first-order condition (1.123) and, hence, $\mathbf{C}_0 = -\mathbf{C}_1\mathbf{Z} = -\mathbf{P}\mathbf{D}\mathbf{P}'\mathbf{Z}$ from (1.116). Note that, as before, $\mathbf{C}_1 \equiv \mathbf{C}_1^*$ and similarly $\mathbf{C} \equiv \mathbf{C}^*$. Recall that $\mathbf{D}$ is restricted to contain only positive diagonal entries (eigenvalues). We wish to show that $\mathbf{C}^* \leqslant \bar{\mathbf{C}}$ for all $\bar{\mathbf{C}}$.

From the assumption $\bar{\mathbf{C}}\mathbf{X} = \mathbf{0}$, it follows that $\mathbf{R}_0'\mathbf{X}_0 + \mathbf{R}_1'\mathbf{X}_1 = \mathbf{0}$, so that $\mathbf{R}_0' = -\mathbf{R}_1'\mathbf{Z}$, with $\mathbf{Z} = \mathbf{X}_1\mathbf{X}_0^{-1}$. Thus, the assumption $\bar{\mathbf{C}}\,\bar{\mathbf{C}}' = \mathbf{I}$, such that $\bar{\mathbf{C}}$ has a scalar covariance matrix, implies

$$(\mathbf{C}+\mathbf{R})'(\mathbf{C}+\mathbf{R}) = (\mathbf{C}_0 + \mathbf{R}_0)'(\mathbf{C}_0 + \mathbf{R}_0) + (\mathbf{C}_1 + \mathbf{R}_1)'(\mathbf{C}_1 + \mathbf{R}_1)$$

$$= (\mathbf{C}_1 + \mathbf{R}_1)'(\mathbf{I} + \mathbf{Z}\mathbf{Z}')(\mathbf{C}_1 + \mathbf{R}_1)$$
$$= (\mathbf{C}_1 + \mathbf{R}_1)'\mathbf{M}_{11}^{-1}(\mathbf{C}_1 + \mathbf{R}_1) = \mathbf{I},$$

where the last equality follows from (1.124). From (1.124) and (1.123), $\mathbf{M}_{11}^{-1} = \mathbf{C}_1^{-2}$, and

$$(\mathbf{I} + \mathbf{C}_1^{-1}\mathbf{R}_1)'(\mathbf{I} + \mathbf{C}_1^{-1}\mathbf{R}_1) = \mathbf{I}, \tag{1.132}$$

implying that $\mathbf{C}_1^{-1}\mathbf{R}_1 + (\mathbf{C}_1^{-1}\mathbf{R}_1)'$ is negative semi-definite. Indeed, with $\mathbf{N} := \mathbf{C}_1^{-1}\mathbf{R}_1$ and $\mathbf{v}' \in \mathbb{R}^{T-k}$ an arbitrary (real) nonzero row vector, premultiplying both sides of (1.132) with $\mathbf{v}'$ and postmultiplying by $\mathbf{v}$ gives

$$\mathbf{v}'(\mathbf{I} + \mathbf{N})'(\mathbf{I} + \mathbf{N})\mathbf{v} = \mathbf{v}'\mathbf{v}, \tag{1.133}$$

implying

$$\mathbf{v}'(\mathbf{N} + \mathbf{N}')\mathbf{v} = -\mathbf{v}'\mathbf{N}'\mathbf{N}\mathbf{v} \leqslant 0, \tag{1.134}$$

so that $\mathbf{N} + \mathbf{N}'$ is negative semi-definite.

Recall that the (unconstrained) objective function in (1.119) can be rewritten with $\mathbf{C}_1\mathbf{C}_1' = \mathbf{I}$. Also recall the properties of the trace operator, $\text{tr}(\mathbf{C}_1) = \text{tr}(\mathbf{C}_1')$, $\text{tr}(\mathbf{C}_1\mathbf{C}_1') = \text{tr}(\mathbf{C}_1'\mathbf{C}_1)$ and $\text{tr}(\mathbf{C}_1(\mathbf{Z}\mathbf{Z}')\mathbf{C}_1') = \text{tr}(\mathbf{C}_1\mathbf{C}_1'(\mathbf{Z}\mathbf{Z}'))$. Then the expectation in (1.112) is

$$\begin{aligned}\mathbb{E}[(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)'(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_1)] &= \text{tr}([\mathbf{C}_1(\mathbf{I} + \mathbf{Z}\mathbf{Z}')\mathbf{C}_1' + \mathbf{I} - \mathbf{C}_1 - \mathbf{C}_1']) \\ &= \text{tr}(\mathbf{C}_1\mathbf{C}_1') + \text{tr}(\mathbf{C}_1(\mathbf{Z}\mathbf{Z}')\mathbf{C}_1') + \text{tr}(\mathbf{I}) - 2\text{tr}(\mathbf{C}_1) \\ &= 2\text{tr}(\mathbf{I}) + \text{tr}(\mathbf{I}(\mathbf{Z}\mathbf{Z}')) - 2\text{tr}(\mathbf{C}_1).\end{aligned}$$

It follows that the unconstrained optimization problem as a function only of $\mathbf{C}_1$ is equal to

$$-\min_{\mathbf{C}_1} \text{tr}(\mathbf{C}_1) = \max_{\mathbf{C}_1} \text{tr}(\mathbf{C}_1) = \max_{\mathbf{C}_1} \text{tr}\left(\sum_{k=1}^{T-k} \frac{1}{d_k}\mathbf{p}_k\mathbf{p}_k'\right), \tag{1.135}$$

where the last equality follows from the spectral decomposition $\mathbf{C}_1 = \mathbf{P}\mathbf{D}\mathbf{P}'$; see (1.123). The objective function of the maximization problem (1.135) applied to $\mathbf{R}_1$ is then given as

$$\begin{aligned}\text{tr}(\mathbf{R}_1) = \text{tr}(\mathbf{C}_1\mathbf{N}) &= \text{tr}\left(\sum_{k=1}^{T-k} \frac{1}{d_k}\mathbf{p}_k\mathbf{p}_k'\mathbf{N}\right) = \text{tr}\left(\sum_{k=1}^{T-k} \frac{1}{d_k}\mathbf{p}_k'\mathbf{N}\mathbf{p}_k\right) \\ &= \frac{1}{2}\text{tr}\left(\sum_{k=1}^{T-k} \frac{1}{d_k}\mathbf{p}_k'(\mathbf{N} + \mathbf{N}')\mathbf{p}_k\right) \leqslant 0,\end{aligned}$$

so that, by the negative semi-definiteness of $(\mathbf{N} + \mathbf{N}')$, $\mathbf{N} = \mathbf{0}$, or, equivalently, $\mathbf{R} = \mathbf{0}$, are corresponding maxima of the objective function (1.135) given that the eigenvalues d_k, $k = 1, \ldots, T - k$, are positive. Therefore, $\mathbf{C}_1^*$ is a minimum of (1.119) and hence $\mathbf{C}^*$ is a minimum of (1.112).

