1

Response Envelopes

Envelopes, which were introduced by Cook et al. (2007) and developed for the multivariate linear model by Cook et al. (2010), encompass a class of methods for increasing efficiency in multivariate analyses without altering traditional objectives. They serve to reshape classical methods by exploiting response–predictor relationships that affect the accuracy of the results but are not recognized by classical methods. Multivariate data are often modeled by combining a selected structural component to be estimated with an error component to account for the remaining unexplained variation. Capturing the desired signal and only that signal in the structural component can be an elusive task with the consequence that, in an effort to avoid missing important information, there may be a tendency to overparameterize, leading to overfitting and relatively soft inferences and interpretations. Essentially a type of targeted dimension reduction that can result in substantial gains in efficiency, envelopes operate by enveloping the signal and thereby account for extraneous variation that might otherwise be present in the structural component.

In this chapter, we consider multivariate (multiresponse) linear regression allowing for the presence of "immaterial variation" (described herein) in the response vector. The possibility of such variation being present in the predictors is considered in Chapter 4, where we develop a connection with partial least squares regression. Section 1.1 contains a very brief review of the multivariate linear model, with an emphasis on aspects that will play a role in later developments. Additional background is available from Muirhead (2005). The envelope model for response reduction is introduced in Section 1.2. Introductory illustrations are given in Section 1.3 to provide intuition, to set the tone for later developments, and to provide running examples. In later sections, we discuss additional properties of the envelope model, maximum likelihood estimation, and the asymptotic variance of the envelope estimator of the coefficient matrix. Most of the technical materials used in this chapter are taken from Cook et al. (2010). Some algebraic details are presented without justification. The missing development is given extensively in Appendix A, which covers the linear algebra of envelopes.

An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics,
First Edition. R. Dennis Cook.

1.1 The Multivariate Linear Model

Consider the multivariate regression of a response vector $\mathbf{Y} \in \mathbb{R}^r$ on a vector of nonstochastic predictors $\mathbf{X} \in \mathbb{R}^p$. The standard linear model for describing a sample $(\mathbf{Y}_i, \mathbf{X}_i)$ can be represented in vector form as

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n, \tag{1.1}$$

where the predictors are centered in the sample $\sum_{i=1}^n \mathbf{X}_i = 0$, the error vectors $\boldsymbol{\varepsilon}_i \in \mathbb{R}^r$ are independently and identically distributed normal vectors with mean 0 and covariance matrix $\boldsymbol{\Sigma} > 0$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ is an unknown vector of intercepts, and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is an unknown matrix of regression coefficients. Centering the predictors facilitates discussion and presentation of some results, but is technically unnecessary. If $\mathbf{X}$ is stochastic, so $\mathbf{X}$ and $\mathbf{Y}$ have a joint distribution, we still condition on the observed values of $\mathbf{X}$ since the predictors are ancillary under model (1.1). The normality requirement for $\boldsymbol{\varepsilon}$ is not essential, as discussed in Section 1.9 and in later chapters.

Let $\mathbb{Y}$ denote the $n \times r$ centered matrix with rows $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$, let $\mathbb{Y}_0$ denote the $n \times r$ uncentered matrix with rows $\mathbf{Y}_i^T$, and let $\mathbb{X}$ denote the $n \times p$ matrix with rows $\mathbf{X}_i^T$, $i = 1, \dots, n$. Also, let $\mathbf{S}_{\mathbf{Y},\mathbf{X}} = \mathbb{Y}^T\mathbb{X}/n$ and

$$\mathbf{S}_{\mathbf{X}} = \mathbb{X}^T\mathbb{X}/n = n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T.$$

Then the maximum likelihood estimator of $\boldsymbol{\alpha}$ is $\bar{\mathbf{Y}}$, and the maximum likelihood estimator of $\boldsymbol{\beta}$, which is also the ordinary least squares estimator, is

$$\mathbf{B} = \mathbb{Y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \mathbb{Y}_0^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \mathbf{S}_{\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}, \tag{1.2}$$

where the second equality follows because the predictors are centered. To see this result, let $\widehat{\mathbf{Y}}_i = \bar{\mathbf{Y}} + \mathbf{B}\mathbf{X}_i$ and $\mathbf{r}_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i$ denote the ith vectors of fitted values and residuals, $i = 1, \dots, n$, and let $\mathbf{D} = \boldsymbol{\beta} - \mathbf{B}$. Then after substituting $\bar{\mathbf{Y}}$ for $\boldsymbol{\alpha}$, the remaining log-likelihood $L(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ to be maximized can be expressed as

$$\begin{aligned}(2/n)L(\boldsymbol{\beta},\boldsymbol{\Sigma}) &= c - \log|\boldsymbol{\Sigma}| - n^{-1}\sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}} - \boldsymbol{\beta}\mathbf{X}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \bar{\mathbf{Y}} - \boldsymbol{\beta}\mathbf{X}_i)\\
&= c - \log|\boldsymbol{\Sigma}| - n^{-1}\sum_{i=1}^n (\mathbf{r}_i - \mathbf{D}\mathbf{X}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{r}_i - \mathbf{D}\mathbf{X}_i)\\
&= c - \log|\boldsymbol{\Sigma}| - n^{-1}\operatorname{tr}\left(\sum_{i=1}^n \mathbf{r}_i\mathbf{r}_i^T\boldsymbol{\Sigma}^{-1}\right)\\
&\quad - n^{-1}\operatorname{tr}\left(\mathbf{D}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T\mathbf{D}^T\boldsymbol{\Sigma}^{-1}\right)\\
&= c - \log|\boldsymbol{\Sigma}| - n^{-1}\operatorname{tr}\left(\sum_{i=1}^n \mathbf{r}_i\mathbf{r}_i^T\boldsymbol{\Sigma}^{-1}\right) - \operatorname{tr}(\mathbf{D}\mathbf{S}_{\mathbf{X}}\mathbf{D}^T\boldsymbol{\Sigma}^{-1}),\end{aligned}$$

where $c = -r\log(2\pi)$ and the last step follows because $\sum_{i=1}^{n} \mathbf{r}_i \mathbf{X}_i^T = 0$. Consequently, $L(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is maximized over $\boldsymbol{\beta}$ by setting $\boldsymbol{\beta} = \mathbf{B}$ so $\mathbf{D} = 0$, leaving the partially maximized log-likelihood

$$(2/n)L(\boldsymbol{\Sigma}) = -r\log(2\pi) - \log|\boldsymbol{\Sigma}| - n^{-1}\,\mathrm{tr}\left(\sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i^T\boldsymbol{\Sigma}^{-1}\right).$$

It follows that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\mathbf{S}_{\mathbf{Y}|\mathbf{X}} := n^{-1}\sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i^T$ and that the fully maximized log-likelihood is

$$\widehat{L} = -(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|.$$

We notice from (1.2) that $\mathbf{B}$ can be constructed by doing r separate univariate linear regressions, one for each element of $\mathbf{Y}$ on $\mathbf{X}$. The coefficients from the jth regression then form the jth row of $\mathbf{B}$, $j = 1, \dots, r$. The stochastic relationships among the elements of $\mathbf{Y}$ are not used in forming these estimators. However, as will be seen later, relationships among the elements of $\mathbf{Y}$ play a central role in envelope estimation. Standard inference on $(\boldsymbol{\beta})_{jk}$, the (j,k) th element of $\boldsymbol{\beta}$, under model (1.1) is the same as inference obtained under the univariate linear regression of Y_j, the jth element of $\mathbf{Y}$, on $\mathbf{X}$. Model (1.1) becomes operational as a multivariate construction when inferring simultaneously about elements in different rows of $\boldsymbol{\beta}$ or when predicting elements of $\mathbf{Y}$ jointly.

The sample covariance matrices of $\mathbf{Y}$, $\widehat{\mathbf{Y}}$, and $\mathbf{r}$ can be expressed as

$$\mathbf{S}_{\mathbf{Y}} = n^{-1}\mathbb{Y}^T\mathbb{Y} = \mathbf{S}_{\mathbf{Y}\circ\mathbf{X}} + \mathbf{S}_{\mathbf{Y}|\mathbf{X}}, \tag{1.3}$$

$$\mathbf{S}_{\mathbf{Y}\circ\mathbf{X}} = n^{-1}\mathbb{Y}^T\mathbf{P}_{\mathbb{X}}\mathbb{Y} = \mathbf{S}_{\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{X},\mathbf{Y}}, \tag{1.4}$$

$$\begin{aligned}\mathbf{S}_{\mathbf{Y}|\mathbf{X}} &= n^{-1}\sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i^T = n^{-1}\mathbb{Y}^T\mathbf{Q}_{\mathbb{X}}\mathbb{Y}, \\ &= \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{Y},\mathbf{X}}^T, \\ &= \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{Y}\circ\mathbf{X}},\end{aligned} \tag{1.5}$$

where $\mathbf{S}_{\mathbf{X}}$ is nonstochastic, $\mathbf{P}_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ denotes the projection onto the column space of $\mathbb{X}$, $\mathbf{Q}_{\mathbb{X}} = \mathbf{I}_n - \mathbf{P}_{\mathbb{X}}$, $\mathbf{S}_{\mathbf{Y}\circ\mathbf{X}}$ is the sample covariance matrix of the fitted vectors $\widehat{\mathbf{Y}}_i$, and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ is the sample covariance matrix of the residuals $\mathbf{r}_i$.

We will occasionally encounter the standardized version of $\mathbf{B}$,

$$\tilde{\mathbf{B}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\mathbf{B}\mathbf{S}_{\mathbf{X}}^{1/2}, \tag{1.6}$$

which corresponds to the estimated coefficient matrix from the ordinary least squares fit of the standardized responses $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\mathbf{Y}$ on the standardized predictors $\mathbf{S}_{\mathbf{X}}^{-1/2}\mathbf{X}$.

The joint distribution of the elements of $\mathbf{B}$ can be found by using the vec operator to stack the columns of $\mathbf{B}$: $\mathrm{vec}(\mathbf{B}) = \{(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\}\mathrm{vec}(\mathbb{Y}_0^T)$, where $\otimes$ denotes the Kronecker product. Since $\mathrm{vec}(\mathbb{Y}_0^T)$ is normally distributed with mean $\mathbf{1}_n \otimes \boldsymbol{\alpha} + (\mathbb{X} \otimes \mathbf{I}_r)\mathrm{vec}(\boldsymbol{\beta})$ and variance $\mathbf{I}_n \otimes \boldsymbol{\Sigma}$, it follows that $\mathrm{vec}(\mathbf{B})$ is normally distributed with mean and variance

$$\mathrm{E}\{\mathrm{vec}(\mathbf{B})\} = \mathrm{vec}(\boldsymbol{\beta}), \tag{1.7}$$

$$\mathrm{var}\{\mathrm{vec}(\mathbf{B})\} = (\mathbb{X}^T\mathbb{X})^{-1} \otimes \boldsymbol{\Sigma} = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}. \tag{1.8}$$

The covariance matrix can also be represented in terms of $\mathbf{B}^T$ by using the $rp \times rp$ commutation matrix $\mathbf{K}_{rp}$ to convert $\mathrm{vec}(\mathbf{B})$ to $\mathrm{vec}(\mathbf{B}^T)$: $\mathrm{vec}(\mathbf{B}^T) = \mathbf{K}_{rp}\,\mathrm{vec}(\mathbf{B})$ and

$$\mathrm{var}\{\mathrm{vec}(\mathbf{B}^T)\} = n^{-1}\mathbf{K}_{rp}(\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma})\mathbf{K}_{rp}^T = n^{-1}\boldsymbol{\Sigma} \otimes \mathbf{S}_{\mathbf{X}}^{-1}.$$

Background on the commutation matrix, vec and related operators is available in Appendix A. The variance $\mathrm{var}\{\mathrm{vec}(\mathbf{B})\}$ is typically estimated by substituting the residual covariance matrix for $\boldsymbol{\Sigma}$,

$$\widehat{\mathrm{var}}\{\mathrm{vec}(\mathbf{B})\} = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \mathbf{S}_{\mathbf{Y}|\mathbf{X}}. \tag{1.9}$$

Let $\mathbf{e}_i \in \mathbb{R}^r$ denote the indicator vector with a 1 in the ith position and 0s elsewhere. Then the covariance matrix for the ith row of $\mathbf{B}$ is

$$\mathrm{var}\{\mathrm{vec}(\mathbf{e}_i^T\mathbf{B})\} = (\mathbf{I}_p \otimes \mathbf{e}_i^T)\mathrm{var}\{\mathrm{vec}(\mathbf{B})\}(\mathbf{I}_p \otimes \mathbf{e}_i) = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma})_{ii}.$$

We see from this that the covariance matrix for the ith row of $\mathbf{B}$ is the same as that from the marginal linear regression of Y_i on $\mathbf{X}$. We refer to the estimator $(\mathbf{B})_{ij}$ divided by its standard error $\{n^{-1}(\mathbf{S}_{\mathbf{X}}^{-1})_{jj}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}})_{ii}\}^{1/2}$ as a Z-score:

$$Z = \frac{(\mathbf{B})_{ij}}{\{n^{-1}(\mathbf{S}_{\mathbf{X}}^{-1})_{jj}(\mathbf{S}_{\mathbf{Y}|\mathbf{X}})_{ii}\}^{1/2}}. \tag{1.10}$$

This statistic will be used from time to time for assessing the magnitude of $(\mathbf{B})_{ij}$, sometimes converting to a p-value using the standard normal distribution.

We will occasionally encounter a conditional variate of the form $\mathbf{N} \mid \mathbf{C}^T\mathbf{N}$, where $\mathbf{N} \in \mathbb{R}^r$ is a normal vector with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Delta}$, and $\mathbf{C} \in \mathbb{R}^{r\times q}$ is a nonstochastic matrix with $q \leq r$. The mean and variance of this conditional form are as follows:

$$\mathrm{E}(\mathbf{N} \mid \mathbf{C}^T\mathbf{N}) = \boldsymbol{\mu} + \mathbf{P}_{\mathbf{C}(\boldsymbol{\Delta})}^T(\mathbf{N} - \boldsymbol{\mu}), \tag{1.11}$$

$$\begin{aligned}\mathrm{var}(\mathbf{N} \mid \mathbf{C}^T\mathbf{N}) &= \boldsymbol{\Delta} - \boldsymbol{\Delta}\mathbf{C}(\mathbf{C}^T\boldsymbol{\Delta}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Delta} \\ &= \boldsymbol{\Delta}\mathbf{Q}_{\mathbf{C}(\boldsymbol{\Delta})} \\ &= \mathbf{Q}_{\mathbf{C}(\boldsymbol{\Delta})}^T\boldsymbol{\Delta}\mathbf{Q}_{\mathbf{C}(\boldsymbol{\Delta})}.\end{aligned} \tag{1.12}$$

The usual log-likelihood ratio statistic for testing that $\boldsymbol{\beta} = 0$ is

$$\Lambda = n \log \frac{|\mathbf{S}_{\mathbf{Y}}|}{|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|}, \tag{1.13}$$

which is asymptotically distributed under the null hypothesis as a chi-square random variable with pr degrees of freedom. We will occasionally use this statistic in illustrations to assess the presence of any detectable dependence of $\mathbf{Y}$ on $\mathbf{X}$. This statistic is sometimes reported with an adjustment that is useful when n is not large relative to r and p (Muirhead 2005, Section 10.5.2).

The Fisher information $\mathbf{J}$ for $(\text{vec}^T(\boldsymbol{\beta}), \text{vech}^T(\boldsymbol{\Sigma}))^T$ in model (1.1) is

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_r \end{pmatrix}, \tag{1.14}$$

where $\mathbf{E}_r$ is the expansion matrix that satisfies $\text{vec}(\mathbf{A}) = \mathbf{E}_r \,\text{vech}(\mathbf{A})$ for $\mathbf{A} \in \mathbb{S}^{r\times r}$, and $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T/n > 0$. It follows from standard likelihood theory that $\sqrt{n}(\text{vec}(\mathbf{B}) - \text{vec}(\boldsymbol{\beta}))$ is asymptotically normal with mean 0 and variance given by the upper left block of $\mathbf{J}^{-1}$,

$$\text{avar}(\sqrt{n}\,\text{vec}(\mathbf{B})) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}. \tag{1.15}$$

Asymptotic normality holds also without normal errors but with some technical conditions: if the errors have finite fourth moments and $\max_{1\le i\le n}(\mathbf{P}_{\mathbb{X}})_{ii} \to 0$, then $\sqrt{n}(\text{vec}(\mathbf{B}) - \text{vec}(\boldsymbol{\beta}))$ converges in distribution to a normal vector with mean 0 (e.g. Su and Cook 2012, Theorem 2).

1.1.1 Partitioned Models and Added Variable Plots

A subset of the predictors may occasionally be of special interest in multivariate regression. Partition $\mathbf{X}$ into two sets of predictors $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_1}$, $p_1 + p_2 = p$, and conformably partition the columns of $\boldsymbol{\beta}$ into $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Then model (1.1) can be rewritten as

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\varepsilon}, \tag{1.16}$$

where $\boldsymbol{\beta}_1$ holds the coefficients of interest. We next reparameterize this model to force the new predictors to be uncorrelated in the sample and to focus attention on $\boldsymbol{\beta}_1$.

Recalling that $\bar{\mathbf{X}} = 0$, let $\widehat{\mathbf{R}}_{1|2} = \mathbf{X}_1 - \mathbf{S}_{\mathbf{X}_1,\mathbf{X}_2}\mathbf{S}_{\mathbf{X}_2}^{-1}\mathbf{X}_2$ denote a typical residual from the ordinary least squares fit of $\mathbf{X}_1$ on $\mathbf{X}_2$, and let $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}_1\mathbf{S}_{\mathbf{X}_1,\mathbf{X}_2}\mathbf{S}_{\mathbf{X}_2}^{-1} + \boldsymbol{\beta}_2$. Then the partitioned model can be reexpressed as

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\beta}_1\widehat{\mathbf{R}}_{1|2} + \boldsymbol{\beta}_2^*\mathbf{X}_2 + \boldsymbol{\varepsilon}. \tag{1.17}$$

In this version of the partitioned model, the parameter vector $\boldsymbol{\beta}_1$ is the same as that in (1.16), while $\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_2^*$ unless $\mathbf{S}_{\mathbf{X}_1,\mathbf{X}_2} = 0$. The predictors – $\widehat{\mathbf{R}}_{1|2}$ and $\mathbf{X}_2$ – in (1.17) are uncorrelated in the sample $\mathbf{S}_{\widehat{\mathbf{R}}_{1|2},\mathbf{X}_2} = 0$, and consequently the maximum likelihood estimator of $\boldsymbol{\beta}_1$ is obtained by regressing $\mathbf{Y}$ on $\widehat{\mathbf{R}}_{1|2}$. The maximum likelihood estimator of $\boldsymbol{\beta}_1$ can also be obtained by regressing $\widehat{\mathbf{R}}_{\mathbf{Y}|2}$, the residuals from the regression of $\mathbf{Y}$ on $\mathbf{X}_2$, on $\widehat{\mathbf{R}}_{1|2}$. A plot of $\widehat{\mathbf{R}}_{\mathbf{Y}|2}$ versus $\widehat{\mathbf{R}}_{1|2}$ is called an added variable plot (Cook and Weisberg 1982). These plots are often used in univariate linear regression ($r = 1$) as general graphical diagnostics for visualizing how hard the data are working to fit individual coefficients.

Added variable plots and the partitioned forms of the multivariate linear model (1.16) and (1.17) will be used in this book from time to time, particularly in Chapter 3.

1.1.2 Alternative Model Forms

Because the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} > 0$ are unconstrained, model (1.1) allows each coordinate of $\mathbf{Y}$ to have a different linear regression on $\mathbf{X}$. It could be necessary in some applications to restrict the elements of $\boldsymbol{\beta}$ so that they are linked over rows and then iterations may be required because closed-form expressions for the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ will not normally be possible. The maximum likelihood estimator of $\boldsymbol{\beta}$ will be in the form of a weighted least squares estimator with weight matrix that depends on $\boldsymbol{\Sigma}$. For instance, suppose that we wish to restrict the coordinate regressions $\mathrm{E}(Y_j \mid \mathbf{X})$ to be parallel. This can be accomplished by requiring $\boldsymbol{\beta}$ to be a rank 1 matrix of the form $\boldsymbol{\beta} = \mathbf{1}_r\mathbf{b}^T$, where $\mathbf{b} \in \mathbb{R}^p$. Then $\boldsymbol{\beta}\mathbf{X} = \mathbf{1}_r\mathbf{X}^T\mathbf{b}$, and the model becomes $\mathbf{Y}_i = \boldsymbol{\alpha} + \mathbf{W}_i\mathbf{b} + \boldsymbol{\varepsilon}_i$, $i = 1, \dots, n$, where $\mathbf{W}_i = \mathbf{1}_r\mathbf{X}_i^T$.

For a second instance, consider a longitudinal study where n independent subjects are each observed at r times t_j, $j = 1, \dots, r$, with the elements of $\mathbf{Y}_i$ corresponding to the ordered sequence of r observations on subject i. Suppose further that $\mathrm{E}((\mathbf{Y}_i)_j \mid t_j) = \alpha + \mathbf{f}^T(t_j)\mathbf{b}$, where $\mathbf{f}(t) \in \mathbb{R}^p$ is a vector-valued user-specified function of time and $\mathbf{b} \in \mathbb{R}^p$. The elements of $\mathrm{E}(\mathbf{Y} \mid t_1, \dots, t_r)$ then correspond to points along the curve $\alpha + \mathbf{f}(t)^T\mathbf{b}$, and the linear model becomes

$$\mathbf{Y}_i = \alpha\mathbf{1} + \mathbf{W}\mathbf{b} + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{W} \in \mathbb{R}^{r\times p}$ with rows $\mathbf{f}^T(t_j)$, $j = 1, \dots, r$. Many variations of this model are available in the literature on longitudinal data (e.g. Weiss 2005). Again, the point here is that the maximum likelihood estimator of $\mathbf{b}$ will in general no longer be an ordinary least squares estimator because the parameters in the individual coordinate regressions are linked. Instead, the maximum likelihood estimator of $\mathbf{b}$ will be in the form of a weighted least squares estimator that depends on $\boldsymbol{\Sigma}$.

