Chapter 15
Collinearity and Alternative Estimates

Collinearity or multicollinearity refers to the problem in regression analysis of the columns of the model matrix being nearly linear dependent. Ideally, this is no problem at all. There are numerical difficulties associated with the actual computations, but there are no theoretical difficulties. If, however, one has any doubts about the accuracy of the model matrix, the analysis could be in deep trouble.

Section 1 discusses what collinearity is and what problems it can cause. Four techniques for dealing with collinearity are examined. These are regression in canonical form, principal component regression, generalized inverse regression, and classical ridge regression. The methods, other than classical ridge regression, are essentially the same. Section 4 presents additional comments on the potential benefits of biased estimation. Section 5 discusses penalized estimation. Finally, Section 6 presents an alternative to least squares estimation. Least squares minimizes the vertical distances between the dependent variable and the regression surface. Section 6 considers minimizing the perpendicular distance between the dependent variable and the regression surface.

15.1 Defining Collinearity

In this section we define the problem of collinearity. The approach taken is to quantify the idea of having columns of the model matrix that are "nearly linearly dependent." The effects of near linear dependencies are examined. The section concludes by establishing the relationship between the definition given here and other commonly used concepts of collinearity.

Suppose we have a regression model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I, \tag{1}$$

where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, and $r(X) = p$. The essence of model (1) is that $\mathrm{E}(Y) \in C(X)$. Suppose that the model matrix consists of some predictor

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*,
Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3_15,

variables, say $x_1, x_2, \ldots, x_p$, that are measured with some small error. A near linear dependence in the observed model matrix X could mean a real linear dependence in the underlying model matrix of variables measured without error. Let X_* be the underlying model matrix. If the columns of X_* are linearly dependent, there exists an infinite number of least squares estimates for the true regression coefficients. If X is nearly linearly dependent, the estimated regression coefficients may not be meaningful and may be highly variable.

The real essence of this particular problem is that $C(X)$ is too large. Generally, we hope that in some sense, $C(X)$ is close to $C(X_*)$. Regression should work well precisely when this is the case. However, when X_* has linearly dependent columns, X typically will not. Thus $r(X) > r(X_*)$. $C(X_*)$ may be close to some proper subspace of $C(X)$, but $C(X)$ has extra dimensions. By pure chance, these extra dimensions could be very good at explaining the Y vector that happened to be observed. In this case, we get an apparently good fit that has no real world significance.

The extra dimensions of $C(X)$ are due to the existence of vectors b such that $X_*b = 0$ but $Xb \neq 0$. If the errors in X are small, then Xb should be approximately zero. We would like to say that a vector w in $C(X)$ is ill-defined if there exists b such that $w = Xb$ is approximately zero, where w is approximately the zero vector if its length is near zero. Unfortunately, multiplying w by a scalar can increase or decrease the length of the vector arbitrarily, while not changing the direction determined within $C(X)$. To rectify this, we can restrict attention to vectors b with $b'b = 1$ (i.e., the length of b is 1), or, equivalently, we make the following:

Definition 15.1.1. A vector $w = Xb$ is said to be ε *ill-defined* if $w'w/b'b = b'X'Xb/b'b < \varepsilon$. The matrix X is ε ill-defined if any vector in $C(X)$ is ε ill-defined. We use the terms "ill-defined" and "ill-conditioned" interchangeably.

The assumption of a real linear dependence in the X_* matrix is a strong one. We now indicate how that assumption can be weakened. Let $X = X_* + \Delta$, where the elements of Δ are uniformly small errors. Consider the vector Xb. (For simplicity, assume $b'b = 1$.) The corresponding direction in the underlying model matrix is X_*b.

Note that $b'X'Xb = b'X_*'X_*b + 2b'\Delta'X_*b + b'\Delta'\Delta b$. The vector $\Delta'b$ is short; so if Xb and X_*b are of reasonable size, they have about the same length. Also

$$b'X_*'Xb = b'X_*'X_*b + b'X_*'\Delta b, \tag{2}$$

where $b'X_*'\Delta b$ is small; so the angle between Xb and X_*b should be near zero. (For any two vectors x and y, let θ be the angle between x and y. Then $x'y = \sqrt{x'x}\sqrt{y'y}\cos\theta$.) On the other hand, if Xb is ill-defined, X_*b will also be small, and the term $b'X_*'\Delta b$ could be a substantial part of $b'X_*'Xb$. Thus, the angle between Xb and X_*b could be substantial. In that case, the use of the direction Xb is called into question because it may be substantially different from the underlying direction X_*b. In practice, we generally cannot know if the angle between Xb and X_*b is large. Considerable care must be taken when using a direction in $C(X)$ that is ill-defined.