1.B Appendix: The Recursive Residuals

Here we provide more detail on the recursive residuals in (1.105). Let $\widehat{\boldsymbol{\beta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{Y}_j$ be the o.l.s. estimator obtained by using only the first j, $j \geqslant k$, observations, where $\mathbf{Y}_j$ is the $j \times 1$ vector of the first j

elements of $\mathbf{Y}$, and $\mathbf{X}_j$ is the $j \times k$ matrix of the first j rows of $\mathbf{X}$. As shown in Brown et al. (1975, p. 152), the $\widehat{\boldsymbol{\beta}}_j$, $j = k+1, \ldots, T$, can be obtained recursively.

In particular, writing $\mathbf{X}_j'\mathbf{X}_j = \mathbf{X}_{j-1}'\mathbf{X}_{j-1} + \mathbf{x}_j\mathbf{x}_j'$, where $\mathbf{x}_j'$ is the jth row of $\mathbf{X}$, we can apply (1.70) with $\mathbf{A} = \mathbf{X}_{j-1}'\mathbf{X}_{j-1}$, $\mathbf{B} = \mathbf{x}_j$ and scalar $\mathbf{D} = 1$, to get

$$(\mathbf{X}_j'\mathbf{X}_j)^{-1} = (\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1} - \frac{(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}}{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}. \tag{1.136}$$

Postmultiplying (1.136) by $\mathbf{x}_j$ and simplifying easily yields

$$(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{x}_j = \frac{(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}. \tag{1.137}$$

Next, from (1.6) and that $\widehat{\boldsymbol{\beta}}_{j-1} = (\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{X}_{j-1}'\mathbf{Y}_{j-1}$, write

$$\begin{aligned}\mathbf{X}_j'\mathbf{X}_j\widehat{\boldsymbol{\beta}}_j &= \mathbf{X}_j'\mathbf{Y}_j = \mathbf{X}_{j-1}'\mathbf{Y}_{j-1} + \mathbf{x}_jY_j = \mathbf{X}_{j-1}'\mathbf{X}_{j-1}\widehat{\boldsymbol{\beta}}_{j-1} + \mathbf{x}_jY_j \\ &= (\mathbf{X}_{j-1}'\mathbf{X}_{j-1} + \mathbf{x}_j\mathbf{x}_j')\widehat{\boldsymbol{\beta}}_{j-1} + \mathbf{x}_jY_j - \mathbf{x}_j\mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1} \\ &= \mathbf{X}_j'\mathbf{X}_j\widehat{\boldsymbol{\beta}}_{j-1} + \mathbf{x}_j(Y_j - \mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1}),\end{aligned}$$

premultiply with $(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ and finally use (1.137) to get

$$\widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}_{j-1} + \frac{(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j(Y_j - \mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1})}{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}, \quad j = k+1, \ldots, T. \tag{1.138}$$

The standardized quantities

$$V_j = \frac{Y_j - \mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1}}{\sqrt{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}}, \quad j = k+1, \ldots, T, \tag{1.139}$$

are defined to be the recursive residuals.

Let $\mathbf{V} = (V_{k+1}, \ldots, V_T)'$. We wish to derive the distribution of $\mathbf{V}$. Clearly, $\mathbb{E}[V_j] = 0$. For the variance, as Y_j and $\widehat{\boldsymbol{\beta}}_{j-1}$ are independent for $j = k+1, \ldots, T$, and recalling (1.8),

$$\begin{aligned}\mathbb{V}(V_j) &= \frac{1}{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}(\mathbb{V}(Y_j) + \mathbf{x}_j'\mathbb{V}(\widehat{\boldsymbol{\beta}}_{j-1})\mathbf{x}_j) \\ &= \frac{1}{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}(\sigma^2 + \sigma^2\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j) = \sigma^2.\end{aligned}$$

Vector $\mathbf{V}$ has a normal distribution, because $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$, and each V_j can be expressed as

$$V_j = \frac{\epsilon_j - \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\sum_{k=1}^{j-1}\mathbf{x}_k\epsilon_k}{\sqrt{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j}}. \tag{1.140}$$

To see this, note that $\mathbf{X}_{j-1}'(\mathbf{Y}_{j-1} - \mathbf{X}_{j-1}\,\boldsymbol{\beta}) = \sum_{k=1}^{j-1}\mathbf{x}_k\epsilon_k$ and hence for the numerator of V_j

$$Y_j - \mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1} = \epsilon_j - \mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1} + \mathbf{x}_j'\boldsymbol{\beta}$$

$$= \epsilon_j - \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{X}_{j-1}'(\mathbf{Y}_{j-1} - \mathbf{X}_{j-1}\,\boldsymbol{\beta})$$
$$= \epsilon_j - \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\sum_{k=1}^{j-1}\mathbf{x}_k\epsilon_k.$$

For the covariances of $\mathbf{V}$, let N_j be the numerator in (1.140). For $j < i$, $\mathbb{E}[N_jN_i]$ is

$$\mathbb{E}(\epsilon_j\epsilon_i) - \mathbb{E}\left[\epsilon_j\mathbf{x}_i'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\sum_{k=1}^{i-1}\mathbf{x}_k\epsilon_k\right] - \mathbb{E}\left[\epsilon_i\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\sum_{k=1}^{j-1}\mathbf{x}_k\epsilon_k\right]$$
$$+\,\mathbb{E}\left[\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\left(\sum_{k=1}^{j-1}\mathbf{x}_k\epsilon_k\right)\mathbf{x}_i'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\left(\sum_{k=1}^{i-1}\mathbf{x}_k\epsilon_k\right)\right].$$

This, in turn, is

$$-\sigma^2\mathbf{x}_i'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\mathbf{x}_j + \sigma^2\sum_{k=1}^{j-1}[\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_k\mathbf{x}_i'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\mathbf{x}_k] \tag{1.141}$$

$$= -\sigma^2\mathbf{x}_i'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\mathbf{x}_j + \sigma^2\sum_{k=1}^{j-1}[\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_k\mathbf{x}_k'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\mathbf{x}_i] \tag{1.142}$$
$$= -\sigma^2\mathbf{x}_i'(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\mathbf{x}_j + \sigma^2[\mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})(\mathbf{X}_{i-1}'\mathbf{X}_{i-1})^{-1}\mathbf{x}_i] = 0,$$

so that $\mathbf{V} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_{T-k})$.