We will employ model (1.1) in this book, unless stated otherwise, although as discussed in Chapter 7, envelopes can apply regardless of the form of the model.

1.2 Envelope Model for Response Reduction

The motivation for response envelopes comes from allowing for the possibility that there are linear combinations of the response vector whose distribution is invariant to changes in the nonstochastic predictor vector. We refer to such linear combinations of $\mathbf{Y}$ as X-invariants. If X-invariants exist, then allowing for them in model (1.1) can result in substantial reduction in estimative variation. The linear transformation $\mathbf{G}^T\mathbf{Y}$, where $\mathbf{G} \in \mathbb{R}^{r\times q}$ with $q \leq r$, is X-invariant if and only if $\mathbf{A}^T\mathbf{G}^T\mathbf{Y}$ is X-invariant for any nonstochastic full-rank matrix $\mathbf{A} \in \mathbb{R}^{q\times q}$. Consequently, a specific transformation $\mathbf{G}$ is not identifiable but span($\mathbf{G}$) is identifiable, which leads us to consider subspaces rather than individual coordinates. The envelope model arises by parameterizing the multivariate linear model (1.1) in terms of the smallest subspace $\mathcal{E}$ of $\mathbb{R}^r$ with the properties that, for all relevant $\mathbf{x}_1$ and $\mathbf{x}_2$,

$$\text{(i) } \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (\mathbf{X} = \mathbf{x}_1) \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (\mathbf{X} = \mathbf{x}_2) \quad \text{and} \quad \text{(ii) } \mathbf{P}_{\mathcal{E}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid \mathbf{X}, \tag{1.18}$$

where $\mathbf{P}_{\mathcal{E}}$ is the projection onto $\mathcal{E}$ and $\mathbf{Q}_{\mathcal{E}} = \mathbf{I}_r - \mathbf{P}_{\mathcal{E}}$. These properties serve to identify parametrically the X-invariant part of $\mathbf{Y}$, $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$. Condition (i) stipulates that the marginal distribution of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ must be unaffected by changes in $\mathbf{X}$. It holds if and only if span($\boldsymbol{\beta}$) $\subseteq \mathcal{E}$, because then

$$\mathbf{Q}_{\mathcal{E}}\mathbf{Y} = \mathbf{Q}_{\mathcal{E}}\boldsymbol{\alpha} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\beta}\mathbf{X} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\varepsilon} = \mathbf{Q}_{\mathcal{E}}\boldsymbol{\alpha} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\varepsilon}.$$

Condition (ii) requires that $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ be unaffected by changes in $\mathbf{X}$ through an association with $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$, and it holds if and only if

$$\text{cov}(\mathbf{P}_{\mathcal{E}}\mathbf{Y}, \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid \mathbf{X}) = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{E}} = 0.$$

Conditions (i) and (ii) together imply that any dependence of $\mathbf{Y}$ on $\mathbf{X}$ must be concentrated in $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$, the X-variant part of $\mathbf{Y}$ that is material to the regression, while $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ is X-invariant and thus immaterial. The next two definitions, which do not require model (1.1), formalize the construction of an envelope in general.

Definition 1.1 A subspace $\mathcal{R} \subseteq \mathbb{R}^r$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{S}^{r\times r}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.

This definition of a reducing subspace is equivalent to that used by Cook et al. (2010), as described in Appendix A. It is common in the literature on invariant subspaces and functional analysis, although the underlying notion of "reduction" differs from the usual understanding in statistics. Here it is used to guarantee condition (ii) of (1.18) since the decomposition holds if and only if $\mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}} = 0$. The following definition makes use of reducing subspaces.

Definition 1.2 Let $\mathbf{M} \in \mathbb{S}^{r\times r}$ and let $\mathcal{B} \subseteq$ span($\mathbf{M}$). Then the $\mathbf{M}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$.

Definition 1.2 is the definition of an envelope introduced by Cook et al. (2007, 2010). It formalizes the construction of the smallest subspace that satisfies conditions (1.18) by asking for the intersection of all subspaces that envelop span($\boldsymbol{\beta}$), and thus that satisfy condition (i), from among those that satisfy condition (ii). Further discussion of this definition is available in Section A.2. We will often identify the subspace $\mathcal{B}$ as the span of a specified matrix $\mathbf{U}$: $\mathcal{B} = \text{span}(\mathbf{U})$. To avoid proliferation of notation in such cases, we will occasionally use the matrix as the argument to $\mathcal{E}_{\mathbf{M}}$: $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) := \mathcal{E}_{\mathbf{M}}(\text{span}(\mathbf{U}))$.

The actual construction of $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ can be characterized in terms of $\mathcal{B}$ and the spectral structure of $\mathbf{M}$. Suppose that $\mathbf{M}$ has $q \leq r$ distinct eigenvalues with projections onto the corresponding eigenspaces represented by $\mathbf{P}_k$, $k = 1, \ldots, q$. Then, as shown by Cook et al. (2010, see also Proposition A.2),

$$\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \sum_{k=1}^{q} \mathbf{P}_k \mathcal{B}. \tag{1.19}$$

This result shows that $\mathcal{B}$ is enveloped by using the eigenspaces of $\mathbf{M}$. Figure 1.1 gives a schematic representation of (1.19) when $r = 3$. The three axes of the plot represent three eigenvectors $\boldsymbol{\ell}_k$ of $\mathbf{M}$ with corresponding eigenvalues φ_k, $k = 1, 2, 3$. Three different possibilities for a one-dimensional $\mathcal{B}$ are represented by the spanning vectors $\boldsymbol{\beta}_j$, so $\mathcal{B}_j = \text{span}(\boldsymbol{\beta}_j)$, $j = 1, 2, 3$. The ordering of the eigenvalues is irrelevant for this discussion, but equality among the eigenvalues is relevant.

$\boldsymbol{\beta}_1$: Since $\boldsymbol{\beta}_1$ aligns with $\boldsymbol{\ell}_1$, $\text{span}(\boldsymbol{\beta}_1) = \text{span}(\boldsymbol{\ell}_1)$ and thus $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_1) = \mathcal{B}_1$ regardless of any equalities among the eigenvalues.

$\boldsymbol{\beta}_2$: Since $\boldsymbol{\beta}_2$ falls in the $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$-plane, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_2)$ depends on the corresponding eigenvalues. If $\varphi_1 = \varphi_2$, then regardless of the value for φ_3, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_2) = \mathcal{B}_2$. But if $\varphi_1 \neq \varphi_2$, then $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_2) = \text{span}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$.

$\boldsymbol{\beta}_3$: In the final case represented in Figure 1.1, $\boldsymbol{\beta}_3$ does not fall in any subspace spanned by a subset of the eigenvectors represented in the figure, except for the full space $\mathbb{R}^3 = \text{span}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3)$. Consequently, if the eigenvalues are distinct, then $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_3) = \mathbb{R}^3$. If $\varphi_1 \neq \varphi_2 = \varphi_3$, then there are two eigenspaces $\text{span}(\boldsymbol{\ell}_1)$ and $\text{span}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_3)$, and $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_3) = \text{span}(\boldsymbol{\ell}_1, \mathbf{P}_{2,3}\boldsymbol{\beta}_3)$, where $\mathbf{P}_{2,3}$ is the projection onto $\text{span}(\boldsymbol{\ell}_2, \boldsymbol{\ell}_3)$. If $\varphi_1 = \varphi_2 = \varphi_3$, then there is only one eigenspace and $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_3) = \mathcal{B}_3$.

Back to model (1.1), let $\mathcal{B} = \text{span}(\boldsymbol{\beta})$. The response projection $\mathbf{P}_{\mathcal{E}}$ is then defined as the projection onto $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, which by construction is the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\mathcal{B}$. Model (1.1) can be parameterized in terms of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ by using a basis. Let $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$, and let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r \times r}$ be an orthogonal matrix with $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$, $\text{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, and $\text{span}(\boldsymbol{\Gamma}_0) = \mathcal{E}_{\boldsymbol{\Sigma}}^{\perp}(\mathcal{B})$. Then the envelope model can be written as

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \tag{1.20}$$

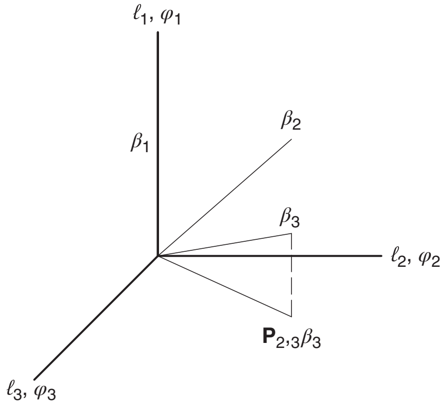


Figure 1.1 Schematic representation of an envelope.

The coefficient vector $\boldsymbol{\beta} = \boldsymbol{\Gamma\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ carries the coordinates of $\boldsymbol{\beta}$ relative to the basis matrix $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite matrices. To see how this model reflects the X-invariant part of $\mathbf{Y}$, multiply both sides by $\boldsymbol{\Gamma}_0^T$ to get $\boldsymbol{\Gamma}_0^T\mathbf{Y} = \boldsymbol{\Gamma}_0^T\boldsymbol{\alpha} + \boldsymbol{\Gamma}_0^T\boldsymbol{\varepsilon}$, where $\text{var}(\boldsymbol{\Gamma}_0^T\boldsymbol{\varepsilon}) = \boldsymbol{\Omega}_0$. We see from this representation that the marginal distribution of $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ does not depend on $\mathbf{X}$, so condition (i) of (1.18) is met. Further, since $\text{cov}(\boldsymbol{\Gamma}^T\mathbf{Y}, \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0 = 0$, we see also that condition (ii) of (1.18) holds.

The parameterization of (1.20) is intended to facilitate the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. We normally do not attempt to reify or infer about the constituent parameters $\boldsymbol{\Gamma}$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Omega}_0$. The values of $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Omega}_0$ depend on the choice of $\boldsymbol{\Gamma}$ to represent $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, and so may be difficult to interpret even if there is a desire to do so, while $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ depend on $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ but not on the particular basis. The basis matrix $\boldsymbol{\Gamma}$ is not identifiable in model (1.20) since, for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u\times u}$, replacing $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}\mathbf{O}$ leads to an equivalent model. For example, $\boldsymbol{\beta} = \boldsymbol{\Gamma\eta} = (\boldsymbol{\Gamma}\mathbf{O})(\mathbf{O}^T\boldsymbol{\eta})$, so replacing $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}\mathbf{O}$ and $\boldsymbol{\eta}$ with $\mathbf{O}^T\boldsymbol{\eta}$ leads to an equivalent expression for $\boldsymbol{\beta}$. However, the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \text{span}(\boldsymbol{\Gamma})$ is identifiable, which allows us to estimate $\boldsymbol{\beta}$. Additional properties of model (1.20) are discussed in Section 1.4.2.

Separate application of either of the two conditions in (1.18) does not necessarily lead to progress, but see the discussion of reduced-rank regression in Section 9.2.1. A subspace that reduces $\boldsymbol{\Sigma}$ may have no useful connection with $\mathcal{B}$. If $\mathcal{B}$ is a proper subspace of $\mathbb{R}^r$, then there can be many subspaces $\mathcal{S}$ that contain $\mathcal{B}$, while any particular subspace may not reduce $\boldsymbol{\Sigma}$. It is typically when the two conditions of (1.18) are used in concert that we obtain substantial gains in efficiency. The X-invariant component $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ induces extraneous variation into the maximum likelihood estimator of $\boldsymbol{\beta}$ under model (1.1). Envelopes distinguish between $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ and $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ in the estimation process and, as discussed in Section 1.5, the envelope estimator of $\boldsymbol{\beta}$ may then be more efficient than $\mathbf{B}$, as the variation from the X-invariant part of $\mathbf{Y}$ is effectively removed. Given the dimension u of the envelope and assuming normal errors, the envelope model can be fitted by maximum likelihood. The details of this are described later in Section 1.5; for now, we note that the maximum likelihood estimator of $\boldsymbol{\beta}$ is just the projection of $\mathbf{B}$ onto the estimated envelope, $\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\mathcal{E}}}\mathbf{B}$. The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is described in Section 1.6.

Definitions 1.1 and 1.2 are quite general – they do not require normality or even a linear model. Consequently, the envelope is still well defined if the linear model holds, but the errors are not normal. In that case, the maximum likelihood estimators under normality are still $\sqrt{n}$-consistent (Section 1.9), and the variance of $\widehat{\boldsymbol{\beta}}$ can be estimated by using the residual bootstrap (Section 1.11).

Figure 1.1 is applicable in the context of model (1.20) with $r = 3$ responses, $p = 1$ predictor, the $\boldsymbol{\ell}_j$s and φ_js interpreted as the eigenvectors and values of $\boldsymbol{\Sigma}$, and the $\boldsymbol{\beta}_j$s interpreted as three different possibilities for the coefficient vector. Those three configurations represent three ways in which an X-invariant can occur. Conversely, if part of $\mathbf{Y}$ is X-invariant, then a setting similar to those represented in the figure must hold.

The dimension $u \in \{0, 1, \dots, r\}$ can be selected based on any of the standard methods, including information criteria such as Akaike's information criterion (AIC) and Bayes information criterion (BIC), likelihood ratio testing, cross-validation, or a hold out sample, as described later in Section 1.10. If $u = 0$, then $\boldsymbol{\beta} = 0$ and there is no dependence of $\mathbf{Y}$ on $\mathbf{X}$, a setting that is often tested in practice. On the other extreme, if $u = r$, then $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathbb{R}^r$, the envelope model reduces to the usual model (1.1), and the distribution of every linear combination of $\mathbf{Y}$ responds to changes in $\mathbf{X}$, a conclusion that might also be useful in some applications.

1.3 Illustrations

Before expanding our discussion of the envelope model (1.20), we give in this section illustrations to provide intuition about the working mechanism of envelope estimation and its potential advantages over the standard estimators. These will also be used as running examples to illustrate various phases of an envelope analysis as they are developed in later sections.

1.3.1 A Schematic Example

Consider comparing the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of two bivariate normal populations, $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. This problem can be cast into the framework of model (1.1) by letting $\mathbf{Y} = (Y_1, Y_2)^T$ denote the bivariate response vector and letting X be an indicator variable taking value $X = 0$ in population 1 and $X = 1$ in population 2. Parameterizing so that $\boldsymbol{\alpha} = \boldsymbol{\mu}_1$ is the mean of the first population and $\boldsymbol{\beta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ is the mean difference, we have the multivariate linear model

$$\mathbf{Y} = \boldsymbol{\mu}_1 + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)X + \boldsymbol{\varepsilon} = \boldsymbol{\alpha} + \boldsymbol{\beta}X + \boldsymbol{\varepsilon}.$$

Starting with a sample, $(\mathbf{Y}_i, X_i)$, $i = 1, \dots, n$, having independent response vectors, the standard estimator of $\boldsymbol{\beta}$ is just the difference in the sample means for the two populations $\mathbf{B} = \bar{\mathbf{Y}}_2 - \bar{\mathbf{Y}}_1$ and the corresponding estimator of $\boldsymbol{\Sigma}$ is the pooled intra-sample covariance matrix. As in the general multivariate normal model (1.1), this estimator of $\boldsymbol{\beta}$ does not make use of the dependence between the responses and is equivalent to performing two univariate regressions of Y_j on X, $j = 1, 2$. Bringing envelopes into play can lead to a very different estimator of $\boldsymbol{\beta}$, one with substantially smaller variation than the maximum likelihood estimator $\mathbf{B}$ under model (1.1). In the remainder of this section, we illustrate one way this can happen.

A standard analysis will likely be sufficient when the populations are well separated, as illustrated in Figure 1.2, even with a larger number of responses. The two ellipses in that figure represent the two normal populations indicated by the predictor values $X = 0$ and $X = 1$. However, a standard analysis may not do as well when the populations are close, as illustrated in Figure 1.3. Without loss of generality, we set $\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 = 0$, so that $\mathcal{B} = \mathrm{span}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \mathrm{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathrm{span}(\boldsymbol{\beta})$. The left panel represents a standard likelihood analysis under model (1.1). For inference on β_2, the second element of $\boldsymbol{\beta}$, a standard analysis directly projects the data "y" onto the Y_2 axis following the dashed line marked A and then proceeds with inference based on the resulting univariate samples. The curves along the horizontal axis in the left panel stand for the projected distributions from the two populations. A standard analysis might involve constructing a two-sample t-test on samples drawn from these populations. There is a considerable overlap between the two projected distributions, so it may take a large sample size to infer that $\beta_2 \neq 0$ in a standard analysis. This illustration is based on β_2 to facilitate visualization; the same conclusions could be reached using a different linear combination of the elements of $\boldsymbol{\beta}$.

The two populations depicted in Figure 1.3 have the same eigenvectors, as they must because they have equal covariance matrices. We saw in (1.19) that an envelope can be constructed as $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \sum_{i=1}^{q} \mathbf{P}_i \mathcal{B}$. There are only two

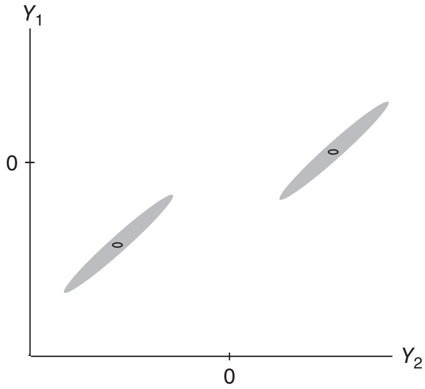


Figure 1.2 Graphical illustration of a relatively uncomplicated scenario. The axes are centered responses.

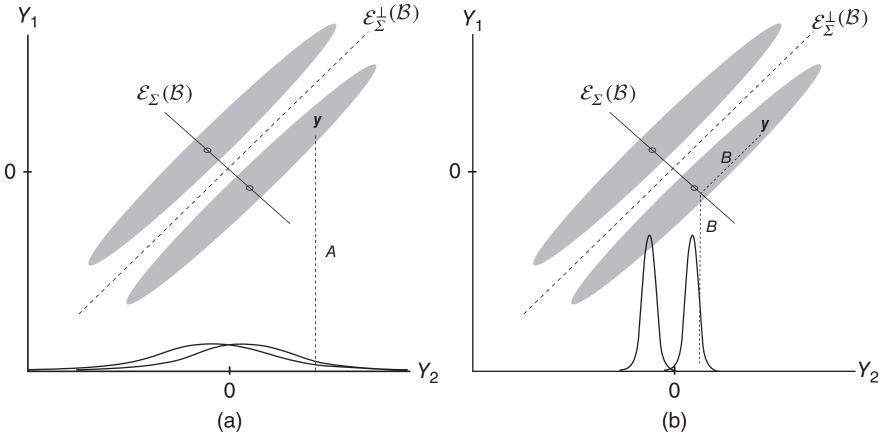

Y_1
$\mathcal{E}_{\Sigma}^{\perp}(\mathcal{B})$
$\mathcal{E}_{\Sigma}(\mathcal{B})$
y
0
A
Y_2
0
(a)
Y_1
$\mathcal{E}_{\Sigma}^{\perp}(\mathcal{B})$
$\mathcal{E}_{\Sigma}(\mathcal{B})$
y
B
0
B
Y_2
0
(b)

Figure 1.3 Graphical illustration envelope estimation. (a) Standard analysis; (b) envelope analysis.

eigenspaces in Figure 1.3, so the envelope must have dimension $u = 0$, $u = 1$, or $u = 2$. Since $\mathcal{B}$ equals the second eigenspace, we have $u = 1$ and $\mathcal{B} = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, although in higher dimensions only $\mathcal{B} \subseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is required. Accordingly, the eigenvector corresponding to the smaller eigenvalue is marked by the notation for the envelope $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, and the first eigenvector is marked by notation for the orthogonal complement of the envelope $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$. Condition (i) of (1.18) holds because the two populations have equal distributions when projected onto $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$; that is, $\mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (X = 0) \sim \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid (X = 1)$, where $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is shortened to $\mathcal{E}$ for use in subscripts. Since the populations are normal and $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$ are spanned by eigenvectors, we also have condition (ii) of (1.18), $\mathbf{P}_{\mathcal{E}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid X$.

The maximum likelihood envelope estimator of β_2 (see Section 1.5) can be formed by first projecting the data onto $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ to remove the X-invariant component $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ and extract the X-variant part of the response $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$, and then projecting onto the horizontal axis, as illustrated by the paths marked "B" in Figure 1.3b. Figure 1.3b also shows the resulting projected distributions corresponding to the projected distributions in Figure 1.3a. Now the projected distributions are well separated, and the envelope estimator of β_2 should be much more efficient than the estimator represented in Figure 1.3a. The estimative gain represented by passing from the standard estimator in Figure 1.3a to the envelope estimator in Figure 1.3b is reflected by the difference in the magnitude of the variances $\text{var}(\mathbf{\Gamma}^T\mathbf{Y} \mid X) = \mathbf{\Omega}$ and $\text{var}(\mathbf{\Gamma}_0^T\mathbf{Y} \mid X) = \mathbf{\Omega}_0$. The distributions in Figure 1.3 were constructed so that $\mathbf{\Omega} \ll \mathbf{\Omega}_0$, and consequently we anticipate substantial estimative gains.