So far we have not discussed whether the columns of X should be adjusted for their mean values or rescaled. I think that one should not be dogmatic about this issue; however, it should be noted that if the squared length of a column of X is less than ε, that direction will be ε ill-defined, regardless of what other vectors are in $C(X)$. The question of standardizing the columns of X arises frequently.

The intercept term is frequently handled separately from all other variables in techniques for dealing with collinearity. The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1} + e_i \tag{3}$$

$$Y = [J, Z] \begin{bmatrix} \beta_0 \\ \beta_* \end{bmatrix} + e$$

is often rewritten as

$$y_i = \alpha + \beta_1 (x_{i1} - \bar{x}_{\cdot 1}) + \ldots + \beta_{p-1} (x_{i\,p-1} - \bar{x}_{\cdot\,p-1}) + e_i, \tag{4}$$

or

$$Y = \left[J, \left(I - \frac{1}{n} J_n^n \right) Z \right] \begin{bmatrix} \alpha \\ \beta_* \end{bmatrix} + e,$$

where $\bar{x}_{\cdot j} = n^{-1} \sum_{i=1}^n x_{ij}$. It is easily seen that β_* is the same in both models, but $\beta_0 \neq \alpha$. We dispense with the concept of the underlying model matrix and write

$$X_* = \left(I - \frac{1}{n} J_n^n \right) Z.$$

Because of orthogonality, $\hat{\alpha} = \bar{y}_{\cdot}$, and β_* can be estimated from the model

$$Y_* = X_* \beta_* + e, \tag{5}$$

where $Y_*' = [y_1 - \bar{y}_{\cdot}, y_2 - \bar{y}_{\cdot}, \ldots, y_n - \bar{y}_{\cdot}]$.

Frequently, the scales of the x variables are also standardized. Let $q_j^2 = \sum_{i=1}^n (x_{ij} - \bar{x}_{\cdot j})^2$. Model (5) is equivalent to

$$Y_* = X_* \mathrm{Diag}(q_j^{-1}) \gamma_* + e, \tag{6}$$

where $\gamma_* = \mathrm{Diag}(q_j) \beta_*$. In model (6), the model matrix is $X_* \mathrm{Diag}(q_j^{-1})$.

Model (3) is rarely used in techniques for dealing with collinearity. Usually model (6) is assumed, and sometimes model (5). To retain full generality, our discussion uses model (3), but all the results apply to models (5) and (6). Note that the matrix $X_*' X_*$ is proportional to the sample covariance matrix of the Xs when $(x_{i1}, \ldots, x_{i\,p-1})$, $i = 1, \ldots, n$, is thought of as a sample of size n from a $p-1$ dimensional random vector. $\mathrm{Diag}(q_j^{-1}) X_*' X_* \mathrm{Diag}(q_j^{-1})$ is the sample correlation matrix.

We now present the relationship between ε ill-defined matrices and three other commonly used methods of identifying ill-conditioned matrices.

One of the main tools in the examination of collinearity is the examination of the eigenvalues of $X'X$. We discuss the relationship between Definition 15.1.1 and an eigen-analysis of $X'X$.

Recall that X has linearly dependent columns if and only if $X'X$ is singular, which happens if and only if $X'X$ has a zero eigenvalue. One often-used (but I think unintuitive) definition is that the columns of X are nearly linearly dependent if $X'X$ has at least one small eigenvalue. Suppose that $v_1, \ldots, v_p$ is an orthogonal set of eigenvectors for $X'X$ corresponding to the eigenvalues $\delta_1, \delta_2, \ldots, \delta_p$. Then $v_1, \ldots, v_p$ form a basis for $\mathbf{R}^p$. If $\delta_i < \varepsilon$, then $\delta_i = v_i'X'Xv_i/v_i'v_i$; so Xv_i is a direction in $C(X)$ that is ε ill-defined. Conversely, if an ε ill-defined vector exists, we show that at least one of the δ_is must be less than ε. Let $w = Xd$ be ε ill-defined. Write $d = \sum_{i=1}^p \alpha_i v_i$. Then $w'w = d'X'Xd = \sum_{i=1}^p \alpha_i^2 \delta_i$ and $d'd = \sum_{i=1}^p \alpha_i^2$. Since $\sum_{i=1}^p \alpha_i^2 \delta_i / \sum_{i=1}^p \alpha_i^2 < \varepsilon$, and since the δ_is are all nonnegative, at least one of the δ_is must be less than ε. We have proved:

Theorem 15.1.2. The matrix X is ε ill-defined if and only if $X'X$ has an eigenvalue less than ε.

The orthogonal eigenvectors of $X'X$ lead to a useful breakdown of $C(X)$ into orthogonal components. Let $\delta_1, \ldots, \delta_r$ be eigenvalues of at least ε and $\delta_{r+1}, \ldots, \delta_p$ eigenvalues less than ε. It is easily seen that $Xv_{r+1}, \ldots, Xv_p$ form an orthogonal basis for a subspace of $C(X)$ in which all vectors are ε ill-defined. The space spanned by $Xv_1, \ldots, Xv_r$ is a subspace in which none of the vectors are ε ill-defined. These two spaces are orthogonal complements with respect to $C(X)$. (Note that by taking a linear combination of a vector in each of the orthogonal subspaces one can get a vector that is ε ill-defined, but is in neither subspace.)

A second commonly used method of identifying ill-conditioned matrices was presented by Belsley, Kuh, and Welsch (1980). See also Belsley (1991). They use the *condition number* (*CN*) as the basis of their definition of collinearity. If the columns of X are rescaled to have length 1, the condition number is

$$CN \equiv \sqrt{\max_i \delta_i \Big/ \min_i \delta_i}.$$

Large values of the condition number indicate that X is ill-conditioned.

Theorem 15.1.3.

(a) If $CN > \sqrt{p/\varepsilon}$, then X is ε ill-defined.
(b) If X is ε ill-defined, then $CN > \sqrt{1/\varepsilon}$.

PROOF. See Exercise 15.2. □

If the columns of X have not been rescaled to have length 1, the *CN* is inappropriate. If X has two orthogonal columns, with one of length 10^3 and one of length 10^{-3},

the condition number is 10^6. Rescaling would make the columns of X orthonormal, which is the ideal of noncollinearity. (Note that such an X matrix is also ill-defined for any $\varepsilon > 10^{-6}$.)

Finally, in Section 14.2 the tolerance was used to measure collinearity. In Section 14.2 the motivation was to exclude from consideration any variables that were too similar to other variables already in the model. The relationship between the concept of tolerance and ill-defined vectors is complex. The tolerance is defined as $T = 1 - R^2$. R^2, as a measure of predictive ability, examines the predictive ability after correcting *all* variables for their means. While this is usually appropriate, in the current discussion it would be more appropriate to define tolerance as

$$T = \text{sum of squares error}/\text{sum of squares total(uncorrected)}.$$

If it is decided to adjust all variables for their means, this becomes the usual definition. We will use this alternative definition of T.

In a more subtle way, T also adjusts for the scale of $X_p = [x_{1p}, \ldots, x_{np}]'$. The scale adjustment comes because T is the ratio of two squared lengths. This issue arises again in our later analysis.

Rewrite model (14.2.1) as

$$X_p = X\beta + e. \tag{7}$$

The new dimension added to $C(X)$ by including the variable x_p in the regression is the vector of residuals from fitting model (7), i.e., $X_p - X\hat{\beta}$. Our question is whether $X_p - X\hat{\beta}$ is ill-defined within $C(X, X_p)$.

Let M be the perpendicular projection operator onto $C(X)$. By definition,

$$\begin{aligned} T &= X_p'(I-M)X_p / X_p'X_p \\ &= X_p'(I-M)X_p / [X_p'(I-M)X_p + X_p'MX_p]. \end{aligned}$$

T is small when $X_p'(I-M)X_p$ is small relative to $X_p'MX_p$.