1.C Appendix: Solutions

1) For the model $Y_t = \beta_1 + \beta_2X_t + \epsilon_t$, $t = 1, \ldots, T$, with $\hat{\epsilon}_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2X_t$, setting $\partial\mathsf{S}(\boldsymbol{\beta})/\partial\beta_1$ to zero gives $0 = -2\sum_{t=1}^{T}\widehat{\epsilon}_t$ or

$$\widehat{\beta}_1 = \bar{Y} - \widehat{\beta}_2\bar{X}. \tag{1.143}$$

Using this in the equation $0 = \partial\mathsf{S}(\boldsymbol{\beta})/\partial\beta_2 = -2\sum_{t=1}^{T}X_t\widehat{\epsilon}_t$ and simplifying yields

$$\widehat{\beta}_2 = \frac{\sum_{t=1}^{T}X_tY_t - T\bar{X}\bar{Y}}{\sum_{t=1}^{T}X_t^2 - T\bar{X}^2} = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^{T}(X_t - \bar{X})^2} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2}, \tag{1.144}$$

where $\hat{\sigma}_{X,Y}$ denotes the sample covariance between X and Y,

$$\hat{\sigma}_{X,Y} = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y}),$$

and $\hat{\sigma}_X^2 = \hat{\sigma}_{X,X}$. From the first derivative equations, it follows that $\sum\widehat{\epsilon}_t = \sum X_t\widehat{\epsilon}_t = 0$. Also, as $\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2X_t$, it is easy to verify using (1.143) that

$$\widehat{Y}_t - \bar{Y} = \widehat{\beta}_2(X_t - \bar{X}). \tag{1.145}$$

Define the standardized variables $x_t = (X_t - \bar{X})/\hat{\sigma}_X$ and $y_t = (Y_t - \bar{Y})/\hat{\sigma}_Y$ (so that $\bar{x} = \bar{y} = 0$, $\hat{\sigma}_x^2 = \hat{\sigma}_y^2 = 1$ and $\sum x_t^2 = \sum y_t^2 = T - 1$) and consider the regression $y_t = \alpha_1 + \alpha_2 x_t + \varepsilon_t$. Then (1.143) implies $\hat{\alpha}_1 = 0$ and (1.144) implies

$$\hat{\alpha}_2 = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x \hat{\sigma}_y} = \hat{\sigma}_{x,y} = \frac{1}{T-1}\sum_{t=1}^{T} x_t y_t = \frac{(T-1)^{-1}}{\hat{\sigma}_X \hat{\sigma}_Y}\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y}) = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y} = \hat{\rho},$$

where $\hat{\rho} = \hat{\rho}_{X,Y}$ is the sample correlation between X and Y, with $|\hat{\rho}| \leqslant 1$. Thus, we can write

$$\hat{y}_t = \hat{\alpha}_1 + \hat{\alpha}_2 x_t = \hat{\rho} x_t,$$

and squaring and summing both sides yields $\hat{\rho}^2 = \sum \hat{y}_t^2 / \sum x_t^2$. The R^2 statistic is then

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum (\hat{y}_t - \bar{y})^2}{\sum (y_t - \bar{y})^2} = \frac{\hat{\rho}^2 \sum x_t^2}{\sum y_t^2} = \hat{\rho}^2.$$

Using (1.145) and (1.144), R^2 for the original model is

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\hat{\beta}_2^2 \sum (X_t - \bar{X})^2}{\sum (Y_t - \bar{Y})^2} = \hat{\beta}_2^2 \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} = \frac{\hat{\sigma}_{X,Y}^2}{\hat{\sigma}_X^2 \hat{\sigma}_Y^2} = \hat{\rho}^2,$$

i.e., the same as for the regression with standardized components.

2) We need to show

$$\sum_{t=1}^{T}(Y_t - \bar{Y})^2 = \sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2 + \sum_{t=1}^{T}(\hat{Y}_t - \bar{Y})^2.$$

From (1.143) and (1.145), we get

$$\hat{Y}_t = \bar{Y} + \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2}(X_t - \bar{X}),$$

and using

$$\hat{\sigma}_{X,Y} = \frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y}) = \frac{1}{T}\sum_{t=1}^{T} X_t Y_t - \bar{X}\bar{Y},$$

simple algebra shows that

$$\sum_{t=1}^{T}(Y_t - \bar{Y})^2 = \sum_{t=1}^{T} Y_t^2 - T\bar{Y}^2,$$

$$\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{T} Y_t^2 - T\bar{Y}^2 - T\frac{\hat{\sigma}_{X,Y}^2}{\hat{\sigma}_X^2},$$

$$\sum_{t=1}^{T}(\hat{Y}_t - \bar{Y})^2 = T\frac{\hat{\sigma}_{X,Y}^2}{\hat{\sigma}_X^2},$$

proving the result.

3) From the appropriate partition

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix},$$

(1.110) implies that, with $\mathbf{U} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}$ and $\mathbf{V} = (\mathbf{X}_2'\mathbf{X}_2)^{-1}$,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{W}^{-1} & -\mathbf{W}^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{V} \\ -\mathbf{V}\mathbf{X}_2'\mathbf{X}_1\mathbf{W}^{-1} & \mathbf{V} + \mathbf{V}\mathbf{X}_2'\mathbf{X}_1\mathbf{W}^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{V} \end{pmatrix}$$

with $\mathbf{W} = \mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2\mathbf{V}\mathbf{X}_2'\mathbf{X}_1 = \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1$, where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$. Then

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

gives

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{W}^{-1}\mathbf{X}_1' - \mathbf{W}^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{V}\mathbf{X}_2')\mathbf{Y} = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y},$$

as in (1.22), and

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_2 &= (-\mathbf{V}\mathbf{X}_2'\mathbf{X}_1\mathbf{W}^{-1}\mathbf{X}_1' + (\mathbf{V} + \mathbf{V}\mathbf{X}_2'\mathbf{X}_1\mathbf{W}^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{V})\mathbf{X}_2')\mathbf{Y} \\ &= (\mathbf{V}\mathbf{X}_2' + \mathbf{V}\mathbf{X}_2'\mathbf{X}_1\mathbf{W}^{-1}\mathbf{X}_1'(\mathbf{X}_2\mathbf{V}\mathbf{X}_2' - \mathbf{I}))\mathbf{Y} \\ &= \mathbf{V}\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}) \\ &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1). \end{aligned}$$

4) Observe that, as $\mathbf{T} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k]$ in (1.42) is an orthonormal basis for $\mathcal{S}$, all vectors in $\mathcal{S}$ can be represented by linear combinations of these $\mathbf{w}_i$. In particular, if $[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k]$ is a (different) basis for $\mathcal{S}$, then we can write $\mathbf{H} = \mathbf{T}\mathbf{A}$, where $\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2\ \ldots\ \mathbf{h}_k]$ and $\mathbf{A}$ is a full rank $k \times k$ matrix. As $\mathbf{T}'\mathbf{T} = \mathbf{I}$ and $\mathbf{H}'\mathbf{H} = \mathbf{I}$, we have $\mathbf{I} = \mathbf{H}'\mathbf{H} = \mathbf{A}'\mathbf{T}'\mathbf{T}\mathbf{A} = \mathbf{A}'\mathbf{A}$, so that $\mathbf{A}$ is orthogonal with $\mathbf{A}' = \mathbf{A}^{-1}$. Then $\mathbf{H}\mathbf{H}' = \mathbf{T}\mathbf{A}\mathbf{A}'\mathbf{T}' = \mathbf{T}\mathbf{T}'$, showing that $\mathbf{P}_{\mathcal{S}}$ is unique. Matrix $\mathbf{A}$ can be computed as $(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{H}$. In Matlab, we can see this with the code in Listing 1.13.