There are two noteworthy caveats to our discussion of Figure 1.3. First, because the response is two-dimensional, the only nontrivial envelope must have dimension 1 and thus must align with one of the eigenvectors of $\mathbf{\Sigma}$. This is not required in higher dimensions, as shown by (1.19). Second, the illustration is based on the true envelope. The envelope needs to be estimated in practice, which has the effect of causing it to wobble in Figure 1.3b. That wobble produces increased variation in the envelope distributions of Figure 1.3b. As shown in Proposition 1.1, regardless of the degree of wobble, the asymptotic variance of the envelope estimator will not exceed the asymptotic variance of the standard estimator, which is reflected by the distributions in Figure 1.3a (Cook et al. 2010). In other words, the envelope estimator will always do at least as well as the standard maximum likelihood estimator asymptotically.

1.3.2 Compound Symmetry

Suppose that the elements $(\mathbf{\Sigma})_{ij}$ of the residual covariance matrix for envelope model (1.20) have the form $(\mathbf{\Sigma})_{ii} = \sigma^2$ and $(\mathbf{\Sigma})_{ij} = \sigma^2\rho$ for $-(r-1)^{-1} < \rho < 1$ and $i \neq j$. Then $\mathbf{\Sigma}$ can be decomposed as

$$\begin{aligned}\mathbf{\Sigma}\sigma^{-2} &= \rho\mathbf{1}_r\mathbf{1}_r^T + (1-\rho)\mathbf{I}_r\\ &= (1+(r-1)\rho)\mathbf{P}_{\mathbf{1}_r} + (1-\rho)\mathbf{Q}_{\mathbf{1}_r}.\end{aligned}$$

In consequence, $\mathbf{\Sigma}$ has two eigenspaces $\mathrm{span}(\mathbf{1}_r)$ and $\mathrm{span}^{\perp}(\mathbf{1}_r)$ with corresponding eigenvalues $\varphi_1 = \sigma^2(1+(r-1)\rho)$ and $\varphi_2 = \sigma^2(1-\rho)$. It follows from (1.19) that the envelope is the sum of the projections of $\mathcal{B}$ onto $\mathrm{span}(\mathbf{1}_r)$ and $\mathrm{span}^{\perp}(\mathbf{1}_r)$: $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}(\mathbf{P}_{\mathbf{1}_r}\boldsymbol{\beta}) + \mathrm{span}(\mathbf{Q}_{\mathbf{1}_r}\boldsymbol{\beta})$, which has dimension at most $1 + \min(p, r-1)$.

Compound symmetry, which is a common name given to the covariance structure considered in this section, is frequently used for analysis of longitudinal data where the response vector is made up of repeated measures on a single individual over time (Weiss 2005), as in the cattle weights illustration introduced in Section 1.3.4.

1.3.3 Wheat Protein: Introductory Illustration

The classic wheat protein data (Fearn 1983) contain measurements on protein content and the logarithms of near-infrared reflectance at six wavelengths across the range 1680–2310 nm measured on each of $n = 50$ samples of ground wheat. To illustrate the ideas associated with Figure 1.3 in data analysis, we use $r = 2$ wavelengths as responses $\mathbf{Y} = (Y_1, Y_2)^T$ and convert the continuous measure of protein content into a categorical predictor X, indicating low and high protein (24 and 26 samples, respectively).

The mean difference $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ corresponds to the parameter vector $\boldsymbol{\beta}$ in model (1.1), with X representing a binary indicator: $X = 0$ for high protein and $X = 1$ for low protein wheat. For these data, which are shown in Figure 1.4, $\mathbf{B}^T = (7.5, -2.1)$ with standard errors (SEs) 8.6 and 9.5 (Figure 1.3a). There is

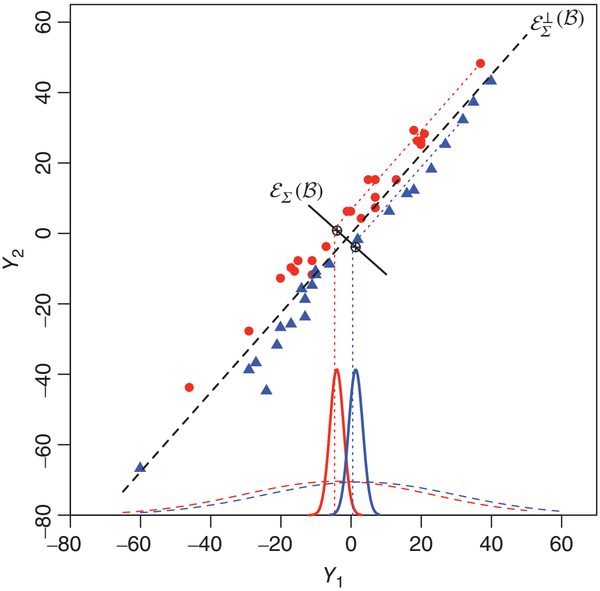


Figure 1.4 Wheat protein data with the estimated envelope superimposed.

no indication from these marginal results that $\mathbf{Y}$ depends on X, while the likelihood ratio test statistic (1.13) has the value 27.5 on 2 degrees of freedom and thus indicates otherwise. The simultaneous occurrence in a standard analysis of relatively small Z-scores and a relatively large likelihood ratio statistic is often a clue that an envelope analysis will be advantageous, although these conditions are certainly not necessary.

The envelope estimate is $\widehat{\boldsymbol{\beta}}^T = (5.1, -4.7)$ with asymptotic standard errors of 0.51 and 0.46 (Figure 1.3b). To more fully appreciate the magnitude of this drop in standard errors, we would need for a standard analysis a sample size of $n \sim 20\ 000$ to reduce the standard error from 9.4 to 0.46. Figure 1.4 shows the projected distributions of the data like those from Figure 1.3.

1.3.4 Cattle Weights: Initial Fit

Roundworm is an intestinal parasite that takes nutrients from animals and lowers their resistance to disease, resulting in relatively low body weights. Kenward (1987) presented data[1] from a randomized experiment to compare two treatments for controlling roundworm in cattle. The two treatments were randomly assigned to 60 cows, with 30 cows per treatment. Their weights in

[1] The data were obtained from the website for Weiss (2005) http://rem.ph.ucla.edu/~rob/mld/data.html.

kilograms were recorded at the beginning of the study prior to treatment application and at 10 times during the study corresponding to weeks $2, 4, 6, \ldots, 18$ and 19; that is, at two-week intervals except the last that was over a one-week interval. The goal of the experiment was to determine if the treatments have a differential effect on weight and, if so, to estimate the time point at which the difference was first manifested. The experimenter evidently anticipated a lag between the time of treatment application and manifestation of their effects.

The basal weights, which were taken before treatment, could be used as covariates in an effort to control some of the variations between animals. In this illustration, we take a simpler approach and neglect the basal weights in our analysis. (Basal weights are included as covariates in an illustrative analysis in Chapter 3.) Since treatments were assigned randomly this should not introduce a bias, although we might incur variation that could otherwise be removed. Figure 1.5a shows a profile plot, also called a parallel coordinate plot, of animal weight against week, beginning with the first posttreatment measurements in week 2. Each line traces the weight of an animal over time, with the two treatments represented by different colors. Profile plots are useful for seeing clusters over time, but for these data no clusters or treatment effects seem apparent visually. Figure 1.5b shows a profile plot of mean weight by week and treatment on the same scale as Figure 1.5a. The two mean profiles are roughly parallel until week 12, cross between weeks 12 and 14, and then cross back between weeks 18 and 19. Judged against the variation of the individual profiles in Figure 1.5a and recognizing that the intra-animal weights are surely positively correlated, these visual representations hint that there is no detectable difference between the two treatments from these data.

Let $\mathbf{Y}_i \in \mathbb{R}^{10}$, $i = 1, \ldots, 60$, be the vector of weight measurements for each animal over time, and let $X_i = -0.5$ or 0.5 indicate the two treatments, so $\sum_{i=1}^{60} X_i = 0$. Our interest lies in the regression coefficient $\boldsymbol{\beta}$ from the multivariate linear regression $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta} X + \boldsymbol{\varepsilon}$, where it is still assumed that $\boldsymbol{\varepsilon} \sim N_{10}(0, \boldsymbol{\Sigma})$. Recall that $\mathbf{B}$ denotes the ordinary least squares estimator of $\boldsymbol{\beta}$, which is also the maximum likelihood estimator under normality. For these data $\mathbf{B}$ is simply the difference in the mean vectors for the two treatments,

$$\mathbf{B} = \bar{\mathbf{Y}} \mid (X = 0.5) - \bar{\mathbf{Y}} \mid (X = -0.5),$$

and the corresponding estimator of $\boldsymbol{\alpha}$ is the grand mean $\bar{\mathbf{Y}}$. The plot in Figure 1.5b can also be described as a profile plot of the fitted weights $\widehat{\mathbf{Y}} = \bar{\mathbf{Y}} + \mathbf{B}X$. The coefficient estimates $\mathbf{B}$ and their Z-scores (1.10) are given in the second and third columns of Table 1.1. This fit supports the visual impression from Figure 1.5, as the largest absolute Z-score is only 1.30. On the other hand, the likelihood ratio statistic (1.13) for the hypothesis $\boldsymbol{\beta} = 0$ has the value 26.9 on 10 degrees of freedom, which suggests differences somewhere that are not manifested by the marginal Z-scores of Table 1.1. As mentioned in the wheat protein illustration, the simultaneous occurrence of relatively small

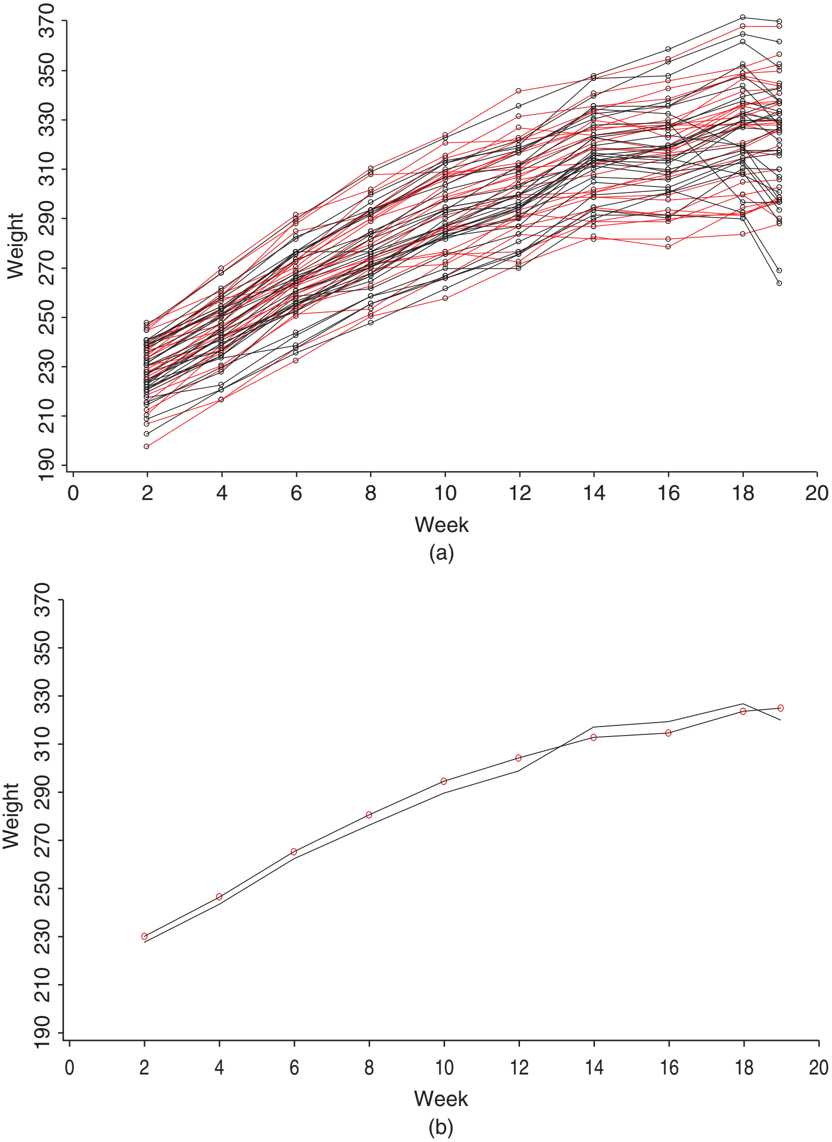


Figure 1.5 Cattle data: profile plots of weight for 60 cows over 10 time periods. Colors designate the two treatments. (a) Profile plots for individual cows by treatment. (b) Profile plots of average weight by treatment.

Table 1.1 Cattle data: Estimated coefficients for the standard model and envelope model with $u = 1$.

Week	B	Z-score	$\widehat{\beta}$	Z-score	R
2	2.43	0.83	−2.17	−2.48	3.32
4	3.33	1.05	−0.48	−0.65	4.27
6	3.13	0.89	0.88	1.23	4.89
8	4.73	1.22	2.38	2.82	4.56
10	4.73	1.14	2.89	4.14	5.94
12	5.50	1.30	5.40	5.30	4.15
14	−4.80	−1.11	−5.09	−5.55	4.69
16	−4.53	−0.97	−4.62	−5.36	5.40
18	−2.87	−0.54	−3.67	−4.06	5.86
19	5.00	0.86	4.21	4.92	6.78

The Z-score is the estimate divided by its standard error (se$(\cdot)$) as defined in (1.10), and $R = \mathrm{se}\{(\mathbf{B})_j\}/\mathrm{se}\{(\widehat{\boldsymbol{\beta}})_j\}$.

Z-scores and a relatively large likelihood ratio statistic is often a good clue that an envelope analysis will offer advantages, although again these conditions are not necessary. We next turn to an envelope analysis.

Using the LRT(.05) method of Section 1.10, we first estimated that $u = 1$, so it was inferred that the treatment difference is manifested in only one linear combination $\boldsymbol{\Gamma}^T\mathbf{Y}$ of the response vector. Then with $u = 1$, we estimated a basis $\widehat{\boldsymbol{\Gamma}}$ for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ using the method of Section 1.5. The fourth and fifth columns of Table 1.1 give the envelope coefficient estimates $\widehat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{B}$ and their Z-scores determined by using the standard errors described in (1.33). Making use of a Bonferroni inequality at level 0.05 to adjust for multiple testing, differential treatment effects are indicated by at least week 10 and persist thereafter. The final column of Table 1.1 gives the ratios of the standard errors for the elements of $\mathbf{B}$ to those of $\widehat{\boldsymbol{\beta}}$. We see from these results that the standard errors for the elements of $\mathbf{B}$ were 3.32–6.78 times those of $\widehat{\boldsymbol{\beta}}$. These ratios represent a substantial increase in efficiency of the envelope estimator relative to the usual maximum likelihood estimator. To more fully appreciate the magnitude of this drop in standard errors, we would need $n \sim 1500$ for a standard analysis to achieve the standard errors from an envelope analysis with $n = 60$.

Figure 1.6 shows a profile plot on the same scale as Figure 1.5 of the fitted weights, $\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}X$ from the envelope model with $u = 1$. Comparing to Figure 1.5b, the corresponding profile plot of the fitted weights from the standard model, we see that the two plots agree well after about week 10, but before then the fitted weights for the two treatments are closer from the

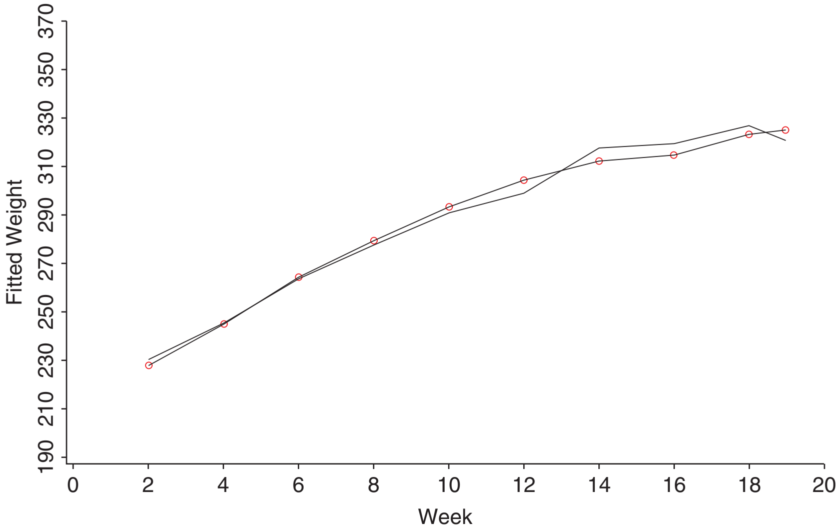


Figure 1.6 Cattle data: profile plot of fitted weights from the envelope model with $u = 1$. Colors designate the two treatments.

envelope fit than from the standard fit. This supports the prior notion of a lag between treatment application and effect.

This example has 10 responses, so an overall graphical construction like that shown in Figure 1.3 is not possible. However, marked plots of the weights for week w versus the weights for week $w + 2$ provide some intuition on the structure of the data. For instance, the plot for week 12 weight versus week 14 weight given in Figure 1.7 suggests a clear difference in weights and exhibits the

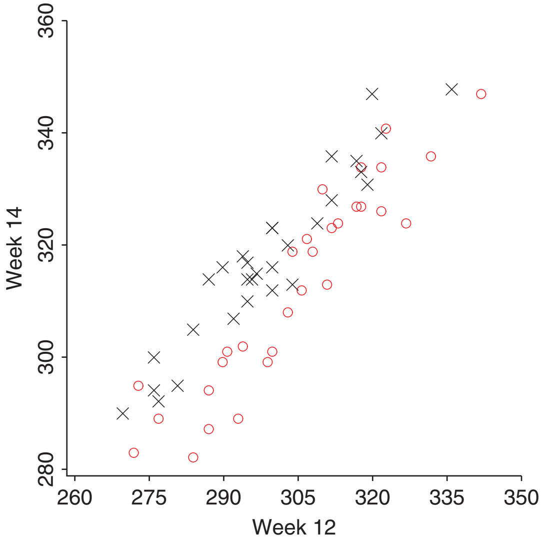


Figure 1.7 Cattle data: scatterplot of week 12 weight versus week 14 weight with treatments marked.

envelope structure represented schematically in Figure 1.3. A formal envelope analysis of these bivariate data indicates that $u = 1$ (Section 1.10), leading to envelope standard errors that are about 5.7 times smaller than those from the standard model and highly significant coefficient estimates. This level of reduction is commensurate with those shown previously in Table 1.1.

1.4 More on the Envelope Model

In this section, we revisit and expand upon the envelope model (1.20) introduced in Section 1.2.

1.4.1 Relationship with Sufficiency

Suppose we know a subspace $\mathcal{S}$ that contains $\mathcal{B}$, so that condition (i) of (1.18) holds; we are not yet enforcing condition (ii) of (1.18). The estimator of $\boldsymbol{\beta}$ based on $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$ is $\mathbf{P}_{\mathcal{S}}\mathbf{B} = \mathbf{S}_{\mathbf{P}_{\mathcal{S}}\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}$. If $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$ is to capture the part of $\mathbf{Y}$ that is material to the estimation of $\boldsymbol{\beta}$, then it is reasonable to require that the conditional distribution of $\mathrm{vec}(\mathbf{B}) \mid \mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B})$ not depend on $\boldsymbol{\beta}$.

Since $\mathrm{vec}(\mathbf{B}) \mid \mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B})$ is normally distributed, we need to consider only its mean and variance. Its variance does not depend on $\boldsymbol{\beta}$, but its mean may do so. Writing $\mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B}) = (\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}})\mathrm{vec}(\mathbf{B})$, we have from (1.11) that

$$\mathrm{E}\{\mathrm{vec}(\mathbf{B}) \mid \mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B})\} = \mathrm{vec}(\boldsymbol{\beta}) + \mathbf{P}^T_{\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}(\mathbf{V})}\{\mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\boldsymbol{\beta})\},$$

where $\mathbf{V} = \mathrm{var}\{\mathrm{vec}(\mathbf{B})\}$ as given in (1.8). Because $\mathbf{P}_{\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}(\mathbf{V})}(\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}) = \mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}$ and $\mathcal{B} \subseteq \mathcal{S}$, we have

$$\begin{aligned}\mathrm{E}\{\mathrm{vec}(\mathbf{B}) \mid \mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B})\} &= (\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}} + \mathbf{I}_p \otimes \mathbf{Q}_{\mathcal{S}})\mathrm{E}\{\mathrm{vec}(\mathbf{B}) \mid \mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B})\} \\ &= \mathrm{vec}(\boldsymbol{\beta}) + (\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}})\mathbf{P}^T_{\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}(\mathbf{V})}\{\mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\boldsymbol{\beta})\} \\ &\quad + (\mathbf{I}_p \otimes \mathbf{Q}_{\mathcal{S}})\mathbf{P}^T_{\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}(\mathbf{V})}\{\mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\boldsymbol{\beta})\} \\ &= \mathrm{vec}(\mathbf{P}_{\mathcal{S}}\mathbf{B}) + (\mathbf{I}_p \otimes \mathbf{Q}_{\mathcal{S}})\mathbf{P}^T_{\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}(\mathbf{V})}\{\mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\boldsymbol{\beta})\}.\end{aligned}$$

The second term on the right-hand side will be free of $\boldsymbol{\beta}$ if and only if $\mathbf{P}_{\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}(\mathbf{V})}(\mathbf{I}_p \otimes \mathbf{Q}_{\mathcal{S}}) = 0$; that is, if and only if $\mathbf{I}_p \otimes \mathbf{P}_{\mathcal{S}}$ is orthogonal to $\mathbf{I}_p \otimes \mathbf{Q}_{\mathcal{S}}$ in the $\mathbf{V}$ inner product. This holds if and only if $\mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}} = 0$, so $\mathcal{S}$ must reduce $\boldsymbol{\Sigma}$. Consequently, we are led back to condition (ii) of (1.18).