The residual vector, $X_p - X\hat{\beta}$, is ε ill-defined in $C(X, X_p)$ if

$$\varepsilon > \frac{(X_p - X\hat{\beta})'(X_p - X\hat{\beta})}{1 + \hat{\beta}'\hat{\beta}} = \frac{X_p'(I-M)X_p}{1 + \hat{\beta}'\hat{\beta}}.$$

By examining the regression in canonical form (see Section 5 and Exercise 15.3), it is easy to see that

$$\delta_1 \frac{X_p'(I-M)X_p}{\delta_1 + X_p'MX_p} \leq \frac{X_p'(I-M)X_p}{1 + \hat{\beta}'\hat{\beta}} \leq \delta_p \frac{X_p'(I-M)X_p}{\delta_p + X_p'MX_p},$$

where δ_1 and δ_p are the smallest and largest eigenvalues of $X'X$. Note that for positive a and b, $a/(a+b) < \varepsilon$ implies $a/(\delta_p + b) < a/b < \varepsilon/(1-\varepsilon)$; so it follows immediately that if $T < \varepsilon$, $X_p - X\hat{\beta}$ is $\delta_p\varepsilon/(1-\varepsilon)$ ill-defined. On the other hand, if $X_p - X\hat{\beta}$ is ε ill-defined, some algebra shows that

$$T < \frac{\delta_1 \varepsilon}{(\delta_1 + \varepsilon) X_p' X_p} + \frac{\varepsilon}{\delta_1 + \varepsilon}.$$

By picking ε small, the tolerance can be made small. This bound depends on the squared length of X_p. In practice, however, if X_p is short, X_p may not have small tolerance, but the vector of residuals may be ill-defined. If X_p is standardized so that $X_p' X_p = 1$, this problem is eliminated. Standardizing the columns of X also ensures that there are some reasonable limits on the values of δ_1 and δ_p. (One would assume in this context that X is not ill-defined.)

Exercise 15.1 Show that any linear combination of ill-defined orthonormal eigenvectors is ill-defined. In particular, if $w = X(av_i + bv_j)$, then

$$\frac{w'w}{(av_i + bv_j)'(av_i + bv_j)} \leq \max(\delta_i, \delta_j).$$

Exercise 15.2 Prove Theorem 15.1.3.

Hint: For a model matrix with columns of length 1, $\text{tr}(X'X) = p$. It follows that $1 \leq \max_i \delta_i \leq p$.

15.2 Regression in Canonical Form and on Principal Components

Regression in canonical form involves transforming the Y and β vectors so that the model matrix is particularly nice. Regression in canonical form is closely related to two procedures that have been proposed for dealing with collinearity. To transform the regression problem we need

Theorem 15.2.1. *The Singular Value Decomposition.*
Let X be an $n \times p$ matrix with rank p. Then X can be written as

$$X = ULV',$$

where U is $n \times p$, L is $p \times p$, V is $p \times p$, and

$$L = \text{Diag}(\lambda_j).$$

The λ_js are the square roots of the eigenvalues (singular values) of $X'X$ (i.e., $\lambda_j^2 = \delta_j$). The columns of V are orthonormal eigenvectors of $X'X$ with

$$X'XV = VL^2,$$

and the columns of U are p orthonormal eigenvectors of XX' with

$$XX'U = UL^2.$$

PROOF. We can pick L and V as indicated in the theorem. Note that since $X'X$ is nonsingular, $\lambda_j > 0$ for all j. We need to show that we can find U so that the theorem holds. If we take $U = XVL^{-1}$, then

$$XX'U = XX'XVL^{-1} = XVL^2L^{-1} = XVL = XVL^{-1}L^2 = UL^2;$$

so the columns of U are eigenvectors of XX' corresponding to the λ_i^2s. The columns of U are orthonormal because the columns of V are orthonormal and

$$U'U = L^{-1}V'X'XVL^{-1} = L^{-1}V'VL^2L^{-1} = L^{-1}IL = I.$$

Having found U we need to show that $X = ULV'$. Note that

$$U'XV = U'XVL^{-1}L = U'UL = L,$$

thus

$$UU'XVV' = ULV'.$$

Note that $VV' = I$ and, by Proposition B.54, $C(U) = C(X)$; so by Theorem B.35, UU' is the perpendicular projection operator onto $C(X)$. Hence, $X = ULV'$. □

We can now derive the canonical form of a regression problem. Consider the linear model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I. \tag{1}$$