5) Let $\mathbf{M} = \mathbf{I}_T - \mathbf{P}_{\mathcal{S}}$ with $\dim(\mathcal{S}) = k, k \in \{1, 2, \ldots, T-1\}$. Via the spectral decomposition, let $\mathbf{H}$ be an orthogonal matrix whose rows consist of the eigenvectors of $\mathbf{M}$. From Theorem 1.2, $\mathbf{H}$ can be partitioned as

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

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are of sizes $(T-k) \times T$ and $k \times T$, respectively, and such that

$$\mathbf{H}\mathbf{M}\mathbf{H}' = \begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{pmatrix} \mathbf{M} \begin{pmatrix} \mathbf{H}_1' & \mathbf{H}_2' \end{pmatrix} = \begin{pmatrix} \mathbf{H}_1\mathbf{M}\mathbf{H}_1' & \mathbf{H}_1\mathbf{M}\mathbf{H}_2' \\ \mathbf{H}_2\mathbf{M}\mathbf{H}_1' & \mathbf{H}_2\mathbf{M}\mathbf{H}_2' \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{T-k} & \mathbf{0}_{(T-k)\times k} \\ \mathbf{0}_{k\times(T-k)} & \mathbf{0}_{k\times k} \end{pmatrix}.$$

Then $\mathbf{0} = \mathbf{H}_2\mathbf{M}\mathbf{H}_2' = \mathbf{H}_2\mathbf{M}'\mathbf{M}\mathbf{H}_2' = (\mathbf{M}\mathbf{H}_2')'\mathbf{M}\mathbf{H}_2'$ implies that $\mathbf{H}_2\mathbf{M} = \mathbf{M}\mathbf{H}_2' = \mathbf{0}$ or

$$\mathbf{0} = (\mathbf{I} - \mathbf{P}_{\mathcal{S}})\mathbf{H}_2' \Leftrightarrow \mathbf{H}_2' = \mathbf{P}_{\mathcal{S}}\mathbf{H}_2'.$$

```
1 T=rand(4,2); T=orth(T); Q=[1,2;3,4]; H=T*Q; H=orth(H);
2 A=inv(T'*T)*T'*H; H-T*A, A'*A
```

Program Listing 1.13: Computes $\mathbf{A} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{H}$.

As $\mathbf{H}_2'$ is unchanged by projecting it onto $\mathcal{S}$, the rows of $\mathbf{H}_2$ are in $\mathcal{S}$. From this, and the fact that the rows of $\mathbf{H}$ are orthogonal,

$$\begin{aligned}\mathbf{H}_1\mathbf{H}_2' = \mathbf{0} &\Leftrightarrow \mathbf{H}_1\mathbf{P}_{\mathcal{S}}\mathbf{y} = \mathbf{0} \quad \forall\, \mathbf{y} \in \mathbb{R}^T \\ &\Leftrightarrow \mathbf{H}_1(\mathbf{I}\mathbf{y} - \mathbf{P}_{\mathcal{S}}\mathbf{y}) = \mathbf{H}_1\mathbf{y} \quad \forall\, \mathbf{y} \in \mathbb{R}^T \\ &\Leftrightarrow \mathbf{H}_1\mathbf{M}\mathbf{y} = \mathbf{H}_1\mathbf{y} \quad \forall\, \mathbf{y} \in \mathbb{R}^T \\ &\Leftrightarrow \mathbf{H}_1\mathbf{M} = \mathbf{H}_1.\end{aligned}$$

Postmultiplying $\mathbf{H}'\mathbf{H} = \mathbf{I}_T$ by $\mathbf{M}$ gives $\mathbf{H}_1'\mathbf{H}_1\mathbf{M} + \mathbf{H}_2'\mathbf{H}_2\mathbf{M} = \mathbf{M}$ or, as $\mathbf{H}_1\mathbf{M} = \mathbf{H}_1$ and $\mathbf{H}_2\mathbf{M} = \mathbf{0}$,

$$\mathbf{H}_1'\mathbf{H}_1 = \mathbf{M}. \tag{1.146}$$

Recall that the rows of $\mathbf{H}$ are orthonormal, so that

$$\mathbf{H}\mathbf{H}' = \begin{pmatrix}\mathbf{H}_1 \\ \mathbf{H}_2\end{pmatrix}\begin{pmatrix}\mathbf{H}_1' & \mathbf{H}_2'\end{pmatrix} = \begin{pmatrix}\mathbf{H}_1\mathbf{H}_1' & \mathbf{H}_1\mathbf{H}_2' \\ \mathbf{H}_2\mathbf{H}_1' & \mathbf{H}_2\mathbf{H}_2'\end{pmatrix} = \mathbf{I}_T = \begin{pmatrix}\mathbf{I}_{T-k} & \mathbf{0}_{(T-k)\times k} \\ \mathbf{0}_{k\times(T-k)} & \mathbf{I}_{k\times k}\end{pmatrix}$$

and, in particular,

$$\mathbf{H}_1\mathbf{H}_1' = \mathbf{I}_{T-k}. \tag{1.147}$$

The result follows from (1.146) and (1.147).

6) Let $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$. Direct substitution gives

$$\mathbf{H}\widehat{\boldsymbol{\gamma}} = \mathbf{H}[\widehat{\boldsymbol{\beta}} + \mathbf{A}\mathbf{H}'[\mathbf{H}\mathbf{A}\mathbf{H}']^{-1}(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}})] = \mathbf{H}\widehat{\boldsymbol{\beta}} + \mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}},$$

so that the first condition is satisfied. To see the second, note that, for every $\mathbf{b} \in \mathbb{R}^k$ such that $\mathbf{H}\mathbf{b} = \mathbf{h}$, we can write

$$\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}\|^2, \tag{1.148}$$

because the cross term $(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}) = \widehat{\boldsymbol{\epsilon}}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \mathbf{b}) = 0$ from (1.61). Because the first term in (1.148) does not depend on $\mathbf{b}$ or $\widehat{\boldsymbol{\gamma}}$, it suffices to show that

$$\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 \leqslant \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}\|^2. \tag{1.149}$$

First note that the cross term $(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}})'(\mathbf{X}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\mathbf{b})$ vanishes because, from (1.69),

$$\begin{aligned}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\gamma}})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\gamma}} - \mathbf{b}) &= -(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}})'[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\gamma}} - \mathbf{b}) \\ &= -(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}})'[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}']^{-1}(\mathbf{H}\widehat{\boldsymbol{\gamma}} - \mathbf{H}\mathbf{b}) = \mathbf{0},\end{aligned}$$

as $\mathbf{H}\widehat{\boldsymbol{\gamma}} = \mathbf{h} = \mathbf{H}\mathbf{b}$. Thus, the right-hand side of (1.149) is

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \mathbf{b})\|^2 = \|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\gamma}} - \mathbf{b})\|^2 = \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2 + \|\mathbf{X}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\mathbf{b}\|^2,$$

and, as $\|\mathbf{X}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\mathbf{b}\|^2$ is non-negative, (1.149) is true. Strict equality holds when $\mathbf{X}\widehat{\boldsymbol{\gamma}}$ equals $\mathbf{X}\mathbf{b}$, but as $\mathbf{X}$ is of full rank, this holds if and only if $\widehat{\boldsymbol{\gamma}} = \mathbf{b}$.