The general point here is that conditions (i) and (ii) of (1.18) are designed to insure that $\mathbf{P}_{\mathcal{S}}\mathbf{B}$ is sufficient for $\boldsymbol{\beta}$ when $\mathcal{S}$ is known; that is, $\mathbf{P}_{\mathcal{S}}\mathbf{Y}$ is a sufficient reduction of $\mathbf{Y}$. The context here is distinct from classical sufficiency because $\mathcal{S}$ is unknown and must be estimated.

1.4.2 Parameter Count

The total number of real parameters required for the envelope model (1.20) is

$$\begin{aligned}N_u &= r + pu + u(r-u) + u(u+1)/2 + (r-u)(r-u+1)/2 \qquad (1.21) \\ &= r + pu + r(r+1)/2.\end{aligned}$$

This count arises as follows. The first addend on the right-hand side of (1.21) corresponds to the intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$, and the second addend corresponds to the unconstrained coordinate matrix $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$. The last two addends correspond to $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Their parameter counts arise because, for any integer $k > 0$, it takes $k(k+1)/2$ numbers to specify a nonsingular $k \times k$ symmetric matrix. The third addend $u(r-u)$, which gives the parameter count for $\boldsymbol{\Gamma}$, arises as follows. As mentioned previously, the matrix $\boldsymbol{\Gamma}$ is not identifiable since, for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u\times u}$, replacing $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}\mathbf{O}$ results in an equivalent model. However, $\text{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is identifiable. The parameter space for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is the Grassmannian $\mathcal{G}(u,r)$ of dimension u in $\mathbb{R}^r$: $\mathcal{G}(u,r)$ is the collection of all u-dimensional subspaces of $\mathbb{R}^r$. From basic properties of Grassmann manifolds it is known that $u(r-u)$ real parameters are needed to specify an element of $\mathcal{G}(u,r)$ uniquely. Once $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is determined, so is its orthogonal complement $\text{span}(\boldsymbol{\Gamma}_0)$, and no additional free parameters are required. The difference between the total parameter count for the standard model (1.1) with $r = u$ and the envelope model (1.20) with $u < r$ is therefore $p(r-u)$. Further discussion of Grassmannians is available from Edelman et al. (1998).

The number of real parameters $u(r-u)$ needed to identify a subspace can be seen intuitively as follows. ru parameters are needed to specify uniquely an unconstrained matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$. But when dealing with subspaces, $\text{span}(\boldsymbol{\Gamma}) = \text{span}(\boldsymbol{\Gamma}\mathbf{A})$ for any nonsingular $\mathbf{A} \in \mathbb{R}^{u\times u}$. Since it requires u^2 parameters to determine an $\mathbf{A}$, specifying a subspace takes $ru - u^2 = u(r-u)$ parameters.

1.4.3 Potential Gains

A specific envelope model is identified by the value of u. All envelope models are nested within the standard model (1.1), but two envelope models with different values of u are not necessarily nested. To see this, notice that the number of real parameters needed to specify an element of $\mathcal{G}(1,r)$ is the same as that for $\mathcal{G}(r-1,r)$. If $u = r$, then $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathbb{R}^r$, the envelope model degenerates to the standard model (1.1) and enveloping offers no gain. If $r \leq p$ and $\dim(\mathcal{B}) = r$, then again the envelope model reduces to the standard model. However, if (i) $r > p$ or (ii) if $\dim(\mathcal{B}) < r \leq p$, then efficiency gains are possible. These gains can arise in two ways. The first is through the parameter count. Since the number of parameters in the envelope model is less than that in the standard model, we can expect some efficiency gains from parsimony. But the second source is where we have the potential to realize massive gains.

To explore envelope gains and in anticipation of maximum likelihood estimation discussed in Section 1.5, consider the maximum likelihood estimator of $\boldsymbol{\beta}$ when $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is known and represented by a semi-orthogonal basis matrix $\boldsymbol{\Gamma}$. In this case, the maximum likelihood estimator of $\boldsymbol{\beta}$ under the envelope model

(1.20) is again the projection of the standard estimator onto $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$: $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{B}$. Using (1.8), the variance of this estimator is

$$\begin{aligned}\text{var}\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\} &= \text{var}\{\text{vec}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{B})\}\\ &= (\mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}})\text{var}\{\text{vec}(\mathbf{B})\}(\mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}})\\ &= n^{-1}(\mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}})(\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma})(\mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}})\\ &= n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma\Omega\Gamma}^T. \end{aligned} \tag{1.22}$$

Comparing this to the variance of the standard estimator $\mathbf{B}$,

$$\text{var}\{\text{vec}(\mathbf{B})\} - \text{var}\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\} = n^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T \geq 0. \tag{1.23}$$

From this we conclude that if the variance of the X-invariant part of $\mathbf{Y}$, $\text{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}) = \boldsymbol{\Omega}_0$ is large relative to the variance of the X-variant part of $\mathbf{Y}$, $\text{var}(\boldsymbol{\Gamma}^T\mathbf{Y}) = \boldsymbol{\Omega}$, then the gain from the envelope model can be substantial. Using the spectral norm $\|\cdot\|$ as a measure of overall size, the envelope model may be particularly advantageous when $\|\boldsymbol{\Gamma\Omega\Gamma}^T\| = \|\boldsymbol{\Omega}\| \ll \|\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\| = \|\boldsymbol{\Omega}_0\|$. The envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ will normally be estimated in practice, and these results will then be mitigated by the variability in its estimator. Nevertheless, experience has shown that they are a useful indicator of the kinds of regressions in which envelopes offer substantial gains.

1.5 Maximum Likelihood Estimation

1.5.1 Derivation

In this section, we discuss the derivation of the maximum likelihood estimators of the parameters in envelope model (1.20), assuming that the dimension $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$ of the envelope is known.

The log-likelihood $L_u(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ under model (1.20) with known u can be expressed as

$$\begin{aligned} L_u &= -(nr/2)\log(2\pi) - (n/2)\log|\boldsymbol{\Gamma\Omega\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T|\\ &\quad -(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma\eta}\mathbf{X}_i)^T(\boldsymbol{\Gamma\Omega\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma\eta}\mathbf{X}_i)\\ &= -(nr/2)\log(2\pi) - (n/2)\log|\boldsymbol{\Omega}| - (n/2)\log|\boldsymbol{\Omega}_0|\\ &\quad -(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma\eta}\mathbf{X}_i)^T(\boldsymbol{\Gamma\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\Gamma\eta}\mathbf{X}_i), \end{aligned}$$

where the second equality arises by applying the third and fourth conclusions of Corollary A.1. Also, while the likelihood function depends on $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, we have written it in terms of the semi-orthogonal basis matrix $\boldsymbol{\Gamma}$ to facilitate the derivation. The original derivation by Cook et al. (2010) uses projections instead of bases.

The maximum likelihood estimator of $\boldsymbol{\alpha}$ is $\hat{\boldsymbol{\alpha}} = \bar{\mathbf{Y}}$ because the predictors are centered. Substituting this into the likelihood function and then decomposing $\mathbf{Y}_i - \bar{\mathbf{Y}} = \mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{Y}_i - \bar{\mathbf{Y}}) + \mathbf{Q}_{\boldsymbol{\Gamma}}(\mathbf{Y}_i - \bar{\mathbf{Y}})$ and simplifying, we arrive at the first partially maximized log-likelihood,

$$\begin{aligned} L_u^{(1)}(\boldsymbol{\eta}, \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) = &-(nr/2)\log(2\pi) + L_u^{(11)}(\boldsymbol{\eta}, \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}) \\ &+ L_u^{(12)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}_0), \end{aligned}$$

where

$$\begin{aligned} L_u^{(11)} = &-(n/2)\log|\boldsymbol{\Omega}| \\ &-(1/2)\sum_{i=1}^{n}\{\boldsymbol{\Gamma}^T(\mathbf{Y}_i - \bar{\mathbf{Y}}) - \boldsymbol{\eta}\mathbf{X}_i\}^T\boldsymbol{\Omega}^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{Y}_i - \bar{\mathbf{Y}}) - \boldsymbol{\eta}\mathbf{X}_i\}, \\ L_u^{(12)} = &-(n/2)\log|\boldsymbol{\Omega}_0| - (1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbf{Y}_i - \bar{\mathbf{Y}}). \end{aligned}$$

Holding $\boldsymbol{\Gamma}$ fixed, $L_u^{(11)}$ can be seen as the log-likelihood for the multivariate regression of $\boldsymbol{\Gamma}^T(\mathbf{Y}_i - \bar{\mathbf{Y}})$ on $\mathbf{X}_i$, and thus $L_u^{(11)}$ is maximized over $\boldsymbol{\eta}$ at the value $\boldsymbol{\eta} = \boldsymbol{\Gamma}^T\mathbf{B}$. Substituting this into $L_u^{(11)}$ and simplifying, we obtain a partially maximized version of $L_u^{(11)}$

$$L_u^{(21)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}) = -(n/2)\log|\boldsymbol{\Omega}| - (1/2)\sum_{i=1}^{n}(\boldsymbol{\Gamma}^T\mathbf{r}_i)^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{r}_i,$$

where, as defined in Section 1.1, $\mathbf{r}_i$ is the ith residual vector from the fit of the standard model (1.1). From this it follows immediately that, still with $\boldsymbol{\Gamma}$ fixed, $L_u^{(21)}$ is maximized over $\boldsymbol{\Omega}$ at $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}$. Consequently, we arrive at the third partially maximized log-likelihood $L_u^{(31)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = -(n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| - nu/2$. By similar reasoning, the value of $\boldsymbol{\Omega}_0$ that maximizes $L_u^{(21)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}), \boldsymbol{\Omega}_0)$ is $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0$. This leads to the maximization of $L_u^{(22)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = -(n/2)\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0| - n(r-u)/2$.

Combining the above steps, we arrive at the partially maximized form

$$\begin{aligned} L_u^{(2)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = &-(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| \\ &-(n/2)\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0|. \end{aligned}$$

Next, since $\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0| = \log|\mathbf{S}_{\mathbf{Y}}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}|$ (Lemma A.13), we can express $L_u^{(2)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$ as a function of $\boldsymbol{\Gamma}$ alone:

$$\begin{aligned} L_u^{(2)}(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})) = &-(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}}| \\ &-(n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| - (n/2)\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}|. \end{aligned} \quad (1.24)$$

Summarizing, the maximum likelihood estimators $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and of the remaining parameters are determined as

$$\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \text{span}\{\arg\min_{\mathbf{G}} (\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|)\}, \tag{1.25}$$

$$\begin{aligned}\widehat{\boldsymbol{\eta}} &= \widehat{\mathbf{\Gamma}}^T\mathbf{B},\\ \widehat{\boldsymbol{\beta}} &= \widehat{\mathbf{\Gamma}}\,\widehat{\boldsymbol{\eta}} = \mathbf{P}_{\widehat{\mathcal{E}}}\mathbf{B},\\ \widehat{\mathbf{\Omega}} &= \widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\mathbf{\Gamma}},\\ \widehat{\mathbf{\Omega}}_0 &= \widehat{\mathbf{\Gamma}}_0^T\mathbf{S}_{\mathbf{Y}}\widehat{\mathbf{\Gamma}}_0,\\ \widehat{\mathbf{\Sigma}} &= \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{\Gamma}}^T + \widehat{\mathbf{\Gamma}}_0\widehat{\mathbf{\Omega}}_0\widehat{\mathbf{\Gamma}}_0^T,\end{aligned} \tag{1.26}$$

where $\min_{\mathbf{G}}$ is over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{r\times u}$, $\widehat{\mathbf{\Gamma}}$ is any semi-orthogonal basis matrix for $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$, and $\widehat{\mathbf{\Gamma}}_0$ is any semi-orthogonal basis matrix for the orthogonal complement of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$. The fully maximized log-likelihood for fixed u is then

$$\begin{aligned}\widehat{L}_u = &-(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}}|\\ &-(n/2)\log|\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\mathbf{\Gamma}}| - (n/2)\log|\widehat{\mathbf{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}^{-1}\widehat{\mathbf{\Gamma}}|.\end{aligned} \tag{1.27}$$

The optimization required in (1.25) can be sensitive to starting values. Discussions of starting values and optimization algorithms are given in Chapter 6 within a broader context. The estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{\Sigma}}$ are invariant to the selection of basis $\widehat{\mathbf{\Gamma}}$ and thus are unique, but the remaining estimators $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Omega}}_0$ are basis dependent and thus not unique.

1.5.2 Cattle Weights: Variation of the *X*-Variant Parts of Y

The envelope estimates in Table 1.3.4 were constructed based on (1.25) and the corresponding estimates of the other parameters. We commented at the end of Section 1.4.2 that an envelope analysis will be particularly advantageous when $||\mathbf{\Omega}||$ is substantially smaller than $||\mathbf{\Omega}_0||$. From the fit of the envelope model with $u = 1$ to the cattle data, we obtain $||\widehat{\mathbf{\Omega}}|| = 27.8$ and $||\widehat{\mathbf{\Omega}}_0|| = 2351.5$, which conforms qualitatively to the gains in Table 1.1 and the structure shown in Figure 1.7.

In an envelope model with $u = 1$, $\boldsymbol{\beta}$ is constrained to lie in a one-dimensional reducing subspace of $\mathbf{\Sigma}$. If the eigenvalues of $\mathbf{\Sigma}$ are all distinct, then $\boldsymbol{\beta}$ will align with an eigenvector of $\mathbf{\Sigma}$, as described previously in (1.19). The estimate $\widehat{\mathbf{\Gamma}}$ will not correspond to an eigenvector of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$, but informally we might expect $\widehat{\mathbf{\Gamma}}$ to fall close to an eigenvector because of the result shown in (1.19). We can often obtain intuition about the fit of an envelope model by considering the correlations between $\widehat{\mathbf{\Gamma}}^T\mathbf{Y}$ and the elements of $\mathbf{L}^T\mathbf{Y}$, where the columns of $\mathbf{L} = (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_r)$ are the ordered eigenvectors of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. For the cattle data, the two largest absolute correlations 0.97 and 0.48 occur with the fourth and sixth eigenvectors, giving $\boldsymbol{\ell}_4^T\mathbf{Y}$ and $\boldsymbol{\ell}_6^T\mathbf{Y}$. Evidently, the treatment differences are largely associated with the fourth eigenvector of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$.

1.5.3 Insights into $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$

In this section, we provide different forms for $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ that might aid intuition. First, $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B})$ can be reexpressed as

$$\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \arg\min_{\mathcal{S}\in\mathcal{G}(u,r)} \{\log|\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathcal{S}}|_0 + \log|\mathbf{Q}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\mathcal{S}}|_0\}, \tag{1.28}$$

where $|\cdot|_0$ denotes the product of the nonzero eigenvalues of the matrix argument, and minimization $\min_{\mathcal{S}\in\mathcal{G}(u,r)}$ is over the Grassmannian $\mathcal{G}(u,r)$ of dimension u in $\mathbb{R}^r$. This form highlights the fact that only the subspace is being estimated, not a particular basis.

We next reexpress the objective function in (1.28) to show how it reflects the constraints on $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ in envelope model (1.20). Let $\mathbf{G}$ and $\mathbf{G}_0$ be semi-orthogonal basis matrices for a subspace $\mathcal{S}$ and its orthogonal complement $\mathcal{S}^{\perp}$. Then it follows from Lemma A.14 that $|\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathcal{S}}|_0 = |\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}|$ and $|\mathbf{Q}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\mathcal{S}}|_0 = |\mathbf{G}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}_0|$. Consequently,

$$\begin{aligned}\log|\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathcal{S}}|_0 + \log|\mathbf{Q}_{\mathcal{S}}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\mathcal{S}}|_0 &= \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}_0| \\ &= \log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| + \log|\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}}|.\end{aligned}$$

From (1.5), we can express

$$\begin{aligned}\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}} &= \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} + \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}\circ\mathbf{X}} \\ &= \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} + \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y},\mathbf{X}}^T \\ &= \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} + \mathbf{B}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\mathbf{S}_{\mathbf{X}}\mathbf{B}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^T \\ &= \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} + \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^{1/2}\tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^T\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^{1/2},\end{aligned}$$

where

$$\tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^{-1/2}\mathbf{B}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{1/2}$$

is the standardized version of the estimated coefficient matrix $\mathbf{B}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}$ from the fit of $\mathbf{G}_0^T\mathbf{Y}$ on $\mathbf{X}$ (1.6). Consequently,

$$|\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}}| = |\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}| \times \left|\mathbf{I}_{r-u} + \tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^T\right|$$

and optimization (1.28) can now be reexpressed as

$$\begin{aligned}\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \operatorname{span}\Bigg\{\arg\min_{\mathbf{G}}\Bigg(&\log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| + \log|\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}| \\ &+ \log\left|\mathbf{I}_{r-u} + \tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\tilde{\mathbf{B}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}^T\right|\Bigg)\Bigg\},\end{aligned} \tag{1.29}$$

where the minimization is over semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{r\times u}$. Since $|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| \times |\mathbf{S}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}| \geq |\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|$ with equality if and only if span($\mathbf{G}$) reduces $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ (see Lemma A.14), the sum of the first two terms in the objective function of (1.29) is minimized when the columns of $\mathbf{G}$ are any u eigenvectors of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. Thus, the role of these terms is to enforce the constraint on $\boldsymbol{\Sigma}$ in model (1.20). The third term measures bias using the standardized coefficients from the regression of $\mathbf{G}_0^T\mathbf{Y}$ on $\mathbf{X}$, corresponding to the population constraint that $\mathcal{B} \subseteq$ span($\mathbf{G}$). The third term will equal 0 for any subspace span($\mathbf{G}$) that contains span($\mathbf{B}$), since then $\mathbf{G}_0^T\mathbf{B} = 0$. In short, the objective function balances the two constraints on the envelope model (1.20), minimizing bias while insuring that the solution is “close” to a reducing subspace of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$.

1.5.4 Scaling the Responses

Like principal component regression, ridge regression, partial least squares, and other methods, the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are not invariant or equivariant under rescaling of $\mathbf{Y}$. For instance, rescale $\mathbf{Y}$ by a nonsingular diagonal matrix $\mathbf{D}$ with positive diagonal elements, $\mathbf{Y} \mapsto \mathbf{DY}$. Expressed in terms of the parameters from the regression of $\mathbf{Y}$ on $\mathbf{X}$, the envelope for the regression of $\mathbf{DY}$ on $\mathbf{X}$ is $\mathcal{E}_{\mathbf{D\Sigma D}}(\mathbf{D}\mathcal{B})$. Then generally $\mathcal{E}_{\mathbf{D\Sigma D}}(\mathbf{D}\mathcal{B}) \neq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ with possibly different dimensions $\dim(\mathcal{E}_{\mathbf{D\Sigma D}}(\mathbf{D}\mathcal{B})) \neq \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$. Let $\widehat{\boldsymbol{\beta}}_{\mathbf{D}}$ be the envelope estimator of $\boldsymbol{\beta}$ from the regression of $\mathbf{DY}$ on $\mathbf{X}$. Then generally we do not have $\widehat{\boldsymbol{\beta}}_{\mathbf{D}} = \boldsymbol{\beta}$ or $\widehat{\boldsymbol{\beta}}_{\mathbf{D}} = \mathbf{D}\widehat{\boldsymbol{\beta}}$. For these reasons, envelope methods based on the constructions of this chapter tend to work best when the responses are in the same or similar scales, although this is not required. Nearly all of the examples in this book are of that type. Similar comments hold for the predictor envelopes discussed in Chapter 4.

In Chapter 8, we extend the envelope model to allow for simultaneous estimation of a rescaling matrix like $\mathbf{D}$. Until then we stay close to the basic envelope construction described in this chapter.

1.6 Asymptotic Distributions

Asymptotic variances of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ can in principle be determined from the inverse of the Fisher information matrix. However, we encounter complications when attempting to apply this general procedure to model (1.20) because $\boldsymbol{\Gamma}$ is not identifiable due to its overparameterization. Results from Shapiro (1986, Proposition 4.1) allow for overparameterization and show how to find the asymptotic distribution of an estimable function of the parameters. In this section, we sketch the process of determining the asymptotic distribution of the estimable functions $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ of the parameters in model (1.20). Additional details are available from Cook et al. (2010).

The general procedure is the same for other envelope models described in later chapters.