Write $X = ULV'$ as in Theorem 15.2.1 and write $U_* = [U, U_1]$, where the columns of U_* are an orthonormal basis for $\mathbf{R}^n$. Transform model (1) to

$$U_*'Y = U_*'X\beta + U_*'e. \tag{2}$$

Let $Y_* = U_*'Y$ and $e_* = U_*'e$. Then

$$\mathrm{E}(e_*) = 0, \quad \mathrm{Cov}(e_*) = \sigma^2 U_*'U_* = \sigma^2 I.$$

Using Theorem 15.2.1 again,

$$U_*'X\beta = U_*'ULV'\beta = \begin{bmatrix} U' \\ U_1' \end{bmatrix} ULV'\beta = \begin{bmatrix} L \\ 0 \end{bmatrix} V'\beta.$$

Reparameterizing by letting $\gamma = V'\beta$ gives the canonical regression model

$$Y_* = \begin{bmatrix} L \\ 0 \end{bmatrix} \gamma + e_*, \quad \mathrm{E}(e_*) = 0, \quad \mathrm{Cov}(e_*) = \sigma^2 I. \tag{3}$$

Since this was obtained by a nonsingular transformation of model (1), it contains all of the information in model (1).

Estimation of parameters becomes trivial in this model:

$$\hat{\gamma} = (L^{-1}, 0)Y_* = L^{-1}U'Y,$$

$$\mathrm{Cov}(\hat{\gamma}) = \sigma^2 L^{-2},$$

$$RSS = Y_*' \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix} Y_*,$$

$$RMS = \sum_{i=p+1}^{n} y_{*i}^2 \Big/ (n-p).$$

In particular, the estimate of γ_j is $\hat{\gamma}_j = y_{*j}/\lambda_j$, and the variance of $\hat{\gamma}_j$ is σ^2/λ_j^2. The estimates $\hat{\gamma}_j$ and $\hat{\gamma}_k$ have zero covariance if $j \neq k$.

Models (1) and (3) are also equivalent to

$$Y = UL\gamma + e, \tag{4}$$

so, writing $U = [u_1, \ldots, u_p]$ and $V = [v_1, \ldots, v_p]$, γ_j is the coefficient in the direction $\lambda_j u_j$. If λ_j is small, $\lambda_j u_j = Xv_j$ is ill-defined, and the variance of $\hat{\gamma}_j$, σ^2/λ_j^2, is large. Since the variance is large, it will be difficult to reject $H_0 : \gamma_j = 0$. If the data are consistent with $\gamma_j = 0$, life is great. We conclude that there is no evidence of an effect in the ill-defined direction $\lambda_j u_j$. If $H_0 : \gamma_j = 0$ is rejected, we have to weigh the evidence that the direction u_j is important in explaining Y against the evidence that the ill-defined direction u_j should not be included in $C(X)$.

Regression in canonical form can, of course, be applied to models (15.1.5) and (15.1.6), where the predictor variables have been standardized.

Exercise 15.3 Show the following.

(a) $Y'MY = \sum_{i=1}^{p} y_{*i}^2$.

(b) $Y'(I-M)Y = \sum_{i=p+1}^{n} y_{*i}^2$.

(c) $\hat{\beta}'\hat{\beta} = \sum_{i=1}^{p} y_{*i}^2/\lambda_i^2$.

(d) If $\lambda_1^2 \leq \cdots \leq \lambda_p^2$, then

$$\lambda_1^2 \frac{Y'(I-M)Y}{\lambda_1^2 + Y'MY} \leq \frac{Y'(I-M)Y}{1+\hat{\beta}'\hat{\beta}} \leq \lambda_p^2 \frac{Y'(I-M)Y}{\lambda_p^2 + Y'MY}.$$

15.2.1 Principal Component Regression

If the direction u_j is ill-defined, we may decide that the direction should not be used for estimation. Not using the direction u_j amounts to setting $\gamma_j = 0$ in model (4). If ill-defined directions are not to be used, and if ill-defined is taken to mean that $\lambda_j < \varepsilon$

for some small value of ε, then we can take as our estimate of γ, $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)'$, where

$$\tilde{\gamma}_j = \begin{cases} \hat{\gamma}_j, & \text{if } \lambda_j \geq \varepsilon \\ 0, & \text{if } \lambda_j < \varepsilon \end{cases}.$$

As an estimate of β in the original model (1) we can use $\tilde{\beta} = V\tilde{\gamma}$. This is reasonable because $V'\beta = \gamma$; so $V\gamma = VV'\beta = \beta$.