7) From the definition of $\langle\cdot,\cdot\rangle$, for any $\mathbf{v} \in \mathbb{R}^n$, $\langle \mathbf{v}, \mathbf{v}\rangle = \sum_{i=1}^n v_i^2 \geqslant 0$. For the second part,

$$\begin{aligned}\langle \mathbf{u} - a\mathbf{v}, \mathbf{u} - a\mathbf{v}\rangle &= \sum_{i=1}^n (u_i - av_i)^2 = \sum_{i=1}^n u_i^2 - 2a\sum_{i=1}^n u_i v_i + a^2\sum_{i=1}^n v_i^2 \\ &= \langle \mathbf{u}, \mathbf{u}\rangle - 2a\langle \mathbf{u}, \mathbf{v}\rangle + a^2\langle \mathbf{v}, \mathbf{v}\rangle,\end{aligned}$$

so that, with $a = \langle \mathbf{u}, \mathbf{v} \rangle / \langle \mathbf{v}, \mathbf{v} \rangle$,

$$0 \leqslant \langle \mathbf{u}, \mathbf{u} \rangle - 2a\langle \mathbf{u}, \mathbf{v} \rangle + a^2 \langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle - 2\frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{v}, \mathbf{v} \rangle} + \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{v}, \mathbf{v} \rangle} = \langle \mathbf{u}, \mathbf{u} \rangle - \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{v}, \mathbf{v} \rangle},$$

or

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leqslant \langle \mathbf{u}, \mathbf{u} \rangle \langle \mathbf{v}, \mathbf{v} \rangle.$$

As both sides are positive, taking square roots gives the inequality $\langle \mathbf{u}, \mathbf{v} \rangle \leqslant \|\mathbf{u}\| \ \|\mathbf{v}\|$, where $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$.

8) (Theorem 1.2)
From idempotency, for any eigenvalue λ and corresponding eigenvector $\mathbf{x}$,

$$\lambda \mathbf{x} = \mathbf{P}\mathbf{x} = \mathbf{P}\mathbf{P}\mathbf{x} = \mathbf{P}\lambda\mathbf{x} = \lambda\mathbf{P}\mathbf{x} = \lambda^2\mathbf{x},$$

which implies that $\lambda = \lambda^2$, so that the only solutions are $\lambda = 0$ or 1 (there are no complex solutions, though note that, from the assumption of symmetry, all eigenvalues are real anyway). Also from symmetry, the number of nonzero eigenvalues of $\mathbf{P}$ equals $\text{rank}(\mathbf{P}) = k$, proving (i).
For (ii), form the spectral decomposition of $\mathbf{P}$ as $\mathbf{UDU}'$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix with k ones and $T - k$ zeros. Using the fact that (for conformable matrices) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$,

$$k = \text{rank}(\mathbf{P}) = \text{tr}(\mathbf{D}) = \text{tr}(\mathbf{UDU}') = \text{tr}(\mathbf{P}).$$

9) (Theorem 1.6)
a) For convenience, we restate (1.50) from the proof in the text: Take as a basis for $\mathbb{R}^T$ the vectors

$$\overbrace{\underbrace{\mathbf{r}_1, \ldots, \mathbf{r}_r}_{\mathcal{S}_0 \text{ basis}}, \overbrace{\mathbf{s}_{r+1}, \ldots, \mathbf{s}_s}^{\mathcal{S}\backslash\mathcal{S}_0 \text{ basis}}}^{\mathcal{S} \text{ basis}}, \underbrace{\mathbf{z}_{s+1}, \ldots, \mathbf{z}_T}_{\mathcal{S}^\perp \text{ basis}} \tag{1.150}$$

and let $\mathbf{y} = \mathbf{r} + \mathbf{s} + \mathbf{z}$, where $\mathbf{r} \in \mathcal{S}_0$, $\mathbf{s} \in \mathcal{S}\backslash\mathcal{S}_0$ and $\mathbf{z} \in \mathcal{S}^\perp$ are orthogonal.
b) Let $\mathbf{Q} = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}_0}$. From Theorem 1.4, if $\mathbf{Q}$ is symmetric and idempotent, then it is the projection matrix onto $\mathcal{C}(\mathbf{Q})$, but it is clearly symmetric and, from the first part of the theorem,

$$\mathbf{QQ} = \mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}_0} - \mathbf{P}_{\mathcal{S}_0}\mathbf{P}_{\mathcal{S}} + \mathbf{P}_{\mathcal{S}_0}\mathbf{P}_{\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}_0}.$$

For $\mathcal{C}(\mathbf{Q}) = \mathcal{S}\backslash\mathcal{S}_0$, it must be that, for $\mathbf{s} \in \mathcal{S}\backslash\mathcal{S}_0$ and $\mathbf{w} \in (\mathcal{S}\backslash\mathcal{S}_0)^\perp$, $\mathbf{Qs} = \mathbf{s}$ and $\mathbf{Qw} = \mathbf{0}$. As $\mathcal{S}\backslash\mathcal{S}_0 \subset \mathcal{S}$, $\mathbf{P}_{\mathcal{S}}\mathbf{s} = \mathbf{s}$ and, as $\mathbf{s} \perp \mathcal{S}_0$, $\mathbf{P}_{\mathcal{S}_0}\mathbf{s} = \mathbf{0}$, showing that $\mathbf{Qs} = \mathbf{s}$. Next, from (1.150), $\mathbf{w}$ can be expressed as

$$\mathbf{w} = c_1\mathbf{r}_1 + \cdots + c_r\mathbf{r}_r + c_{s+1}\mathbf{z}_{s+1} + \cdots + c_T\mathbf{z}_T$$

for some constants $c_i \in \mathbb{R}$. As $\mathbf{z}_i \perp \mathcal{S}$ (which implies $\mathbf{z}_i \perp \mathcal{S}_0 \subset \mathcal{S}$), $\mathbf{P}_{\mathcal{S}_0}\mathbf{w} = \mathbf{P}_{\mathcal{S}}\mathbf{w} = c_1\mathbf{r}_1 + \cdots + c_r\mathbf{r}_r$ so that $\mathbf{Qw} = \mathbf{0}$. Thus, $\mathcal{C}(\mathbf{Q}) = \mathcal{S}\backslash\mathcal{S}_0$ and $\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0} = \mathbf{Q} = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}_0}$. Note that this is a special case of the earlier result