We begin by defining the vector $\boldsymbol{\phi}$ of parameters associated with model (1.20) along with the estimable functions $\mathbf{h}(\boldsymbol{\phi})$. We do not include the intercept $\boldsymbol{\alpha}$ because it is not typically of interest and its maximum likelihood estimator is asymptotically independent of the others in the model. Let

$$\boldsymbol{\phi} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\eta}) \\ \mathrm{vec}(\boldsymbol{\Gamma}) \\ \mathrm{vech}(\boldsymbol{\Omega}) \\ \mathrm{vech}(\boldsymbol{\Omega}_0) \end{pmatrix} := \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \boldsymbol{\phi}_3 \\ \boldsymbol{\phi}_4 \end{pmatrix} \tag{1.30}$$

and

$$\mathbf{h}(\boldsymbol{\phi}) = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}\boldsymbol{\eta}) \\ \mathrm{vech}(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) \end{pmatrix} := \begin{pmatrix} \mathbf{h}_1(\boldsymbol{\phi}) \\ \mathbf{h}_2(\boldsymbol{\phi}) \end{pmatrix}.$$

We will also require the gradient matrix,

$$\begin{aligned} \mathbf{H} &:= \begin{pmatrix} \partial \mathbf{h}_1/\partial \boldsymbol{\phi}_1^T & \cdots & \partial \mathbf{h}_1/\partial \boldsymbol{\phi}_4^T \\ \partial \mathbf{h}_2/\partial \boldsymbol{\phi}_1^T & \cdots & \partial \mathbf{h}_2/\partial \boldsymbol{\phi}_4^T \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{22} & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{(r-u)} \end{pmatrix}, \end{aligned}$$

where $\mathbf{H}_{22} = 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_r - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)$, and $\mathbf{C}_r$ and $\mathbf{E}_u$ are the contraction and expansion matrices that connect the vec and vech operators. The derivatives needed for $\mathbf{H}$ are also required in other asymptotic calculations. These and related derivatives are summarized in Section A.6 along with basic properties of the contraction and expansion matrices in Appendix A.5.

Because of the overparameterization in $\boldsymbol{\Gamma}$, $\mathbf{H}$ is not of full rank and standard likelihood methods cannot be applied directly. But $\mathbf{h}$ is estimable, allowing us to use Shapiro (1986, Proposition 4.1) to conclude that

Proposition 1.1 *Under the envelope model (1.20) with normal errors and known* $u = \dim\{\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})\}$, $\sqrt{n}\{\mathbf{h}(\widehat{\boldsymbol{\phi}}) - \mathbf{h}(\boldsymbol{\phi})\}$ *is asymptotically normal with mean* $\mathbf{0}$ *and covariance matrix*

$$\mathrm{avar}\{\sqrt{n}\mathbf{h}(\widehat{\boldsymbol{\phi}})\} = \mathbf{P}_{\mathbf{H}(\mathbf{J})}\mathbf{J}^{-1}\mathbf{P}_{\mathbf{H}(\mathbf{J})}^T = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T, \tag{1.31}$$

where $\mathbf{J}$ *is the information matrix (1.14) for* $(\mathrm{vec}^T(\boldsymbol{\beta}), \mathrm{vech}^T(\boldsymbol{\Sigma}))$ *in the model (1.1). Since* $\mathbf{J}^{-1} - \mathrm{avar}(\sqrt{n}\mathbf{h}(\widehat{\boldsymbol{\phi}})) \geq 0$, *the envelope estimator never does worse than the standard estimator.*

Additionally, $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *is asymptotically normal with mean 0 and variance*

$$\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\widehat{\boldsymbol{\beta}})\} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{U}^{\dagger}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T), \tag{1.32}$$

where

$$\begin{aligned}\mathbf{U} &= \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathrm{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_u \otimes \mathbf{I}_{r-u} \\ &= \boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathrm{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} + (\boldsymbol{\Omega}^{1/2} \otimes \boldsymbol{\Omega}_0^{-1/2} - \boldsymbol{\Omega}^{-1/2} \otimes \boldsymbol{\Omega}_0^{1/2})^2.\end{aligned}$$

These results can be used in practice to construct an asymptotic standard error for $(\widehat{\boldsymbol{\beta}})_{ij}$, $i = 1, \dots, r$, $j = 1, \dots, p$, by first substituting estimates for the unknown quantities on the right-hand side of (1.32) to obtain an estimated asymptotic variance $\widehat{\mathrm{avar}}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}})\}$. The estimated asymptotic variance $\widehat{\mathrm{avar}}\{\sqrt{n}(\widehat{\boldsymbol{\beta}})_{ij}\}$ is then the corresponding diagonal element of $\widehat{\mathrm{avar}}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}})\}$, and its *asymptotic standard error* is

$$\mathrm{se}\{(\widehat{\boldsymbol{\beta}})_{ij}\} = \frac{[\widehat{\mathrm{avar}}\{\sqrt{n}(\widehat{\boldsymbol{\beta}})_{ij}\}]^{1/2}}{\sqrt{n}}, \tag{1.33}$$

with corresponding Z-score equal to $(\widehat{\boldsymbol{\beta}})_{ij}/\mathrm{se}\{(\widehat{\boldsymbol{\beta}})_{ij}\}$. This is the method that is used to obtain standard errors in the previous illustrations. The bootstrap can also be used, as described in Section 1.11.

If $u = r$, then $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$, and the second addend on the right-hand side of (1.32) does not appear. The first addend on the right-hand side of (1.32) is the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$, the envelope estimator of $\boldsymbol{\beta}$ when $\boldsymbol{\Gamma}$ is known (cf. (1.22)),

$$\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\} = \boldsymbol{\Sigma}_{\mathrm{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T.$$

The second addend can be interpreted as the "cost" of estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. The total on the right does not exceed $\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\mathbf{B})\}$; that is,

$$\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\mathbf{B})\} - \mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}})\} \geq 0.$$

The asymptotic variance (1.32) can be reexpressed informatively as

$$\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}})\} = \mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\} + \mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})\}. \tag{1.34}$$

As mentioned previously, the first addend $\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\}$ on the right-hand side is the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$, and in the second addend $\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})\}$ is the asymptotic variance of the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}}$ of $\boldsymbol{\beta}$ when $\boldsymbol{\eta}$ is known, both corresponding to asymptotic variances in multivariate linear models.

We next consider a special case to gain intuition about the gains offered by envelopes over the standard method. In preparation, write the asymptotic variance of the estimator $\mathbf{B}$ of $\boldsymbol{\beta}$ under the standard model in terms of the envelope parameters as

$$\mathrm{avar}\{\sqrt{n}\,\mathrm{vec}(\mathbf{B})\} = \boldsymbol{\Sigma}_{\mathrm{X}}^{-1} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{X}}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma}_{\mathrm{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.$$

Let $\mathbf{D} = \text{avar}\{\sqrt{n}\,\text{vec}(\mathbf{B})\} - \text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}})\}$ denote the difference in asymptotic variances. We then have

$$\mathbf{D} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T - (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{U}^{\dagger}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T) \geq 0.$$

Suppose now that $\boldsymbol{\beta}$ has full column rank p, $\boldsymbol{\Omega} = \omega\mathbf{I}_u$ and $\boldsymbol{\Omega}_0 = \omega_0\mathbf{I}_{r-u}$. Compound symmetry is an instance of this structure (see Section 1.3.2). As a consequence, we see that $\boldsymbol{\Sigma}$ has two eigenspaces, one corresponding to ω and the other to ω_0. Since $\mathcal{B}$ is contained in the eigenspace corresponding to ω, we must have $\mathcal{B} = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, $u = p$, $\boldsymbol{\Gamma} = \boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1/2}$, and $\boldsymbol{\eta} = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2}$. Then after a little algebra, we can write

$$\mathbf{D} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 - \boldsymbol{\eta}^T\{\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\omega_0^{-1} + (\omega\omega_0^{-1} + \omega^{-1}\omega_0 - 2)\mathbf{I}_u\}^{-1}\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T.$$

Simplifying $\omega\omega_0^{-1} + \omega^{-1}\omega_0 - 2 = (\omega - \omega_0)^2/\omega\omega_0$ and using the fact that $\boldsymbol{\eta} = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2} \in \mathbb{R}^{p\times p}$ is nonsingular,

$$\mathbf{D} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 - \{\boldsymbol{\Sigma}_{\mathbf{X}} + \omega^{-1}(\omega - \omega_0)^2(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 \geq 0.$$

Recall that if $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is known, the envelope model will offer substantial gains when $\omega \ll \omega_0$. The second addend on the right-hand side of the last expression shows the cost of estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. From the expression, we see that the cost will be relatively small when again $\omega \ll \omega_0$. It can also be small if $\omega \neq \omega_0$ and $(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}$ is large relative to $\boldsymbol{\Sigma}_{\mathbf{X}}$.

The next corollary summarizes an important special case of this illustration.

Corollary 1.1 *Assume that* $\boldsymbol{\Sigma} = \omega\mathbf{I}_r$ *and that* $\text{rank}(\boldsymbol{\beta}) = p$. *Then the asymptotic variance of the envelope estimator is the same as the asymptotic variance of the usual maximum likelihood estimator,* $\text{avar}\{\sqrt{n}\,\text{vec}(\mathbf{B})\} = \text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}})\}$.

Gains are still possible when $\boldsymbol{\Sigma} = \omega\mathbf{I}_r$ if the rank of $\boldsymbol{\beta}$ is less than p. In that case, $\mathcal{B} = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, $\omega\omega_0^{-1} + \omega^{-1}\omega_0 - 2 = 0$, and

$$\begin{aligned}\mathbf{D} &= \{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \boldsymbol{\eta}^T(\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T)^{-1}\boldsymbol{\eta}\} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 \\ &= \mathbf{Q}_{\boldsymbol{\eta}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T\omega_0 \geq 0.\end{aligned}$$

1.7 Fitted Values and Predictions

The previous asymptotic results can be used to derive the asymptotic distribution of the fitted values, as well as the asymptotic prediction variance. The fitted values at a particular $\mathbf{X}$ can be written as $\widehat{\mathbf{Y}} = \bar{\mathbf{Y}} + \widehat{\boldsymbol{\beta}}\mathbf{X} = \bar{\mathbf{Y}} + (\mathbf{X}^T \otimes \mathbf{I}_r)\text{vec}(\widehat{\boldsymbol{\beta}})$. Hence, the fitted value $\widehat{\mathbf{Y}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mathbf{Y}} - \text{E}(\widehat{\mathbf{Y}})) \xrightarrow{\mathcal{L}} N_r[0, \text{avar}\{\sqrt{n}\,\text{vec}(\bar{\mathbf{Y}} + \widehat{\boldsymbol{\beta}}\mathbf{X})\}]. \tag{1.35}$$

Using (1.34) and the fact that $\bar{\mathbf{Y}}$ and $\widehat{\boldsymbol{\beta}}$ are asymptotically independent, the asymptotic variance in this distribution can be expressed informatively as

$$\begin{aligned}
&\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\bar{\mathbf{Y}} + \widehat{\boldsymbol{\beta}}\mathbf{X})\} \\
&\quad = \boldsymbol{\Sigma} + (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\widehat{\boldsymbol{\beta}})\}(\mathbf{X} \otimes \mathbf{I}_r)) \\
&\quad = \boldsymbol{\Sigma} + (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})\}(\mathbf{X} \otimes \mathbf{I}_r) \\
&\qquad + (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\eta})\}(\mathbf{X} \otimes \mathbf{I}_r) \\
&\quad = \boldsymbol{\Sigma} + \mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}\mathbf{X})\} + \mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\eta}\mathbf{X})\}.
\end{aligned}$$

Consequently, the variance of a fitted value has the same essential decomposition as the variance of $\widehat{\boldsymbol{\beta}}$ discussed previously.

Turning to prediction, suppose that at some value of $\mathbf{X}$, we wish to infer about a new $\mathbf{Y}$, say $\mathbf{Y}_{\mathrm{new}}$, independently of the past observations. Then

$$\begin{aligned}
&\mathrm{E}\{(\widehat{\mathbf{Y}} - \mathbf{Y}_{\mathrm{new}})(\widehat{\mathbf{Y}} - \mathbf{Y}_{\mathrm{new}})^T\} \\
&\quad = \mathrm{E}\{(\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}}))(\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}}))^T\} + \mathrm{E}\{(\mathrm{E}(\widehat{\mathbf{Y}}) - \mathbf{Y}_{\mathrm{new}})(\mathrm{E}(\widehat{\mathbf{Y}}) - \mathbf{Y}_{\mathrm{new}})^T\},
\end{aligned}$$

where the cross-product terms vanish because $\mathbf{Y}_{\mathrm{new}}$ and $\widehat{\mathbf{Y}}$ are independent. Combining this with expression (1.35), we see that the mean squared error of the prediction is approximated by

$$\begin{aligned}
\mathrm{E}\{(\widehat{\mathbf{Y}} - \mathbf{Y}_{\mathrm{new}})(\widehat{\mathbf{Y}} - \mathbf{Y}_{\mathrm{new}})^T\} &= n^{-1}\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}\mathbf{X})\} \\
&\quad + (1 + n^{-1})\boldsymbol{\Sigma} + o(n^{-1}). \qquad (1.36)
\end{aligned}$$

Envelope model (1.20) can be quite effective at reducing $\mathrm{avar}\{\sqrt{n}\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}\mathbf{X})\}$, but it has no impact on the underlying variance $\boldsymbol{\Sigma}$ except for the induced structure. Envelopes give greatest estimative gain when $\mathrm{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}) = \boldsymbol{\Omega}_0$ is large relative to the material variation $\mathrm{var}(\boldsymbol{\Gamma}^T\mathbf{Y} \mid \mathbf{X}) = \boldsymbol{\Omega}$. Nevertheless, the X-invariant variation is still present in $\boldsymbol{\Sigma}$, and consequently the advantages that model (1.20) bring in the estimation of $\boldsymbol{\beta}$ may not be present to the same degree in prediction. This can be seen in the schematic illustration of Figure 1.3, where the distributions represented in the left-hand display contribute to prediction, but not to estimation as shown in the right-hand display. Greater predictive gain might be realized by using partial envelopes for prediction, as discussed in Section 3.4, or by using envelopes for predictor reduction, as discussed in Chapter 4.

1.8 Testing the Responses

1.8.1 Test Development

In some regressions, we may wish to test the hypothesis that the part of $\mathbf{Y}$ represented by the projection $\mathbf{P}_{\mathcal{H}}\mathbf{Y}$ onto a known user-specified subspace $\mathcal{H} \subset \mathbb{R}^r$

holds the entire X-variant part of $\mathbf{Y}$. That is, starting with model (1.20), we may wish to test the hypothesis that

$$\text{(i) } \mathbf{Q}_{\mathcal{H}}\mathbf{Y} \mid (\mathbf{X} = \mathbf{x}_1) \sim \mathbf{Q}_{\mathcal{H}}\mathbf{Y} \mid (\mathbf{X} = \mathbf{x}_2) \text{ and (ii) } \mathbf{P}_{\mathcal{H}}\mathbf{Y} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{H}}\mathbf{Y} \mid \mathbf{X}. \quad (1.37)$$

This is similar to specification (1.18), except here we are not requiring $\mathcal{H}$ to be the smallest subspace. Equivalently, this hypothesis specifies that $\mathbf{Q}_{\mathcal{H}}\mathbf{Y}$ is X-invariant, while allowing for the possibility that $\mathbf{P}_{\mathcal{H}}\mathbf{Y}$ may also contain an X-invariant part of $\mathbf{Y}$. Since $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is defined as the intersection of all subspaces that satisfy (1.37), it follows that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is contained in any subspace that satisfies (1.37) and thus $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) \subseteq \mathcal{H}$. Hypothesis (1.37) can be tested by using the likelihood ratio test statistic $\Lambda_u(\mathcal{H}) = 2(\widehat{L}_u - \widehat{L}_u(\mathcal{H}))$, where $\widehat{L}_u$ is the maximized envelope log-likelihood (1.27), and $\widehat{L}_u(\mathcal{H})$ is the maximized log-likelihood under the hypothesis. As in previous sections, we treat u as known in this development. Methods for selecting u are discussed in Section 1.10.

To construct $\widehat{L}_u(\mathcal{H})$, let $v = \dim(\mathcal{H}) \geq u$ denote the known dimension of $\mathcal{H}$, let $\mathbf{H} \in \mathbb{R}^{r\times v}$ be a semi-orthogonal basis matrix for $\mathcal{H}$, and let $(\mathbf{H}, \mathbf{H}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix. Under hypothesis (1.37), $\mathcal{H}$ is a reducing subspace of $\mathbf{\Sigma}$ that contains $\mathcal{B}$. It follows from Proposition A.6 that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathbf{H}\mathcal{E}_{\mathbf{H}^T\mathbf{\Sigma}\mathbf{H}}(\mathbf{H}^T\mathcal{B})$. Let $\mathbf{h} \in \mathbb{R}^{v\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{H}^T\mathbf{\Sigma}\mathbf{H}}(\mathbf{H}^T\mathcal{B})$, which corresponds to the envelope regression of $\mathbf{H}^T\mathbf{Y}$ on $\mathbf{X}$. Then we have that $\mathbf{\Gamma} = \mathbf{H}\mathbf{h}$ is a basis for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$. Let $\mathbf{\Gamma}_{01} = \mathbf{H}\mathbf{h}_0$ be a semi-orthogonal basis for the orthogonal complement of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ within $\mathcal{H}$, where $\mathbf{h}_0 \in \mathbb{R}^{v\times(v-u)}$ is semi-orthogonal and $\mathbf{h}^T\mathbf{h}_0 = 0$. Envelope model (1.20) can now be expressed under (1.37) as

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \quad (1.38)$$

$$\mathbf{\Sigma} = \mathbf{H}\mathbf{h}\mathbf{\Omega}\mathbf{h}^T\mathbf{H}^T + \mathbf{H}\mathbf{h}_0\mathbf{\Omega}_{01}\mathbf{h}_0^T\mathbf{H}^T + \mathbf{H}_0\mathbf{\Omega}_{00}\mathbf{H}_0^T, \quad (1.39)$$

where the structure of $\mathbf{\Sigma}$ follows because $\mathbf{h}^T\mathbf{H}^T\mathbf{Y}$, $\mathbf{h}_0^T\mathbf{H}^T\mathbf{Y}$, and $\mathbf{H}_0^T\mathbf{Y}$ are mutually independent given $\mathbf{X}$. The terms $\boldsymbol{\alpha}$, $\mathbf{\Omega}$, and $\boldsymbol{\eta}$ that occur in envelope model (1.20) are the same as those under hypothesis (1.37), while $\boldsymbol{\beta} = \mathbf{H}\mathbf{h}\boldsymbol{\eta}$, $\mathbf{\Gamma}_0 = (\mathbf{H}\mathbf{h}_0, \mathbf{H}_0)$, and $\mathbf{\Omega}_0 = \text{bdiag}(\mathbf{\Omega}_{01}, \mathbf{\Omega}_{00})$. We see from (1.38) and (1.39) that $\mathbf{H}^T\mathbf{Y} \mid \mathbf{X}$ follows an envelope model, which is helpful when studying the likelihood.

The log-likelihood

$$L_u(\mathcal{H}) := L_u(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{h}, \mathbf{\Omega}, \mathbf{\Omega}_{01}, \mathbf{\Omega}_{00} \mid \mathcal{H})$$

for this model can now be expressed as

$$L_u(\mathcal{H}) = -(nr/2)\log(2\pi) - (n/2)\log|\mathbf{\Sigma}| \\ -(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)^T\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i),$$

where $\mathbf{\Sigma}$ is as given in (1.39). Continuing,

$$\begin{aligned}L_u(\mathcal{H}) = &-(nr/2)\log(2\pi) - (n/2)\log|\mathbf{\Omega}|\\ &-(n/2)\log|\mathbf{\Omega}_{01}| - (n/2)\log|\mathbf{\Omega}_{00}|\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)^T(\mathbf{H}\mathbf{h}\mathbf{\Omega}^{-1}\mathbf{h}^T\mathbf{H}^T)(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)^T(\mathbf{H}\mathbf{h}_0\mathbf{\Omega}_{01}^{-1}\mathbf{h}_0^T\mathbf{H}^T)(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)^T(\mathbf{H}_0\mathbf{\Omega}_{00}^{-1}\mathbf{H}_0^T)(\mathbf{Y}_i - \boldsymbol{\alpha} - \mathbf{H}\mathbf{h}\boldsymbol{\eta}\mathbf{X}_i).\end{aligned}$$

Rearranging terms to match regressions of $\mathbf{H}^T\mathbf{Y}$ and $\mathbf{H}_0^T\mathbf{Y}$ on $\mathbf{X}$, we get

$$\begin{aligned}L_u(\mathcal{H}) = &-(nr/2)\log(2\pi) - (n/2)\log|\mathbf{\Omega}| - (n/2)\log|\mathbf{\Omega}_{01}|\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{H}^T(\mathbf{Y}_i - \boldsymbol{\alpha}) - \mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)^T\mathbf{h}\mathbf{\Omega}^{-1}\mathbf{h}^T(\mathbf{H}(\mathbf{Y}_i - \boldsymbol{\alpha}) - \mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{H}^T(\mathbf{Y}_i - \boldsymbol{\alpha}))^T\mathbf{h}_0\mathbf{\Omega}_{01}^{-1}\mathbf{h}_0^T(\mathbf{H}^T(\mathbf{Y}_i - \boldsymbol{\alpha}))\\ &-(n/2)\log|\mathbf{\Omega}_{00}| - (1/2)\sum_{i=1}^{n}(\mathbf{H}_0^T(\mathbf{Y}_i - \boldsymbol{\alpha}))^T\mathbf{\Omega}_{00}^{-1}(\mathbf{H}_0^T(\mathbf{Y}_i - \boldsymbol{\alpha})).\end{aligned}$$

Addends 2–5 plus $-(nv/2)\log(2\pi)$ correspond to an envelope likelihood with v responses, as described in Section 1.5.1, and the last two addends plus $-(n(r-v)/2)\log(2\pi)$ correspond to a mean only regression. The sum $-(nv/2)\log(2\pi) - (n(r-v)/2)\log(2\pi) = -(nr/2)\log(2\pi)$ is the first addend.

Let $L_u^{(1)}(\mathbf{h}, \boldsymbol{\eta}, \mathbf{\Omega}, \mathbf{\Omega}_{01} \mid \mathcal{H})$ denote the log-likelihood for the envelope regression of $\mathbf{H}^T\mathbf{Y}$ on $\mathbf{X}$:

$$\begin{aligned}L_u^{(1)} = &-(nv/2)\log(2\pi) - (n/2)\log|\mathbf{\Omega}| - (n/2)\log|\mathbf{\Omega}_{01}|\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{H}^T(\mathbf{Y}_i - \boldsymbol{\alpha}) - \mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)^T\mathbf{h}\mathbf{\Omega}^{-1}\mathbf{h}^T(\mathbf{H}(\mathbf{Y}_i - \boldsymbol{\alpha}) - \mathbf{h}\boldsymbol{\eta}\mathbf{X}_i)\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{H}^T(\mathbf{Y}_i - \boldsymbol{\alpha}))^T\mathbf{h}_0\mathbf{\Omega}_{01}^{-1}\mathbf{h}_0^T(\mathbf{H}^T(\mathbf{Y}_i - \boldsymbol{\alpha})).\end{aligned}$$

Let $L_u^{(2)}(\mathbf{\Omega}_{00} \mid \mathcal{H})$ denote the log-likelihood arising from the mean-only regression $\mathbf{H}_0^T\mathbf{Y} = \mathbf{H}_0^T\boldsymbol{\alpha} + \mathbf{H}_0^T\boldsymbol{\varepsilon}$:

$$\begin{aligned}L_u^{(2)} = &-(n(r-v)/2)\log(2\pi) - (n/2)\log|\mathbf{\Omega}_{00}|\\ &-(1/2)\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha})^T\mathbf{H}_0\mathbf{\Omega}_{00}^{-1}\mathbf{H}_0^T(\mathbf{Y}_i - \boldsymbol{\alpha}).\end{aligned}$$

Then

$$L_u(\mathcal{H}) = L_u^{(1)}(\mathbf{h}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_{01} \mid \mathcal{H}) + L_u^{(2)}(\boldsymbol{\Omega}_{00} \mid \mathcal{H}).$$

and the estimators of the parameters in (1.38) and (1.39) can be found by following the derivation in Section 1.5.1:

$$\begin{aligned}
\hat{\mathbf{h}} &= \arg\min_{\mathbf{G}}\{\log|\mathbf{G}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}^{-1}\mathbf{G}|\},\\
\hat{\boldsymbol{\eta}} &= \hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1},\\
\hat{\boldsymbol{\beta}} &= \mathbf{H}\hat{\mathbf{h}}\hat{\boldsymbol{\eta}},\\
\hat{\boldsymbol{\Omega}} &= \hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{X}}\hat{\mathbf{h}},\\
\hat{\boldsymbol{\Omega}}_{01} &= \hat{\mathbf{h}}_0^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}\hat{\mathbf{h}}_0,\\
\hat{\boldsymbol{\Omega}}_{00} &= \mathbf{S}_{\mathbf{H}_0^T\mathbf{Y}},\\
\hat{\boldsymbol{\Sigma}} &= \mathbf{H}\hat{\mathbf{h}}\hat{\boldsymbol{\Omega}}\hat{\mathbf{h}}^T\mathbf{H}^T + \mathbf{H}\hat{\mathbf{h}}_0\hat{\boldsymbol{\Omega}}_{01}\hat{\mathbf{h}}_0^T\mathbf{H}^T + \mathbf{H}_0\mathbf{S}_{\mathbf{H}_0^T\mathbf{Y}}\mathbf{H}_0^T,
\end{aligned}$$

where the minimum is over all semi-orthogonal matrices $\mathbf{G} \in \mathbb{R}^{v\times u}$. Let $\hat{L}_u^{(1)}$ and $\hat{L}_u^{(2)}$ denote the maximized values of $L_u^{(1)}$ and $L_u^{(2)}$:

$$\begin{aligned}
\hat{L}_u^{(1)} &= -(nv/2)\log(2\pi) - nv/2 - (n/2)\log|\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}|\\
&\quad -(n/2)\log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{X}}\hat{\mathbf{h}}| - (n/2)\log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}^{-1}\hat{\mathbf{h}}|,\\
\hat{L}_u^{(2)} &= -(n(r-v)/2)\log(2\pi) - n(r-v)/2 - (n/2)\log|\mathbf{S}_{\mathbf{H}_0^T\mathbf{Y}}|,\\
\hat{L}_u(\mathcal{H}) &= \hat{L}_u^{(1)} + \hat{L}_u^{(2)}\\
&= -(nr/2)\log(2\pi) - (nr/2) - (n/2)\log|\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}| - (n/2)\log|\mathbf{S}_{\mathbf{H}_0^T\mathbf{Y}}|\\
&\quad -(n/2)\log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{X}}\hat{\mathbf{h}}| - (n/2)\log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}^{-1}\hat{\mathbf{h}}|.
\end{aligned}$$

Combining these results with (1.27), we obtain

$$\begin{aligned}
\Lambda_u(\mathcal{H}) = n\Big\{&\log|\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}| + \log|\mathbf{S}_{\mathbf{H}_0^T\mathbf{Y}}| - \log|\mathbf{S}_{\mathbf{Y}}| + \log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{X}}\hat{\mathbf{h}}|\\
&+ \log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}}^{-1}\hat{\mathbf{h}}| - \log|\hat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\hat{\boldsymbol{\Gamma}}| - \log|\hat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}^{-1}\hat{\boldsymbol{\Gamma}}|\Big\}.
\end{aligned}\tag{1.40}$$

Under the hypothesis, $\Lambda_u(\mathcal{H})$ is distributed asymptotically as a chi-square random variable with $v(r-v)$ degrees of freedom.

1.8.2 Testing Individual Responses

In this section, we discuss tests to determine if the distribution of a selected subset of the responses, $\mathbf{Y}_1 \in \mathbb{R}^s, s < r$, is unaffected by changing the predictors. Without loss of generality, assume that those responses are the first s components in $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$, so we can write a partitioned form of model (1.1) as

$$\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{pmatrix} + \begin{pmatrix}\boldsymbol{\beta}_1\mathbf{X}\\ \boldsymbol{\beta}_2\mathbf{X}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{pmatrix}, \tag{1.41}$$

where $\mathbf{Y}_2 \in \mathbb{R}^{r-s}$, and the partitioning of $\boldsymbol{\beta}$ corresponds to the partitioning of $\mathbf{Y}$, with $\boldsymbol{\beta}_1 \in \mathbb{R}^{s\times p}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{(r-s)\times p}$. The specific requirements on $\mathbf{Y}_1$ under this hypothesis are

$$\text{(i) } \mathbf{Y}_1 \mid (\mathbf{X} = \mathbf{x}_1) \sim \mathbf{Y}_1 \mid (\mathbf{X} = \mathbf{x}_2) \ \text{ and } \ \text{(ii) } \mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2 \mid \mathbf{X}. \tag{1.42}$$

This hypothesis corresponds to (1.37) with $\mathbf{H} = (0, \mathbf{I}_{r-s})^T \in \mathbb{R}^{r\times(r-s)}$, $v = r - s$, and $\mathbf{H}_0 = (\mathbf{I}_s, 0)^T \in \mathbb{R}^{r\times s}$. Assuming that (1.42) holds, the responses in $\mathbf{Y}_1$ are X-invariant. Generally, the response subvector $\mathbf{Y}_2$ may hold X-invariant responses as well because they are not determined solely by the value of the corresponding regression coefficients $\boldsymbol{\beta}_1$. It follows from (1.42(i)) that $\boldsymbol{\beta}_1 = 0$ is a necessary condition for $\mathbf{Y}_1$ to be X-invariant, but it is not sufficient since we also require (1.42(ii)) to hold so information from $\mathbf{Y}_1$ does not contribute to the estimation of $\boldsymbol{\beta}_2$ via an association with $\mathbf{Y}_2$. Additionally, it follows from the discussion in Section 1.8.1 that under (1.42) semi-orthogonal bases $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and $\mathcal{E}_{\boldsymbol{\Sigma}}^{\perp}(\mathcal{B})$ can be constructed as

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0 \\ \boldsymbol{\Gamma}_2 \end{pmatrix}, \quad \boldsymbol{\Gamma}_0 = \begin{pmatrix} \mathbf{I}_s & 0 \\ 0 & \boldsymbol{\Gamma}_{2,0} \end{pmatrix},$$

where $\boldsymbol{\Gamma}_2 \in \mathbb{R}^{(r-s)\times u}$, $\boldsymbol{\Gamma}_{2,0} \in \mathbb{R}^{(r-s)\times(r-s-u)}$, and still $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$. We will return to this line of reasoning when considering sparse envelopes for response selection in Section 7.5. The estimator of $\mathbf{h}$ and the likelihood ratio statistic simplify a bit to

$$\begin{aligned} \hat{\mathbf{h}} &= \arg\min_{\mathbf{h}} \{\log|\mathbf{h}^T\mathbf{S}_{\mathbf{Y}_2|\mathbf{X}}\mathbf{h}| + \log|\mathbf{h}^T\mathbf{S}_{\mathbf{Y}_2}^{-1}\mathbf{h}|\}, \\ \Lambda_u(\mathcal{H}) &= n\{\log|\mathbf{S}_{\mathbf{Y}_2}| + \log|\mathbf{S}_{\mathbf{Y}_1}| - \log|\mathbf{S}_{\mathbf{Y}}| + \log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{Y}_2|\mathbf{X}}\hat{\mathbf{h}}| \\ &\quad + \log|\hat{\mathbf{h}}^T\mathbf{S}_{\mathbf{Y}_2}^{-1}\hat{\mathbf{h}}| - \log|\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\boldsymbol{\Gamma}}| - \log|\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}^{-1}\widehat{\boldsymbol{\Gamma}}|\}. \end{aligned}$$

A brief illustration on the use of this methodology is given in Section 2.8.

We conclude this section by giving additional discussion to emphasize an important difference between inference on responses and coefficients, using results from Su et al. (2016). Suppose that an oracle told us that in fact $\boldsymbol{\beta}_1 = 0$ in (1.41). Should we now estimate $\boldsymbol{\beta}_2$ from the constrained model

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{\beta}_2\mathbf{X} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix} \tag{1.43}$$

or the reduced model

$$\mathbf{Y}_2 = \boldsymbol{\alpha}_2 + \boldsymbol{\beta}_2\mathbf{X} + \boldsymbol{\varepsilon}_2?$$

We address this question by comparing $\mathbf{B}_{2,\mathrm{C}}$, the maximum likelihood estimator of $\boldsymbol{\beta}_2$ from the constrained model, and $\mathbf{B}_{2,\mathrm{R}}$, the maximum likelihood estimator of $\boldsymbol{\beta}_2$ from the reduced model. We need some additional notation for this comparison. Let $\mathbf{B}_1$ and $\mathbf{B}_2$ denote the maximum likelihood estimators from the separate regressions of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ on $\mathbf{X}$. Let $\mathbf{r}_1$ and $\mathbf{r}_2$ denote the residual vectors from the separate regressions of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ on $\mathbf{X}$, and let $\mathbf{B}_{2|1}$ denote the coefficient matrix from the ordinary least squares regression of $\mathbf{r}_2$ on $\mathbf{r}_1$, which is the same as the estimate of $\boldsymbol{\beta}_2$ from the full model (1.41). Let $\boldsymbol{\Sigma}_{jj} = \mathrm{var}(\mathbf{Y}_j \mid \mathbf{X})$, let $\boldsymbol{\Sigma}_{ij} = \mathrm{cov}(\mathbf{Y}_i, \mathbf{Y}_j \mid \mathbf{X})$, $i, j = 1, 2$, and let $\boldsymbol{\Sigma}_{2|1} = \mathrm{var}(\mathbf{Y}_2 \mid \mathbf{Y}_1, \mathbf{X}) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Then (Su et al. 2016),

$$\mathbf{B}_{2,\mathrm{C}} = \mathbf{B}_2 - \mathbf{B}_{2|1}\mathbf{B}_1 \quad \text{with} \quad \mathrm{avar}(\sqrt{n}\mathbf{B}_{2,\mathrm{C}}) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{2|1},$$
$$\mathbf{B}_{2,\mathrm{R}} = \mathbf{B}_2 \quad \text{with} \quad \mathrm{avar}(\sqrt{n}\mathbf{B}_{2,\mathrm{R}}) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{22}.$$

From this we see that $\mathbf{B}_{2,\mathrm{C}} \neq \mathbf{B}_{2,\mathrm{R}}$, with $\mathbf{B}_{2,\mathrm{C}}$ being an adjusted version of $\mathbf{B}_{2,\mathrm{R}}$ that accounts for the conditional correlation between $\mathbf{Y}_1$ and $\mathbf{Y}_2$. More importantly,

$$\mathrm{avar}(\sqrt{n}\mathbf{B}_{2,\mathrm{R}}) - \mathrm{avar}(\sqrt{n}\mathbf{B}_{2,\mathrm{C}}) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \geq 0,$$

so the estimator $\mathbf{B}_{2,\mathrm{C}}$ from the constrained model is always at least as good as the estimator $\mathbf{B}_{2,\mathrm{R}}$ from the reduced model and may be substantially better depending on $\boldsymbol{\Sigma}_{12}$, the two estimators being asymptotically equivalent when $\boldsymbol{\Sigma}_{12} = 0$.

1.8.3 Testing Containment Only

In some regressions, we may wish to test the hypothesis $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) \subseteq \mathcal{H}$ of containment only. This hypothesis satisfies (1.37(i)), but not necessarily (1.37(ii)). Under this hypothesis, we still can represent the basis $\boldsymbol{\Gamma}$ for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ as $\boldsymbol{\Gamma} = \mathbf{H}\mathbf{h}$, but we no longer necessarily have $\mathbf{H}^T\mathbf{Y} \perp\!\!\!\perp \mathbf{H}_0^T\mathbf{Y} \mid \mathbf{X}$. As a consequence, the estimator of $\mathbf{h}$ is determined as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{G}}\{\log|\mathbf{G}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{H}^T\mathbf{Y}|\mathbf{H}_0\mathbf{Y}}^{-1}\mathbf{G}|\},$$

where the second term of the objective function is now the residual covariance matrix from the regression of $\mathbf{H}^T\mathbf{Y}$ on $\mathbf{H}_0^T\mathbf{Y}$, rather than the marginal covariance matrix of $\mathbf{H}^T\mathbf{Y}$. The remaining parameters and the likelihood ratio statistic can be determined following the steps in the previous development.

1.9 Nonnormal Errors

Again consider model (1.1), but now relax the condition that the errors are normally distributed. The structure of an envelope described in Definition 1.2 requires only $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$; it does not require normality. This implies that the coordinate form of the envelope model (1.20) is still applicable with nonnormal errors, although now the condition $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$ is replaced with $\mathrm{cov}(\boldsymbol{\Gamma}^T\mathbf{Y}, \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}) = 0$. Nevertheless, the goal under model (1.20) remains the estimation of $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$.

Lacking knowledge of the distribution of the errors, we need to decide how to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. One natural route is to base estimation on the least squares estimators $\mathbf{B}$ and $\mathbf{S}_{\mathrm{Y|X}}$ by selecting an objective function to fit the mean and variance structures of model (1.20). There are likely many ways to proceed, but one good way is to use the partially maximized log-likelihood $L_2(\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}))$ (1.24) to fill this role for the purpose of estimating the envelope. It can be used straightforwardly since it is a function of only $\mathbf{B}$ and $\mathbf{S}_{\mathrm{Y|X}}$. The remaining parameters are then estimated as described in Section 1.5. Since we are not assuming normality, these estimators no longer inherit optimality properties from general likelihood theory, so a different approach is needed to study them.

Lemma 1.1 *The sample matrices* $\mathbf{B}$, $\mathbf{S}_{\mathrm{Y|X}}$ *and* $\mathbf{S}_{\mathrm{Y}}$ *are* $\sqrt{n}$ *consistent estimators of their population counterparts* $\boldsymbol{\beta}$, $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ *and* $\boldsymbol{\Sigma}_{\mathrm{Y}} = \boldsymbol{\Sigma} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathrm{X}}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T$.

Recall from (1.25) that $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathrm{span}\{\arg\min_{\mathbf{G}}(\log|\mathbf{G}^T\mathbf{S}_{\mathrm{Y|X}}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{S}_{\mathrm{Y}}\mathbf{G}_0|)\}$. It follows from this lemma that $\log|\mathbf{G}^T\mathbf{S}_{\mathrm{Y|X}}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{S}_{\mathrm{Y}}\mathbf{G}_0|$ converges in probability to $\log|\mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T(\boldsymbol{\Sigma} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathrm{X}}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T)\mathbf{G}_0|$. This population objective function is covered by Proposition 6.1, and consequently it follows that

$$\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathrm{span}\{\arg\min_{\mathbf{G}}(\log|\mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathrm{Y}}\mathbf{G}_0|)\},$$

and thus that the normal-theory objective function recovers $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ in the population without actually assuming normality.

Going further, assume that the errors have finite fourth moments and that as $n \to \infty$ the maximum diagonal element of $\mathbf{P}_{\mathbb{X}}$ converges to 0. Under these conditions,

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\mathbf{S}_{\mathrm{Y|X}}) - \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}$$

converges to a normal random vector with mean 0 and nonsingular covariance matrix (Su and Cook 2012). It then follows from Shapiro (1986) that with known u

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\widehat{\boldsymbol{\Sigma}}) - \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}$$

also converges to a normal random vector with mean 0 and nonsingular covariance matrix. Consequently, using the normal likelihood for estimation under nonnormality still produces asymptotically normal $\sqrt{n}$-consistent estimators.

Efficiency gains, as illustrated in Figure 1.3, can still accrue without normality, but now they are judged relative to the least squares estimators $\mathbf{B}$ and $\mathbf{S}_{\mathrm{Y|X}}$ rather than maximum likelihood estimators. However, the normal-theory asymptotic variances given in Section 1.6 are no longer applicable. While expressions for the asymptotic variances can be derived, it will likely be difficult to use them as the basis for estimated variances in practice. The residual bootstrap (see Section 1.11 and Freedman 1981) offers a practically useful alternative.

1.10 Selecting the Envelope Dimension, *u*

The dimension u is in effect a model-selection parameter, rather like the rank of $\boldsymbol{\beta}$ in reduced-rank regression (see Section 9.2.1), the power when transforming the response in linear regression, the degree in polynomial regression, or the number of components in a mixture regression. Our discussion so far has treated u as known, but this will typically not be so in applications.

1.10.1 Selection Methods

In this section, we discuss selecting u by using sequential likelihood ratio testing, an information criterion such as AIC or BIC, or cross-validation.

1.10.1.1 Likelihood Ratio Testing

As mentioned in Section 1.4, two envelope models with different values for u are not necessarily nested, but an envelope model is always nested within the standard model (1.1), which arises when $u = r$. The likelihood ratio for testing an envelope model against the standard model can be cast as a test of the hypothesis $u = u_0$ versus the alternative $u = r$. The likelihood ratio statistic for this hypothesis is $\Lambda(u_0) = 2(\widehat{L}_r - \widehat{L}_{u_0})$, where $\widehat{L}_{u_0}$ is the maximized envelope log-likelihood given in (1.27), and $\widehat{L}_r$ is the maximized log-likelihood under the standard model, $\widehat{L}_r = -(nr/2)\log(2\pi) - nr/2 - (n/2)\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|$, giving

$$\Lambda(u_0) = n\log|\mathbf{S}_{\mathbf{Y}}| + n\log|\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\widehat{\boldsymbol{\Gamma}}| + n\log|\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{Y}}^{-1}\widehat{\boldsymbol{\Gamma}}| - n\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}|. \tag{1.44}$$

Under the null hypothesis, this statistic is distributed asymptotically as a chi-squared random variable with $p(r - u_0)$ degrees of freedom. Schott (2013) obtained improved approximations for the asymptotic null distribution of $\Lambda(u_0)$ by using a saddlepoint expansion and he demonstrated that his method can outperform the chi-squared approximation in some settings. He also demonstrated by simulation that the chi-squared approximation can produce very large significance levels when the sample size is small. Due to its simplicity, we will use the chi-squared approximation in illustrations.

The likelihood ratio test statistic $\Lambda(u_0)$ can be used sequentially to estimate u: Starting with $u_0 = 0$, test the hypothesis $u = u_0$ against $u = r$ at a selected level α. If the hypothesis is rejected, increment u_0 by 1 and test again.

The estimate $\widehat{u}$ of u is the first hypothesized value that is not rejected. We indicate this estimator using the notation LRT(α). When testing $u = 0$ versus $u = r$, the envelope basis $\boldsymbol{\Gamma} = 0$ under the null hypothesis, the two terms in $\Lambda(0)$ involving $\widehat{\boldsymbol{\Gamma}}$ do not appear in (1.44) and the test statistic reduces to

$$\Lambda(0) = n \log \frac{|\mathbf{S}_{\mathrm{Y}}|}{|\mathbf{S}_{\mathrm{Y}|\mathrm{X}}|},$$

which is asymptotically distributed under the null hypothesis as a chi-square random variable with pr degrees of freedom. This is the same as the usual likelihood ratio statistic (1.13) for testing $\boldsymbol{\beta} = 0$ in model (1.1).

1.10.1.2 Information Criteria

The envelope dimension can also be selected by using an information criterion:

$$\widehat{u} = \arg\min_{u}\{-2\widehat{L}_u + h(n)N_u\}, \tag{1.45}$$

where N_u is the number of envelope parameters given in (1.21) and $h(n) = \log n$ for BIC and $h(n) = 2$ for AIC. Theoretical results (Su and Cook 2013, Proposition 4) supported by simulations indicate that AIC will tend to select a model that contains the true model and thus will tend to overestimate u. BIC will select the correct u with probability tending to 1 as $n \to \infty$ (Yang 2005), but it can be slow to respond in small samples. LRT(α) can perform well depending on the sample size, but asymptotically it makes an error with rate α. What constitutes a small or large sample in any particular application depends on other characteristics of the regression, including the strength of the signal. It may be useful to use all three methods in applications, giving a preference to BIC and LRT if there is disagreement, or using the largest estimate of u in cases where it is desirable to be conservative.