$$\mathbf{P}_{\mathcal{S}^\perp} = \mathbf{P}_{\mathbb{R}^T\backslash\mathcal{S}} = \mathbf{P}_{\mathbb{R}^T} - \mathbf{P}_{\mathcal{S}} = \mathbf{I}_T - \mathbf{P}_{\mathcal{S}}.$$

c) As $\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}_0}$,

$$\begin{aligned}\|\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}\mathbf{y}\|^2 &= \|\mathbf{P}_{\mathcal{S}}\mathbf{y} - \mathbf{P}_{\mathcal{S}_0}\mathbf{y}\|^2 = (\mathbf{P}_{\mathcal{S}}\mathbf{y} - \mathbf{P}_{\mathcal{S}_0}\mathbf{y})'(\mathbf{P}_{\mathcal{S}}\mathbf{y} - \mathbf{P}_{\mathcal{S}_0}\mathbf{y}) \\ &= \mathbf{y}'\mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}}\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathcal{S}_0}\mathbf{P}_{\mathcal{S}}\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathcal{S}}\mathbf{P}_{\mathcal{S}_0}\mathbf{y} + \mathbf{y}'\mathbf{P}_{\mathcal{S}_0}\mathbf{P}_{\mathcal{S}_0}\mathbf{y} \\ &= \|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|^2 - \|\mathbf{P}_{\mathcal{S}_0}\mathbf{y}\|^2\end{aligned}$$

using the results from part (a).

d) By expressing (1.150) as

$$\overbrace{\mathbf{r}_1, \ldots, \mathbf{r}_r}^{\mathcal{S}_0}, \overbrace{\underbrace{\mathbf{s}_{r+1}, \ldots \mathbf{s}_s}_{\mathcal{S}_0^\perp\backslash\mathcal{S}^\perp}, \underbrace{\mathbf{z}_{s+1}, \ldots \mathbf{z}_T}_{\mathcal{S}_0^\perp\cap\mathcal{S}^\perp=\mathcal{S}^\perp}}^{\mathcal{S}_0^\perp},$$

it is clear that $\mathcal{S}\backslash\mathcal{S}_0 = \mathcal{S}_0^\perp\backslash\mathcal{S}^\perp$. To verify the last equality,

$$\mathbf{P}_{\mathcal{S}_0^\perp} - \mathbf{P}_{\mathcal{S}^\perp} = (\mathbf{I} - \mathbf{P}_{\mathcal{S}_0}) - (\mathbf{I} - \mathbf{P}_{\mathcal{S}}) = \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}_0} = \mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}.$$

e) This follows easily from (1.150) because $\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}\mathbf{y} \in (\mathcal{S}\backslash\mathcal{S}_0) \subset \mathcal{S}$, so that $\mathbf{P}_{\mathcal{S}}(\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}\mathbf{y})$ remains $\mathbf{P}_{\mathcal{S}\backslash\mathcal{S}_0}\mathbf{y}$. Transposing gives the other equality.

10) For the projection condition, let $\mathbf{x} \in C(\mathbf{X})$. We need to show that $(\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2})\mathbf{x} = \mathbf{x}$. From the hint,

$$\begin{aligned}(\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2})\mathbf{x} &= (\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2})(\mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2) \\ &= \mathbf{P}_{\mathbf{X}_1}(\mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2) + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}(\mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2).\end{aligned}$$

Clearly, $\mathbf{P}_{\mathbf{X}_1}\mathbf{X}_1\boldsymbol{\gamma}_1 = \mathbf{X}_1\boldsymbol{\gamma}_1$, and as $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1$, we have $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{X}_1 = \mathbf{0}$ (as $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$) and $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2$. Thus,

$$\begin{aligned}(\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2})\mathbf{x} &= \mathbf{P}_{\mathbf{X}_1}(\mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2) + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}(\mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2) \\ &= \mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{P}_{\mathbf{X}_1}\mathbf{X}_2\boldsymbol{\gamma}_2 + \mathbf{M}_1\mathbf{X}_2\boldsymbol{\gamma}_2 \\ &= \mathbf{X}_1\boldsymbol{\gamma}_1 + (\mathbf{P}_{\mathbf{X}_1} + \mathbf{M}_1)\mathbf{X}_2\boldsymbol{\gamma}_2 \\ &= \mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2 \\ &= \mathbf{X}\boldsymbol{\gamma} = \mathbf{x},\end{aligned}$$

as $\mathbf{M}_1 = \mathbf{M}_{\mathbf{X}_1} = \mathbf{I} - \mathbf{P}_{\mathbf{X}_1}$.

For the perpendicularity condition, recall that the orthogonal complement of $C(\mathbf{X})$ is

$$C(\mathbf{X})^\perp = \{\mathbf{z} \in \mathbb{R}^T : \mathbf{X}'\mathbf{z} = \mathbf{0}\}. \tag{1.151}$$

Let $\mathbf{u} \in C(\mathbf{X})^\perp$. We need to show that $(\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2})\mathbf{u} = \mathbf{0}$. For the first term, note that, directly from (1.151), $C(\mathbf{X})^\perp \subset C(\mathbf{X}_1)^\perp$, i.e., if $\mathbf{u} \in C(\mathbf{X})^\perp$, then $\mathbf{u} \in C(\mathbf{X}_1)^\perp$, so that $\mathbf{P}_{\mathbf{X}_1}\mathbf{u} = \mathbf{0}$. For the second term, first note that, as $C(\mathbf{X})^\perp \subset C(\mathbf{X}_2)^\perp$, $\mathbf{X}_2'\mathbf{u} = \mathbf{0}$. As

$$\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1 = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1}),$$

the condition $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{u} = \mathbf{0}$ holds if both $\mathbf{X}_2'\mathbf{u} = \mathbf{0}$ and $\mathbf{P}_{\mathbf{X}_1}\mathbf{u} = \mathbf{0}$ hold, and we have just seen that these are both true, and we are done.

11) Write $\mathbf{I} = \mathbf{I} - \mathbf{P} + \mathbf{P}$, and use Theorem B.67 to get $T - \text{rank}(\mathbf{P}) \leqslant \text{rank}(\mathbf{I} - \mathbf{P})$. But, as $\mathbf{P}$ is idempotent, we have $\mathbf{0} = (\mathbf{I} - \mathbf{P})\mathbf{P}$, so from Theorem B.68, $T - \text{rank}(\mathbf{P}) \geqslant \text{rank}(\mathbf{I} - \mathbf{P})$. Together, they imply that $\text{rank}(\mathbf{I} - \mathbf{P}) = T - \text{rank}(\mathbf{P}) = k$.