1.10.1.3 Cross-validation

m-Fold cross-validation is used to select the dimension of the envelope based on prediction performance. For each u, the data are randomly partitioned into m parts of approximately equal size and each part is used in turn for testing prediction performance while the remaining $m - 1$ parts are used for fitting. The dimension selected is the one that minimizes the average prediction errors. A positive definite inner product matrix $\mathbf{M}$ is necessary to map the fitted response vectors to $\mathbb{R}^1$, so the cross-validation criterion is

$$\mathrm{CV}(u) = n^{-1} \sum_{j=1}^{m} \sum_{k=1}^{n_j} (\mathbf{Y}_{jk} - \widehat{\mathbf{Y}}_{jk}^{(u)})^T \mathbf{M} (\mathbf{Y}_{jk} - \widehat{\mathbf{Y}}_{jk}^{(u)}),$$

where j indexes the part, k indexes observations within part j, and $\widehat{\mathbf{Y}}_{jk}^{(u)}$ indicates the fitted vector for a selected u. Then $\widehat{u} = \arg\min_u \mathrm{CV}(u)$. For best results, this procedure should be repeated for several random partitions of the data and the conclusions based on the overall average. Cross-validation will tend to balance variance and bias in its selection of u and so may naturally lead to choices that are different from those indicated by LRT(α) or an information criterion. We use the identity inner product $\mathbf{M} = \mathbf{I}_r$ in illustrations, unless indicated otherwise.

1.10.2 Inferring About rank($\boldsymbol{\beta}$)

The dimension u of the envelope cannot be less than the rank, $k = \text{rank}(\boldsymbol{\beta}) \leq \min(r, p)$, of the population coefficient matrix. For this reason, it can be useful to have some knowledge of k before using the methods of Section 1.10.1 to select u. Bura and Cook (2003) developed a method for estimating k based on a series of chi-squared tests, similar to the LRT method for determining u as discussed in Section 1.10.1.

Let $\hat{\varphi}_1 \geq \hat{\varphi}_2 \geq \cdots \geq \hat{\varphi}_{\min(p,r)}$ denote the singular values of the standardized coefficient matrix

$$\mathbf{B}_{\text{std}} = ((n-p-1)/n)\tilde{\mathbf{B}} = ((n-p-1)/n)\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\mathbf{B}\mathbf{S}_{\mathbf{X}}^{1/2},$$

where $\tilde{\mathbf{B}}$ is as defined previously in (1.6). The statistic for the hypothesis $k = k_0$ is $\Lambda(k_0) = n\sum_{i=k_0+1}^{\min(p,r)} \hat{\varphi}_i^2$. Under the hypothesis $k = k_0$, Bura and Cook showed that $\Lambda(k_0)$ is distributed asymptotically as a chi-squared random variable with $(p-k_0)(r-k_0)$ degrees of freedom. This conclusion is based essentially on only the requirement that $\mathbf{B}$ be asymptotically normal. The statistic $\Lambda(k_0)$ can be used in a sequential manner to provide an estimator of k: beginning with $k_0 = 0$, test $k = k_0$ at a preselected level α. If the test is rejected, increment k_0 by 1 and test again, terminating the first time the hypothesis is not rejected, in which case the current value of k_0 is taken as the estimator of k.

1.10.3 Asymptotic Considerations

It has been a common practice in applied statistics to perform postselection inference, treating the selected model as if it had been known a priori. This practice can be problematic because it neglects the model selection process that can distort classical inference in the known-model context. Nevertheless, finding a general solution is challenging because model selection is often a complex process that defies characterization. In the context of this book, choosing u is the model selection step, although there could also be selection involved in the choice of the original multivariate model, prior to the introduction of envelopes.

The main point of this section is that in some settings it may be appropriate to treat $\widehat{u}$ as if it had been selected a priori, provided that $\Pr(\widehat{u} \neq u)$ is sufficiently small. This is achieved asymptotically if $\Pr(\widehat{u} = u) \to 1$ as $n \to \infty$, as it does with BIC. To state this result in detail, let $\widehat{\boldsymbol{\beta}}_u$ and $\widehat{\boldsymbol{\beta}}_{\widehat{u}}$ denote the envelope estimators of $\boldsymbol{\beta}$ at the true value and estimated values of u, and for any fixed vector $\mathbf{c} \in \mathbb{R}^{pr}$, let $\xi_n(u, \mathbf{c}) = \sqrt{n}\mathbf{c}^T \operatorname{vec}(\widehat{\boldsymbol{\beta}}_u - \boldsymbol{\beta})$. Then we know from Proposition 1.1 that $\xi_n(u, \mathbf{c})$ is asymptotically normal. The essential idea for the following proposition, which characterizes the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\widehat{u}}$, was provided by Zhang (2017).

Proposition 1.2 *Assume envelope model (1.20) holds. Then for any $\delta \in \mathbb{R}$ and any $\mathbf{c} \in \mathbb{R}^{pr}$,*

$$|\Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) - \Pr(\xi_n(u, \mathbf{c}) > \delta)| \leq \Pr(\widehat{u} \neq u).$$

In addition, if the method of selecting u is consistent, $\Pr(\widehat{u} = u) \to 1$ as $n \to \infty$, then as $n \to \infty$,

$$|\Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) - \Pr(\xi_n(u, \mathbf{c}) > \delta)| \to 0.$$

Proof: The proof of this proposition hinges on the fact that the parameter space for u is discrete.

$$\begin{aligned}
\Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) &= \Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta, \widehat{u} = u) + \Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta, \widehat{u} \neq u) \\
&= \Pr(\xi_n(u, \mathbf{c}) > \delta, \widehat{u} = u) + \Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta, \widehat{u} \neq u) \\
&\leq \Pr(\xi_n(u, \mathbf{c}) > \delta) + \Pr(\widehat{u} \neq u).
\end{aligned}$$

So

$$\Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) - \Pr(\xi_n(u, \mathbf{c}) > \delta) \leq \Pr(\widehat{u} \neq u).$$

Similarly,

$$\begin{aligned}
\Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) &= 1 - \Pr(\xi_n(\widehat{u}, \mathbf{c}) \leq \delta) \\
&= 1 - \Pr(\xi_n(\widehat{u}, \mathbf{c}) \leq \delta, \widehat{u} = u) - \Pr(\xi_n(\widehat{u}, \mathbf{c}) \leq \delta, \widehat{u} \neq u) \\
&\geq 1 - \Pr(\xi_n(u, \mathbf{c}) \leq \delta) - \Pr(\widehat{u} \neq u) \\
&= \Pr(\xi_n(u, \mathbf{c}) > \delta) - \Pr(\widehat{u} \neq u).
\end{aligned}$$

So

$$\Pr(\xi_n(u, \mathbf{c}) > \delta) - \Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) \leq \Pr(\widehat{u} \neq u).$$

Consequently,

$$|\Pr(\xi_n(\widehat{u}, \mathbf{c}) > \delta) - \Pr(\xi_n(u, \mathbf{c}) > \delta)| \leq \Pr(\widehat{u} \neq u) \to 0,$$

which is the desired conclusion. The Cramer–Wold device can be used to extend this to the asymptotic distributions of $\sqrt{n} \operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{u}} - \boldsymbol{\beta})$ and $\sqrt{n} \operatorname{vec}(\widehat{\boldsymbol{\beta}}_u - \boldsymbol{\beta})$. □

To illustrate the implications of Proposition 1.2, the parameter estimates $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\Sigma}}$ from the fit of the cattle data with $u = 3$ were taken as population values. A new set of data was then generated as

$$\mathbf{Y}_i = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\beta}} X_i + \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n^*,$$

where n^* denotes simulation sample size, and the errors $\boldsymbol{\varepsilon}_i$ are independent copies of a $N_{10}(0, \mathbf{I})$ random vector. Because the predictors are fixed, the simulation sample size $n^* = r \times n$ was increased in multiples r of the sample size $n = 60$ for the cattle data, repeating the 60 values of X for each multiple. The BIC estimates of u were then computed for 100 replicates of this scenario for each of the simulation sample sizes $n^* = 2^k n$, $k = 0, 1, \ldots, 6$. The percentage of replicates in which BIC selected the true value $u = 3$ is shown in Figure 1.8 for each n^*.

The simulation results indicate that BIC nearly always selects $u = 3$ at $n^* = 16n$ and thereafter, with $\Pr(\widehat{u} \neq u)$ estimated to be 0.01 at $n^* = 64n$. For regressions where the sample size may not be large enough to instill confidence that $\Pr(\widehat{u} \neq u)$ is sufficiently small, additional intuition and bootstrap methods for gaining data-analytic guidance are discussed later in this chapter. An alternative weighted estimator that does not require choosing a value for u is presented in Section 1.11.4.

Proposition 1.2 may have implications in other model-selection problems as well. Consider, for instance, the common practice of using the Box–Cox method to estimate a power transformation $Y^{(\lambda)}$ of a univariate response Y to induce a linear regression model. Can we reasonably treat the estimated power $\hat{\lambda}$ as being nonstochastic in subsequent inference statements? Or are we obliged to take the variability in $\hat{\lambda}$ into account? This issue was the focus of considerable discussion in the early 1980s (e.g. Bickel and Doksum 1981; Box and Cox 1982;

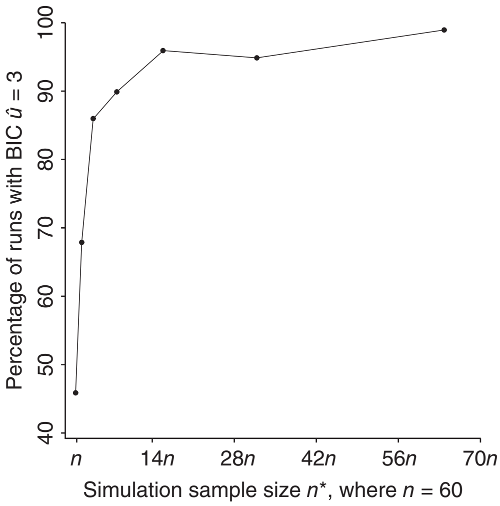


Figure 1.8 Cattle data: simulation results based on the fit of the cattle data with $u = 3$. The horizontal axis denotes the simulation sample size $n^* = 2^k n$, where $n = 60$ is the sample size for the original data.

Hinkley and Runger 1984, with discussion). If the transformation parameter λ is restricted to a small finite set of plausible values and $\Pr(\hat{\lambda} \neq \lambda)$ is sufficiently small, then it may well be reasonable to treat $\hat{\lambda}$ as nonstochastic.

1.10.4 Overestimation Versus Underestimation of *u*

For clarity in this section, we use u to denote the true dimension of the envelope, and let $u_{\rm f}$ denote the value used in fitting. In this discussion, we treat $u_{\rm f}$ as if it had been selected a priori. Overestimation occurs then $u_{\rm f} > u$, while underestimation occurs when $u_{\rm f} < u$. Overestimation is perhaps the less serious error. Fitting with $u_{\rm f} > u$ gives a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$, although the estimator will be more variable than that with $u_{\rm f} = u$. For instance, fitting with $u_{\rm f} = r$ reduces the regression to the standard model (1.1). Underestimation produces inconsistent estimators, and for this reason, it may be the more serious error.

We saw in Section 1.5.3 that the objective function serves to control a measure of bias. To gain intuition into the impact of underestimation of u, consider the population version of the representation of the objective function given in (1.29) allowing for fitting when $u_{\rm f} \neq u$. Let

$$(\tilde{\mathbf{G}}, \tilde{\mathbf{G}}_0) = \arg\min_{\mathcal{O}_{u_{\rm f}}} \Big\{ \log|\boldsymbol{\Sigma}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}| \\ + \log|\mathbf{I}_{r-u_{\rm f}} + \tilde{\boldsymbol{\beta}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\tilde{\boldsymbol{\beta}}^T_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}| \Big\},$$

where the minimum is taken over the set $\mathcal{O}_{u_{\rm f}}$ of all orthogonal $r \times r$ matrices $(\mathbf{G}, \mathbf{G}_0)$ with $\mathbf{G} \in \mathbb{R}^{r\times u_{\rm f}}$ and $\tilde{\boldsymbol{\beta}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} = (\mathbf{G}_0^T\boldsymbol{\Sigma}\mathbf{G}_0)^{-1/2}\mathbf{G}_0^T\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}$ is the population coefficient matrix for the regression of the standardized response $(\mathbf{G}_0^T\boldsymbol{\Sigma}\mathbf{G}_0)^{-1/2}$ $\mathbf{G}_0^T\mathbf{Y}$ on the standardized predictor $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathbf{X}$. The population version of the envelope estimator at $\mathbf{G}$ is $\mathbf{P}_{\mathbf{G}}\boldsymbol{\beta}$, so the bias from underestimation is $\mathbf{Q}_{\mathbf{G}}\boldsymbol{\beta} = \mathbf{G}_0\mathbf{G}_0^T\boldsymbol{\beta}$. Thus, $\tilde{\boldsymbol{\beta}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}$ is a standardized version of the coordinates $\mathbf{G}_0^T\boldsymbol{\beta}$ of this bias.

To facilitate exposition, we consider the case $p = 1$ and for clarity use σ_X^2 instead of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Then

$$B(\mathbf{G}_0) := \tilde{\boldsymbol{\beta}}^T_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}\tilde{\boldsymbol{\beta}}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}} = \boldsymbol{\beta}^T\mathbf{G}_0(\mathbf{G}_0^T\boldsymbol{\Sigma}\mathbf{G}_0)^{-1}\mathbf{G}_0^T\boldsymbol{\beta}\sigma_X^2$$

is a measure of the squared length of the bias at $\mathbf{G}_0$ for the standardized regression,

$$\log\left|\mathbf{I}_{r-u_{\rm f}} + \tilde{\boldsymbol{\beta}}_{\mathbf{G}_0^T\mathbf{Y}|X}\tilde{\boldsymbol{\beta}}^T_{\mathbf{G}_0^T\mathbf{Y}|X}\right| = \log\{1 + B(\mathbf{G}_0)\}$$

and

$$(\tilde{\mathbf{G}}, \tilde{\mathbf{G}}_0) = \arg\min_{\mathcal{O}_{u_{\rm f}}}\{\log|\boldsymbol{\Sigma}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}| + \log\{1 + B(\mathbf{G}_0)\}\}. \tag{1.46}$$

Our goal now is to provide a bound on the bias $B(\tilde{\mathbf{G}}_0)$ that results from the minimization in (1.46). Assume for ease of exposition that the eigenvalues of

$\boldsymbol{\Sigma}$ are distinct and let $\mathcal{V}_m$ denote the collection of all subsets of m eigenvectors of $\boldsymbol{\Sigma}$. Then the columns of any $\mathbf{G} \in \mathcal{V}_{u_\mathrm{f}}$ form a subset of u_f eigenvectors of $\boldsymbol{\Sigma}$, and $\mathbf{G}_0$ is the complementary subset of $r - u_\mathrm{f}$ eigenvectors. A bound on $B(\tilde{\mathbf{G}}_0)$ can now be constructed by minimizing (1.46) over $\mathcal{V}_{u_\mathrm{f}}$ instead of $\mathcal{O}_{u_\mathrm{f}}$. We know from Lemma A.14 that the sum $\log|\boldsymbol{\Sigma}_{\mathbf{G}^T\mathbf{Y}|\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{G}_0^T\mathbf{Y}|\mathbf{X}}|$ is minimized by any $\mathbf{G} \in \mathcal{V}_{u_\mathrm{f}}$ and that its minimum value is $\log|\boldsymbol{\Sigma}|$. Let $\dot{\mathbf{G}}_0 = \arg\min_{\mathbf{G}_0 \in \mathcal{V}_{r-u_\mathrm{f}}} B(\mathbf{G}_0)$. Then $B(\tilde{\mathbf{G}}_0) \le B(\dot{\mathbf{G}}_0)$ since, by construction,

$$\log|\boldsymbol{\Sigma}_{\tilde{\mathbf{G}}^T\mathbf{Y}|\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\tilde{\mathbf{G}}_0^T\mathbf{Y}|\mathbf{X}}| + \log\{1 + B(\tilde{\mathbf{G}}_0)\} \le \log|\boldsymbol{\Sigma}| + \log\{1 + B(\dot{\mathbf{G}}_0)\},$$

which implies that

$$\begin{aligned} 0 &\le \log|\boldsymbol{\Sigma}_{\tilde{\mathbf{G}}^T\mathbf{Y}|\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\tilde{\mathbf{G}}_0^T\mathbf{Y}|\mathbf{X}}| - \log|\boldsymbol{\Sigma}| \\ &\le \log\{1 + B(\dot{\mathbf{G}}_0)\} - \log\{1 + B(\tilde{\mathbf{G}}_0)\}. \end{aligned}$$

We can gain insights about the potential bias $B(\tilde{\mathbf{G}}_0)$ by studying its upper bound $B(\dot{\mathbf{G}}_0)$. Let $(\dot{\mathbf{G}}, \dot{\mathbf{G}}_0)$ be an orthogonal matrix. Since $\dot{\mathbf{G}} \in \mathcal{V}_{u_\mathrm{f}}$ reduces $\boldsymbol{\Sigma}$, we have the decomposition $\boldsymbol{\Sigma} = \dot{\mathbf{G}}\boldsymbol{\Lambda}\dot{\mathbf{G}}^T + \dot{\mathbf{G}}_0\boldsymbol{\Lambda}_0\dot{\mathbf{G}}_0^T$, where the $\boldsymbol{\Lambda}$'s are diagonal matrices of eigenvalues, and thus

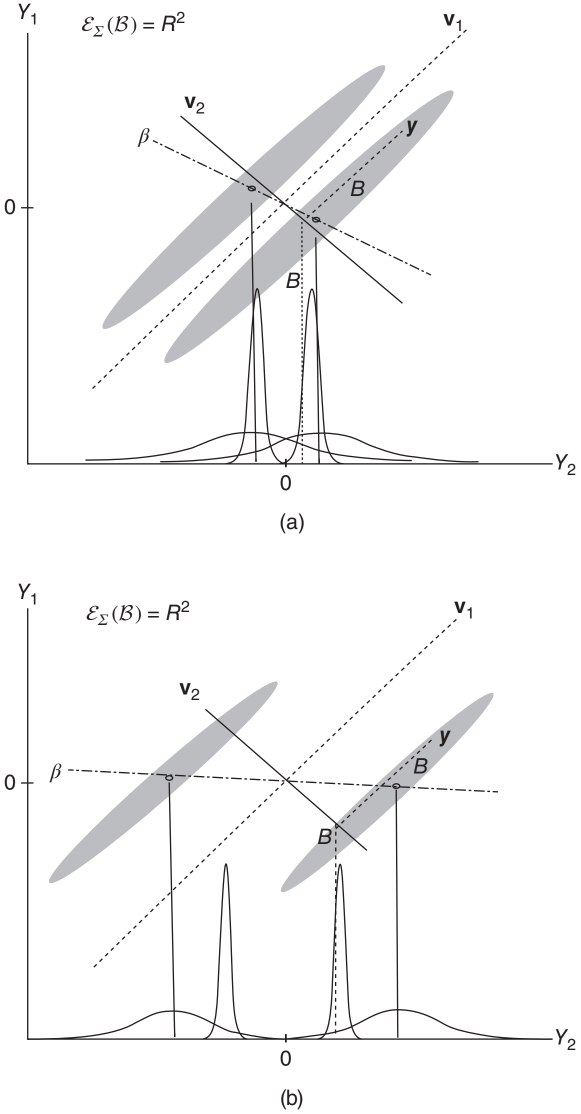
$$B(\dot{\mathbf{G}}_0) = \boldsymbol{\beta}^T\dot{\mathbf{G}}_0\boldsymbol{\Lambda}_0^{-1}\dot{\mathbf{G}}_0^T\boldsymbol{\beta}\sigma_X^2. \tag{1.47}$$

From this representation, we see that the bias bound depends on (i) the length of $\boldsymbol{\beta}$, (ii) the angles between $\boldsymbol{\beta}$ and the eigenvectors of $\boldsymbol{\Sigma}$ that comprise the columns of $\dot{\mathbf{G}}_0$, and (iii) the associated eigenvalues in $\boldsymbol{\Lambda}_0$. Since $\dot{\mathbf{G}}_0$ minimizes the bias, we expect that $\boldsymbol{\beta}$ will be orthogonal to multiple columns of $\dot{\mathbf{G}}_0$ when underestimating the dimension by one or two. Large eigenvalues $\boldsymbol{\Lambda}_0$ can also result in a small bias. As discussed in Section 1.4.2, envelopes result in substantial variance reduction when $\|\boldsymbol{\Omega}_0\| \gg \|\boldsymbol{\Omega}\|$. Consequently, we expect the bias due to underestimation to be small when the corresponding analysis shows substantial variance reduction. These comments are illustrated in Figure 1.9, which was constructed like the stylized illustration of Figure 1.3, except that $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta})$ no longer aligns with the smallest eigenvector of $\boldsymbol{\Sigma}$ and thus the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathbb{R}^2$. The figure illustrates the bias that can result when underestimating u with $u_\mathrm{f} = 1$. In Figure 1.9a, the bias is small because the angle between $\mathcal{B}$ and $\mathbf{v}_2$ is small and $\boldsymbol{\Lambda}_0$, the eigenvalue associated with $\mathbf{v}_1$, is large. The bias is larger in Figure 1.9b because we increased the length of $\boldsymbol{\beta}$ and the angle between $\mathcal{B}$ and $\mathbf{v}_2$. In this case, the dimension selection methods discussed in Section 1.10.1 will likely indicate correctly that $u = 2$, depending on the sample size, so underestimation of u may not be a worrisome issue.