12) For the statement in the hint, to see that $\mathbf{A}^{-1}$ is symmetric,

$$\mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \Leftrightarrow \mathbf{I}' = \mathbf{I} = \mathbf{A}^{-1'}\mathbf{A} \Leftrightarrow \mathbf{I}\mathbf{A}^{-1} = \mathbf{A}^{-1'}\mathbf{A}\mathbf{A}^{-1} \Leftrightarrow \mathbf{A}^{-1} = \mathbf{A}^{-1'}.$$

As $\mathbf{A}$ is symmetric, all its eigenvalues are real, so that $\mathbf{A}$ has spectral decomposition $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}'$ with $\mathbf{U}$ orthonormal and $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ with each d_i real and positive. Then $\mathbf{A}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$ (confirmed by calculating $\mathbf{A}\mathbf{A}^{-1}$) with $\mathbf{D}^{-1} = \text{diag}(d_1^{-1}, \ldots, d_n^{-1})$ with each $d_i^{-1} > 0$, implying that $\mathbf{A}^{-1}$ is also full rank.

To show that $\mathbf{K}$ is positive semi-definite: Let $\mathbf{x}$ be a $k \times 1$ real vector. We have to show that $\mathbf{x}'\mathbf{K}\mathbf{x} \geqslant 0$ for all $\mathbf{x}$ or, with $\mathbf{z} = \mathbf{H}\mathbf{A}\mathbf{x}$ and the fact that $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$ is symmetric, that $\mathbf{z}'(\mathbf{H}\mathbf{A}\mathbf{H}')^{-1}\mathbf{z} \geqslant 0$. But this is true because $\mathbf{H}\mathbf{A}\mathbf{H}'$ (and, thus, $(\mathbf{H}\mathbf{A}\mathbf{H}')^{-1}$) is symmetric and full rank, i.e., $\mathbf{q}'\mathbf{H}\mathbf{A}\mathbf{H}'\mathbf{q} > 0$ for all nonzero $\mathbf{q}$.

Observe that $\mathbf{K}$ is not necessarily positive definite when $J < k$ because $\mathbf{z} = \mathbf{H}\mathbf{A}\mathbf{x}$ could be zero even for nonzero $\mathbf{x}$. This is the case, for example, with

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{x} = \text{null}(\mathbf{H}) = \begin{bmatrix} 0 \\ \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}.$$

If $J = k$ and, as always assumed, $\mathbf{H}$ is full rank, then $\mathbf{H}$ is a square matrix with unique inverse, and $\boldsymbol{\beta}$ is fully specified from the restrictions and the data have no influence on its estimate, i.e., the restriction $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ implies that $\boldsymbol{\beta} = \mathbf{H}^{-1}\mathbf{h}$ and $\widehat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}\mathbf{h}$, which is not stochastic and, thus, has a zero covariance matrix. This agrees with the expression (1.77), because, with $J = k$,

$$\begin{aligned} \mathbf{K} &= \sigma^2\mathbf{A}\mathbf{H}'(\mathbf{H}\mathbf{A}\mathbf{H}')^{-1}\mathbf{H}\mathbf{A} \\ &= \sigma^2\mathbf{A}\mathbf{H}' \; \mathbf{H}'^{-1}\mathbf{A}^{-1}\mathbf{H}^{-1} \; \mathbf{H}\mathbf{A} = \sigma^2\mathbf{A} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \text{Var}(\widehat{\boldsymbol{\beta}}). \end{aligned}$$

13) Using program `ncf.m` to compute the noncentral F c.d.f., the code in Listing 1.14 will do the job.

14)

a) We take $\mathbf{H} = [0 \quad 1 \quad 1 \quad 1]$ and $\mathbf{h} = 1$. The constraint implies, for example, that $\beta_2 = 1 - \beta_3 - \beta_4$, so that $\mathbf{S}$, $\boldsymbol{\eta}$ and $\mathbf{s}$ are given via

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ 1 - \beta_3 - \beta_4 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

b) The model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}\mathbf{S}\boldsymbol{\eta} + \mathbf{X}\mathbf{s} + \boldsymbol{\epsilon}$ or $\mathbf{Y} - \mathbf{X}\mathbf{s} = \mathbf{X}\mathbf{S}\boldsymbol{\eta} + \boldsymbol{\epsilon}$, so that, with $\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}\mathbf{s}$ and $\mathbf{Z} = \mathbf{X}\mathbf{S}$,

$$\widehat{\boldsymbol{\eta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}^* = (\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{s}).$$

```
powneed=0.90; beta=[0 -5 3 5]'; H=[1 -1 0 0; 0 0 1 -1]; sig2=9;
notenough=1; a=5;
while notenough
  a=a+1; n=4*a;
  dum1=[ones(n,1); zeros(n,1)]; dum2=1-dum1;
  time=kron((1:4)',ones(floor(n/4),1));
  c3=kron([1,0]',time); c4=kron([0,1]',time);
  X=[dum1 dum2 c3 c4]; A=inv(X'*X);
  theta=beta'*H'*inv(H*A*H')*H*beta/sig2;
  cutoff = finv(0.95,2,2*n-4); pow=1-ncf(cutoff,2,36,theta,0)
  if pow>=powneed, notenough=0; end
end
T=2*n
```

Program Listing 1.14: Finds minimum T for a given power `powneed` based on the setup in Example 1.11. Here, $T = 2n$, and n is incremented in steps of 4.

From the constraint $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\eta} + \mathbf{s}$,

$$\hat{\boldsymbol{\gamma}} = \mathbf{S}\hat{\boldsymbol{\eta}} + \mathbf{s} = \mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{s}) + \mathbf{s}.$$

c) We have

$$\mathbf{X}\hat{\boldsymbol{\gamma}} = \mathbf{X}\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{s}) + \mathbf{X}\mathbf{s} = \mathbf{P}_{\mathbf{Z}}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}\mathbf{s},$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{X}\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'$ is clearly a projection matrix.

d) Choose $\mathbf{H}$ and $\boldsymbol{\beta}$ in such a way that the partition

$$\mathbf{H}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{[1]} \\ \boldsymbol{\beta}_{[2]} \end{pmatrix} = \mathbf{H}_1\boldsymbol{\beta}_{[1]} + \mathbf{H}_2\boldsymbol{\beta}_{[2]} = \mathbf{h}$$

can be formed for which $\mathbf{H}_1$ is $J \times J$ and nonsingular. (This is always possible because $\mathbf{H}$ is full rank J.) Premultiplying by $\mathbf{H}_1^{-1}$ implies that $\boldsymbol{\beta}_{[1]} = \mathbf{H}_1^{-1}\mathbf{h} - \mathbf{H}_1^{-1}\mathbf{H}_2\boldsymbol{\beta}_{[2]}$ and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{[1]} \\ \boldsymbol{\beta}_{[2]} \end{pmatrix} = \begin{pmatrix} -\mathbf{H}_1^{-1}\mathbf{H}_2 \\ \mathbf{I}_{k-J} \end{pmatrix} \boldsymbol{\beta}_{[2]} + \begin{pmatrix} \mathbf{H}_1^{-1}\mathbf{h} \\ \mathbf{0}_{k-J} \end{pmatrix} = \mathbf{S}\boldsymbol{\eta} + \mathbf{s}.$$

15) From (1.9),

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{\beta}}, \tilde{\sigma}^2; \mathbf{Y}) &= \frac{1}{(2\pi)^{T/2}\tilde{\sigma}^T} \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right\} \\ &= \frac{1}{(2\pi)^{T/2}\tilde{\sigma}^T} \exp\left\{-\frac{1}{2T^{-1}\mathsf{S}(\hat{\boldsymbol{\beta}})}\mathsf{S}(\hat{\boldsymbol{\beta}})\right\} = \frac{\mathrm{e}^{-T/2}}{(2\pi)^{T/2}\tilde{\sigma}^T}, \end{aligned}$$

and, similarly,

$$\mathcal{L}(\hat{\boldsymbol{\gamma}}, \tilde{\sigma}_{\boldsymbol{\gamma}}^2; \mathbf{Y}) = \frac{\mathrm{e}^{-T/2}}{(2\pi)^{T/2}\tilde{\sigma}_{\boldsymbol{\gamma}}^T},$$

so that

$$R = \left(\frac{\tilde{\sigma}_{\boldsymbol{\gamma}}}{\tilde{\sigma}}\right)^{-T} = \left(\frac{\tilde{\sigma}_{\boldsymbol{\gamma}}^2}{\tilde{\sigma}^2}\right)^{-T/2}.$$

The GLRT rejects for small R, i.e., when $\tilde{\sigma}_\gamma^2/\tilde{\sigma}^2$ is large. In terms of sums of squares, R rejects when $\mathsf{S}(\widehat{\boldsymbol{\gamma}})/\mathsf{S}(\widehat{\boldsymbol{\beta}})$ is large, or, equivalently, when

$$\frac{T-k}{J}\left(\frac{\mathsf{S}(\widehat{\boldsymbol{\gamma}})}{\mathsf{S}(\widehat{\boldsymbol{\beta}})}-1\right)=\frac{[\mathsf{S}(\widehat{\boldsymbol{\gamma}})-\mathsf{S}(\widehat{\boldsymbol{\beta}})]/J}{\mathsf{S}(\widehat{\boldsymbol{\beta}})/(T-k)}=\frac{\mathsf{S}(\widehat{\boldsymbol{\gamma}})-\mathsf{S}(\widehat{\boldsymbol{\beta}})}{J\hat{\sigma}^2}=F$$

is large. Thus, the F test and the GLRT are the same.

16) With $\mathbf{G}=(G_1,G_2,G_3)$, $R_3\equiv G_3$, and $\mathbf{R}=(R_1,R_2,R_3)$, the one-to-one transformation of $\mathbf{r}=(r_1,r_2,r_3)$ to $\mathbf{g}=(g_1,g_2,g_3)$ is $g_1=r_1r_3$, $g_2=r_2r_3$, and $g_3=r_3$. The Jacobian is

$$\mathbf{J}=\begin{bmatrix}\partial g_1/\partial r_1 & \partial g_2/\partial r_1 & \partial g_3/\partial r_1\\ \partial g_1/\partial r_2 & \partial g_2/\partial r_2 & \partial g_3/\partial r_2\\ \partial g_1/\partial r_3 & \partial g_2/\partial r_3 & \partial g_3/\partial r_3\end{bmatrix}=\begin{bmatrix}r_3 & 0 & 0\\ 0 & r_3 & 0\\ r_1 & r_2 & 1\end{bmatrix},\quad \det(\mathbf{J})=r_3^2,$$

and, as

$$\begin{aligned}f_{\mathbf{G}}(\mathbf{g})=&\frac{1}{\Gamma(\alpha_1)}\frac{1}{\Gamma(\alpha_2)}\frac{1}{\Gamma(\alpha_3)}\mathbb{I}(g_1>0)\mathbb{I}(g_2>0)\mathbb{I}(g_3>0)\\ &\times g_1^{\alpha_1-1}g_2^{\alpha_2-1}g_3^{\alpha_3-1}\exp(-g_1-g_2-g_3),\end{aligned}$$

the joint density of $\mathbf{R}$ is

$$\begin{aligned}f_{\mathbf{R}}(\mathbf{r})=&f_{\mathbf{G}}(\mathbf{g})|\det(\mathbf{J})|\\ =&\frac{1}{\Gamma(\alpha_1)}\frac{1}{\Gamma(\alpha_2)}\frac{1}{\Gamma(\alpha_3)}r_3^{\alpha_1+\alpha_2+\alpha_3-1}r_1^{\alpha_1-1}r_2^{\alpha_2-1}\exp(-r_3(1+r_1+r_2)).\end{aligned}$$

As $g_3=r_3$, the margin $R_3\sim\text{Gam}(\alpha_3,1)$, and

$$\begin{aligned}f_{(R_1,R_2)|R_3}(r_1,r_2\mid r_3)=&\frac{f_{\mathbf{R}}(\mathbf{r})}{f_{R_3}(r_3)}\\ &\propto r_1^{\alpha_1-1}\exp(-r_3r_1)\times r_2^{\alpha_2-1}\exp(-r_3r_2)\times r_3^{\alpha_1+\alpha_2},\end{aligned}$$

so that, conditional on $R_3=r_3$, the density of R_1 and R_2 factors, and R_1 and R_2 are conditionally independent.

```
1  function I = gam3(a1,a2,a3)
2  up=20; I = dblquad(@RR,0,up,0,up);
3
4    function A=RR(r1,r2)
5      c = gamma(a1+a2+a3) / (gamma(a1)*gamma(a2)*gamma(a3));
6      num = r1.^(a1-1).* r2.^(a2-1);
7      den = (1+r1+r2).^(a1+a2+a3);
8      A = c * num./den;
9    end
10
11 end
```

Program Listing 1.15: Computes the integral in (1.152), confirming it is 1.000. The integral upper limit up would have to be chosen in a more intelligent manner to work for all values of input parameters a_1, a_2, and a_3.

For the joint density of R_1 and R_2, using (1.111), $f_{R_1,R_2}(r_1, r_2)$ is

$$\begin{aligned}
&\int_0^\infty f_{\mathbf{R}}(\mathbf{r})\,\mathrm{d}r_3 \\
&= \frac{1}{\Gamma(\alpha_1)}\frac{1}{\Gamma(\alpha_2)}\frac{1}{\Gamma(\alpha_3)} r_1^{\alpha_1-1} r_2^{\alpha_2-1} \int_0^\infty r_3^{\alpha_1+\alpha_2+\alpha_3-1} \exp(-r_3(1+r_1+r_2))\,\mathrm{d}r_3 \\
&= \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \frac{r_1^{\alpha_1-1} r_2^{\alpha_2-1}}{(1+r_1+r_2)^{\alpha_1+\alpha_2+\alpha_3}}.
\end{aligned} \tag{1.152}$$

The program in Listing 1.15 shows how to use function `dblquad` within Matlab with what they call *nested functions* to perform the integration.