Another perhaps more relevant gauge of bias is to compare $\mathbf{B}(\dot{\mathbf{G}})$ to the squared length of $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}\sigma_X$, which is $\boldsymbol{\beta}$ in the standardized scale akin to $\tilde{\boldsymbol{\beta}}_{\mathbf{G}_0^T\mathbf{Y}|X}$:

$$\frac{B(\dot{\mathbf{G}})}{\|\tilde{\boldsymbol{\beta}}\|^2} = \frac{\boldsymbol{\beta}^T\dot{\mathbf{G}}_0\boldsymbol{\Lambda}_0^{-1}\dot{\mathbf{G}}_0^T\boldsymbol{\beta}\sigma_X^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\sigma_X^2} = \frac{\tilde{\boldsymbol{\beta}}^T\dot{\mathbf{G}}_0\dot{\mathbf{G}}_0^T\tilde{\boldsymbol{\beta}}}{\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}}.$$

Figure 1.9 Illustrations of the potential bias that can result from underestimating *u*. The context is the same as that for Figure 1.3, $\mathbf{v}_1$ and $\mathbf{v}_2$ denote the eigenvectors of $\mathbf{\Sigma}$ and $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta})$. (a) Small bias; (b) large bias.



In this form, we see that the bias depends only on the angles between the standardized coefficients $\tilde{\boldsymbol{\beta}}$ and the eigenvectors in $\dot{\mathbf{G}}_0$.

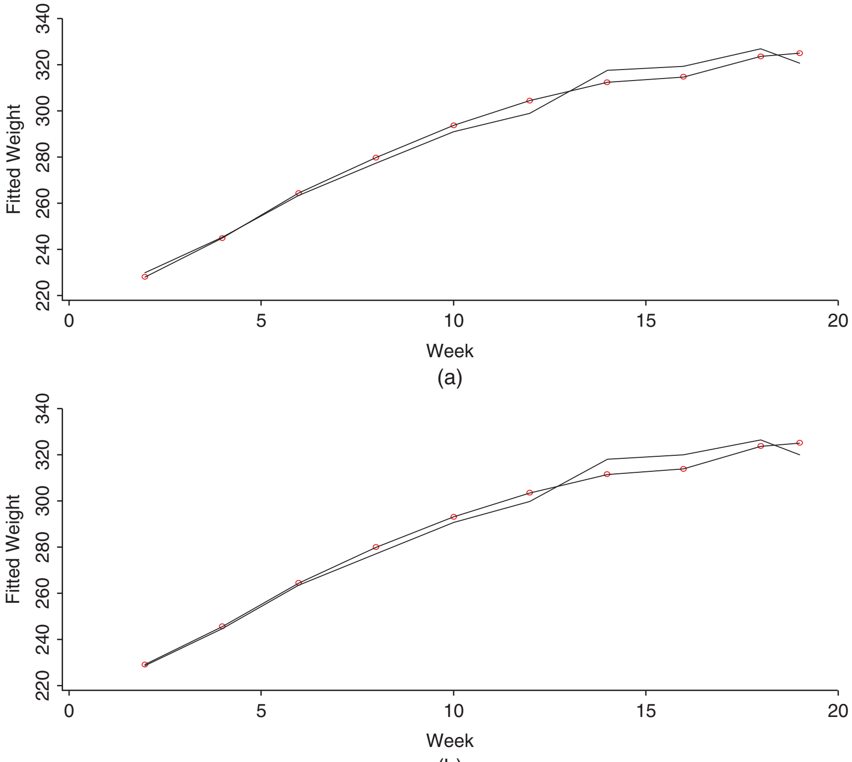
1.10.5 Cattle Weights: Influence of *u*

A first step in the analysis of the cattle data in Section 1.3.4 was to determine the dimension $u = \dim(\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}))$ of the envelope. Using the methods of this section, LRT(0.05) gave $u = 1$, which is the value used in the previous illustrations.

On the other hand, AIC and BIC gave $u = 3$, and fivefold cross-validation gave $u = 5$. In many applications, these methods agree on the value of u, but in some cases they disagree, as in this illustration. We tend to give preference to BIC and LRT since AIC has a propensity to overestimate u. Cross-validation is in effect a different criterion, and it is possible that $u = 5$ is best for prediction.

Nevertheless, when the selection methods disagree, we often find that the essential results are the same across the indicated values of u. To reinforce this point, Figure 1.10 shows the fitted profiles from envelope models with $u = 1$ and $u = 5$. These fitted profiles are very similar and lead to the same inference about their key features: the first detected differential treatment effects are around week 10 and persist thereafter.

Plots of the coordinates of $\mathbf{Q}_{\hat{\Gamma}}\mathbf{Y}$ versus the predictors, which show the estimated X-invariant part of $\mathbf{Y}$, can also be informative diagnostics. In these plots, $\mathbf{Q}_{\hat{\Gamma}}\mathbf{Y}$ should appear independent of the predictors if the model and choice of u are reasonable. Systematic variation with $\mathbf{X}$ may be an indication of model



(a)

(b)

Figure 1.10 Cattle data: fitted profile plots from envelope models with (a) $u = 1$ and (b) $u = 5$.

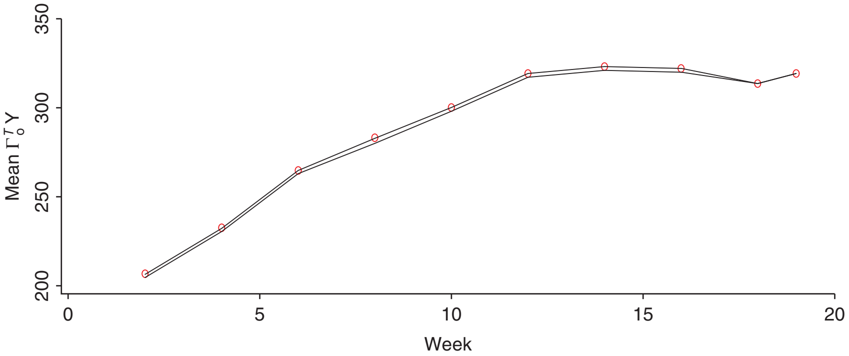


Figure 1.11 Cattle data: mean of immaterial variation $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ for each treatment and time with $u = 3$.

failure. For instance, shown in Figure 1.11 is a plot of the 10 coordinates in the projection $\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\bar{\mathbf{Y}}_j$ of the average response vector $\bar{\mathbf{Y}}_j$ by treatment $j = 1, 2$. The two curves, one for each treatment, are essentially identical and show no treatment clear effects. The variation in the curves themselves reflects the variation in $\mathbf{Y}$ that is common to the two treatments.

1.11 Bootstrap and Uncertainty in the Envelope Dimension

In the previous section, we discussed selection methods, when it might be reasonable to proceed as if $\widehat{u} = u$, overestimation versus underestimation, and the possibility that key aspects of inference are unaffected over a reasonable set of values for u. In this section, we turn to the bootstrap for standard errors and for assistance in addressing the uncertainty in $\widehat{u}$.

1.11.1 Bootstrap for Envelope Models

In this section, we describe how the residual bootstrap can be used to estimate the variance of $\widehat{\boldsymbol{\beta}}$ assuming that the envelope dimension u is known, perhaps relying on Proposition 1.2 to justify setting $u = \widehat{u}$. For emphasis, the envelope estimator at the true value is denoted $\widehat{\boldsymbol{\beta}}_u$ throughout our discussion of the bootstrap.

The variance of $\widehat{\boldsymbol{\beta}}_u$ can be obtained by using the asymptotic results of Section 1.6 or by using the residual bootstrap. Recall from the setup of Section 1.1 that $\mathbf{r}_i$ denotes the ith vector of residuals from the ordinary least squares fit of model (1.1). Let $\mathbf{R} \in \mathbb{R}^{n\times r}$ denote the matrix with residual vectors $\mathbf{r}_i^T$ as its rows, let $\mathbf{R}^*$ denote a resampled residual matrix constructed by sampling with replacement n rows of $\mathbf{R}$, and let $\mathbb{Y}^* = \mathbf{1}_n\widehat{\boldsymbol{\alpha}}^T + \mathbb{X}\widehat{\boldsymbol{\beta}}_u^T + \mathbf{R}^*$ denote the $n \times r$ matrix of bootstrapped responses. Then a single bootstrap envelope estimator $\widehat{\boldsymbol{\beta}}_u^*$ of $\boldsymbol{\beta}$ is found from the envelope fit of model (1.20) to the data $(\mathbb{Y}^*, \mathbb{X})$. Repeating this operation B times gives the bootstrap estimators $\widehat{\boldsymbol{\beta}}_{u,k}^*$, $k = 1, \ldots, B$. Then the sample variance of the $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{u,k}^*)$'s provides a bootstrap estimator of $\mathrm{var}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_u)\}$. The justification of this bootstrap follows from Andrews (2002, pp. 122–124 and Theorem 2) and from Eck (2017).

1.11.2 Wheat Protein: Bootstrap and Asymptotic Standard Errors, *u* Fixed

To illustrate application of the bootstrap, we consider the wheat protein data discussed in Section 1.3.3, but now with responses measured at six wavelengths instead of two. The dimension of the envelope model was estimated to be 1 by BIC, and for the purpose of this bootstrap it is assumed that $u = 1$. Two hundred (B) bootstrap samples were used throughout. The first part of Table 1.2 shows the estimated coefficients under the standard model (1.1) along with good agreement between their bootstrap and asymptotic standard errors. The second part of Table 1.2 shows the estimated envelope coefficients and the corresponding bootstrap and asymptotic standard errors. There is again a good agreement between the standard errors, which are much smaller than those for the standard model. The advantages of the envelope model in this fit are indicated roughly by the sizes of $\widehat{\boldsymbol{\Omega}} = 7.88$ and $\|\widehat{\boldsymbol{\Omega}}_0\| = 6517$. Thus, the envelope model has an apparent advantage because the variation $\widehat{\boldsymbol{\Omega}}_0$ in the estimated X-invariant part of $\mathbf{Y}$ is considerably larger than the variation $\widehat{\boldsymbol{\Omega}}$.

Table 1.2 Wheat protein data: bootstrap and asymptotic standard errors (SEs) of the six elements in $\widehat{\beta}$ under the standard (1.1) and envelope models (1.20) for the wheat protein data with six responses.

***1.* Standard model (1.1)**						
B	3.27	8.03	7.52	−2.06	3.22	0.65
Bootstrap SE	9.87	8.12	8.70	9.65	13.90	5.48
SE	9.78	8.12	8.70	9.49	13.65	5.39
2. Envelope model with $u = 1$						
$\widehat{\beta}_u$	−1.06	4.47	3.68	−5.97	0.69	−1.60
Bootstrap SE	0.35	0.48	0.39	0.64	0.20	0.69
Asymptotic SE (1.33)	0.35	0.43	0.35	0.59	0.21	0.86

1.11.3 Cattle Weights: Bootstrapping u

In this section, we use the cattle data with BIC to illustrate how the bootstrap and other sampling scenarios might be used to gain data-analytic guidance on the choice of u. Part 1 of Table 1.3 gives the BIC bootstrap distribution of u based on 100 residual bootstrap datasets constructed as described in Section 1.11.1. We see from Part 1 of Table 1.3 that most of the mass is concentrated between $u = 1$ and $u = 4$. The question is how we might use these results to help with the choice of u.

Let $\widehat{\boldsymbol{\alpha}}_u$, $\widehat{\boldsymbol{\beta}}_u$, and $\widehat{\boldsymbol{\Sigma}}_u$ denote the estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ from the fit of the envelope model with the indicated value of u and consider generating reference data as

$$\widehat{\mathbf{Y}}_i^{(u)} = \widehat{\boldsymbol{\alpha}}_u + \widehat{\boldsymbol{\beta}}_u X_i + \widehat{\boldsymbol{\Sigma}}_u^{1/2}\boldsymbol{\varepsilon}_i, \quad i = 1, \dots , 60, \tag{1.48}$$

where errors $\boldsymbol{\varepsilon}_i$ are independent copies of a $N_{10}(0, \mathbf{I})$ random vector. Part 2 of Table 1.3 shows the empirical distribution of the BIC estimate of u from 100 replications over the set of errors $\boldsymbol{\varepsilon}_i$, $i = 1, \dots , 60$, with the value of u used in the fit shown in the first column. The distributions for $u = 3, 5, 7$ are similar and also similar to the distribution from the residual bootstrap in Part 1, while the distribution for $u = 1$ is notably different. One possible explanation for this

Table 1.3 Cattle data: distributions of $\widehat{u}$ from BIC based on various sampling scenarios.

$\widehat{u}$	0	1	2	3	4	5	6	Mean	sd
1. Residual bootstrap									
BIC	0	0.11	0.25	0.33	0.22	0.07	0.02	2.95	1.17
2. Normal errors									
$u = 1$	0	0.65	0.28	0.06	0.01	0	0	1.43	0.66
$u = 3$	0	0.11	0.22	0.43	0.20	0.03	0.01	2.85	1.02
$u = 5$	0	0.14	0.28	0.36	0.19	0.02	0.01	2.70	1.05
$u = 7$	0	0.13	0.21	0.36	0.25	0.03	0.02	2.90	1.13
3. Residual bootstrap from normal errors									
$u = 1$	0	0.43	0.27	0.22	0.04	0.04	0	1.99	1.09
$u = 3$	0	0.13	0.25	0.48	0.08	0.05	0.01	2.70	1.02
$u = 5$	0	0.16	0.23	0.33	0.15	0.10	0.03	2.89	1.29
$u = 7$	0	0.11	0.23	0.40	0.19	0.07	0	2.81	1.06

(1) 100 residual bootstrap datasets, (2) fit with the indicated value of u plus normal errors, (3) residual bootstrap from one dataset generated from a fit with the indicated value of u plus normal errors.

is as follows. If the true $u = 3$, then a population fit with $u > 3$ must give a subspace that contains the envelope. Hence, we are again led back to the model with $u = 3$. In other words, the results $u = 3, 5, 7$ are roughly as expected if in fact the true $u = 3$. Part 3 of Table 1.3 shows the BIC bootstrap distribution of $\widehat{u}$ from one dataset generated from (1.48) based on 100 bootstrap samples. These distributions are similar to corresponding distributions shown in Part 2, so it again seems plausible that the true $u = 3$. These results give data-analytic support to the original BIC estimate $\widehat{u} = 3$. The estimate $\widehat{u} = 5$ might be used if we wanted to be conservative.

1.11.4 Bootstrap Smoothing

In this section, we expand our notation a bit and let $\widehat{\boldsymbol{\beta}}_j$ denote the envelope estimator computed by assuming $u = j$. The envelope estimator at the true value is still denoted $\widehat{\boldsymbol{\beta}}_u$. In the previous sections, we discussed how data-analytic methods might be used to help select the envelope dimension. In this section, we discussed one way to avoid choosing a particular dimension and instead use a weighted average of the estimators $\widehat{\boldsymbol{\beta}}_j$, $j = 1, \dots, r$. By constraining $j > 0$, we are considering only regressions in which $\boldsymbol{\beta} \neq 0$, which will be reasonable in numerous applications. Extensions of the following to allow $j = 0$ are straightforward.

Let $b_j = -2\widehat{L}_j + N_j \log(n)$ denote the BIC value (1.45) for the envelope model of dimension j, where $\widehat{L}_j$ is the value of the maximized log-likelihood (1.27), and N_j is the number of parameters (1.21), both for the envelope model of dimension j. The weighted estimator we consider, which was proposed by Eck and Cook (2017), is of the form

$$\widehat{\boldsymbol{\beta}}_w = \sum_{j=1}^{r} w_i \widehat{\boldsymbol{\beta}}_j, \quad \text{where } w_j = \frac{e^{-b_j}}{\sum_{i=1}^{r} e^{-b_i}}. \tag{1.49}$$

Estimators of this form have been advocated in various contexts (Efron 2014; Nguefack-Tsague 2014; Hjort and Claeskens 2003; Burnham and Anderson 2004; Buckland et al. 1997). The estimator (1.49) may be advantageous because it bypasses the need to select a specific dimension and automatically incorporates model uncertainty in inference via BIC. It can be more variable than a specific estimator, however. For instance, if $j \geq u$, then $\widehat{\boldsymbol{\beta}}_j$ corresponds to a true model, and the variability of $\widehat{\boldsymbol{\beta}}_j$ can be noticeably less than that of $\widehat{\boldsymbol{\beta}}_w$, depending on the sample size. The estimator $\widehat{\boldsymbol{\beta}}_w$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$, but analytic expressions of its asymptotic variance are unknown. However, the bootstrap can still be used to assess its variance.

Following the construction of $\widehat{\boldsymbol{\beta}}_w$, generate n bootstrap samples $\widehat{\boldsymbol{\beta}}^*_{w,k}$, $k = 1, \dots, B$, as described in Section 1.11.1, except replace $\widehat{\boldsymbol{\beta}}_u$ with $\widehat{\boldsymbol{\beta}}_w$ so the bootstrap responses are generated as $\mathbb{Y}^* = \mathbf{1}_n \widehat{\boldsymbol{\alpha}}^T + \mathbb{X}\widehat{\boldsymbol{\beta}}_w^T + \mathbf{R}^*$. The following proposition (Eck and Cook 2017) shows that the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_w^*$ is the same as the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_u^*$, the bootstrapped envelope estimator at the true dimension. In particular, the sample variance computed from the $\widehat{\boldsymbol{\beta}}_{w,k}^*$'s provides a $\sqrt{n}$-consistent estimator of the asymptotic variance of the envelope estimator $\widehat{\boldsymbol{\beta}}_u$.

Proposition 1.3 *Assume envelope model (1.20) for some $u = 1, \dots, r$, and that $\mathbf{S}_{\mathbf{X}}$ converges to $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$. Then as $n \to \infty$,*

$$\begin{aligned}\sqrt{n}\left\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_w^*) - \mathrm{vec}(\widehat{\boldsymbol{\beta}}_w)\right\} &= \sqrt{n}\left\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_u^*) - \mathrm{vec}(\widehat{\boldsymbol{\beta}}_u)\right\} \\ &\quad + O_p\left\{n^{(1/2-p)}\right\} + 2(u-1)O_p(1)\sqrt{n}e^{-n|O_p(1)|}.\end{aligned} \tag{1.50}$$

To gain some intuition about the orders in (1.50) write

$$\begin{aligned}\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_w^*) - \mathrm{vec}(\widehat{\boldsymbol{\beta}}_w)\} &= \sqrt{n}\sum_{j=1}^{u-1}\{w_j^*\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_j^*) - w_j\mathrm{vec}(\widehat{\boldsymbol{\beta}}_j)\} \\ &\quad + \sqrt{n}\{w_u^*\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_u^*) - w_u\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_u)\} \\ &\quad + \sqrt{n}\sum_{j=u+1}^{r}\{w_j^*\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_j^*) - w_j\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_j)\}.\end{aligned}$$

Eck and Cook (2017) show that for $j \neq u$, $\sqrt{n}\{w_j^*\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_j^*) - w_j\ \mathrm{vec}(\widehat{\boldsymbol{\beta}}_j)\} \to 0$, so the first and third terms on the right-hand side vanish as $n \to 0$. The term $O_p\{n^{(1/2-p)}\}$ in (1.50) corresponds to the rate at which $\sqrt{n}w_j$ and $\sqrt{n}w_j^*$ converge to 0 for $j = u+1, \dots, r$. This rate is a cost of overestimation of the envelope. It decreases quite fast, particularly when p is not small, because models with $j > u$ are true and thus have no systematic bias due to choosing the wrong dimension. The $2(u-1)\sqrt{n}e^{-n|O_p(1)|}$ term corresponds to the rate at which $\sqrt{n}w_j$ and $\sqrt{n}w_j^*$ vanish for $j = 1, \dots, u-1$. This rate arises from under estimating the envelope space and it is affected by bias arising from choosing the wrong dimension. Because we consider only regressions in which $u \geq 1$, this term is 0 when $u = 1$. When $u = 1$, underestimation is not possible in our context and thus the term vanishes.

1.11.5 Cattle Data: Bootstrap Smoothing

Returning to the cattle data, Table 1.4 gives ratios of the standard errors of the elements of $\mathbf{B}$ to the bootstrap standard errors of the elements of $\widehat{\boldsymbol{\beta}}_w$ (second column) and to the bootstrap standard errors of the elements of $\widehat{\boldsymbol{\beta}}_j$ from

Table 1.4 Cattle data: ratios of standard errors of the elements of **B** to the bootstrap standard errors from the weighted envelope fit and the envelope fits for $u = 1, 2, 3, 4$.

		Envelope dimension			
Week	**Weighted**	**1**	**2**	**3**	**4**
2	1.48	2.38	1.83	1.34	1.19
4	1.61	3.77	1.99	1.44	1.22
6	1.81	3.30	2.48	1.57	1.29
8	2.01	4.07	3.30	1.58	1.31
10	2.09	5.30	3.80	1.59	1.34
12	1.88	3.70	2.59	1.56	1.31
14	2.02	3.94	3.19	1.59	1.33
16	2.07	4.01	3.15	1.58	1.3
18	2.14	5.08	3.38	1.51	1.28
19	1.96	5.26	2.69	1.48	1.23

the envelope fits with $j = 1, 2, 3, 4$ (columns 3–6). The table illustrates that the weighted estimator can be usefully less variable than the standard estimator. It can also be seen from the table that the standard errors for the weighted estimator all lie between those for envelope estimators with $u = 2$ and $u = 3$.