If we take as our original model a standardized version such as (15.1.5) or (15.1.6), the model matrix UL of model (4) has columns that are the principal components of the multivariate data set $(x_{i1}, \ldots, x_{ip-1})'$, $i = 1, \ldots, n$. See Johnson and Wichern (1988) or Christensen (2001) for a discussion of principal components. The procedure outlined here for obtaining an estimate of β is referred to as *principal component regression*. See Christensen (1996a, Section 15.6) for an example.

15.2.2 Generalized Inverse Regression

To deal with collinearity, Marquardt (1970) suggested using the estimate

$$\tilde{\beta} = (X'X)_r^- X'Y,$$

where

$$(X'X)_r^- = \sum_{j=p-r+1}^{p} v_j v_j' / \lambda_j^2,$$

and the λ_js are written so that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$. $(X'X)_r^-$ would be the (Moore–Penrose) generalized inverse of $(X'X)$ if $r(X'X) = r$, i.e., if $0 = \lambda_1 = \cdots = \lambda_{p-r}$. Since $X = ULV'$,

$$\tilde{\beta} = (X'X)_r^- X'Y = \sum_{j=p-r+1}^{p} v_j v_j' VLU'Y / \lambda_j^2 = \sum_{j=p-r+1}^{p} v_j \hat{\gamma}_j.$$

This is the same procedure as principal component regression. Marquardt originally suggested this as an alternative to classical ridge regression, which is the subject of the next section.

15.3 Classical Ridge Regression

Ridge regression was originally proposed by Hoerl and Kennard (1970) as a method to deal with collinearity. Now it is more commonly viewed as a form of penalized likelihood estimation, which makes it a form of Bayesian estimation. In this section, we consider the traditional view of ridge regression. In the penultimate section we relate ridge regression to penalty functions.

Hoerl and Kennard (1970) looked at the mean squared error, $\mathrm{E}[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)]$, for estimating β with least squares. This is the expected value of a quadratic form in $(\hat{\beta}-\beta)$. $\mathrm{E}(\hat{\beta}-\beta)=0$ and $\mathrm{Cov}(\hat{\beta}-\beta)=\sigma^2(X'X)^{-1}$; so by Theorem 1.3.2

$$\mathrm{E}[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)]=\mathrm{tr}[\sigma^2(X'X)^{-1}].$$

If $\lambda_1^2,\ldots,\lambda_p^2$ are the eigenvalues of $(X'X)$, we see that $\mathrm{tr}[(X'X)^{-1}]=\sum_{j=1}^{p}\lambda_j^{-2}$; so

$$\mathrm{E}[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)]=\sigma^2\sum_{j=1}^{p}\lambda_j^{-2}.$$

If some of the values λ_j^2 are small, the mean squared error will be large.

Hoerl and Kennard suggested using the estimate

$$\tilde{\beta}=(X'X+kI)^{-1}X'Y, \tag{1}$$

where k is some fixed scalar. The choice of k will be discussed briefly later. The consequences of using this estimate are easily studied in the canonical regression model. The canonical regression model (15.2.3) is

$$Y_*=\begin{bmatrix}L\\0\end{bmatrix}\gamma+e_*.$$

The ridge regression estimate is

$$\tilde{\gamma}=(L'L+kI)^{-1}[L',0]Y_*=(L^2+kI)^{-1}L^2\hat{\gamma}. \tag{2}$$

In particular,

$$\tilde{\gamma}_j=\frac{\lambda_j^2}{\lambda_j^2+k}\hat{\gamma}_j.$$

If λ_j is small, $\tilde{\gamma}_j$ will be shrunk toward zero. If λ_j is large, $\tilde{\gamma}_j$ will change relatively little from $\hat{\gamma}_j$.

Exercise 15.4 illustrates the relationship between ridge regression performed on the canonical model and ridge regression performed on the usual model. The transformation matrix V is defined as in Section 2.

Exercise 15.4 Use equations (1) and (2) to show that

(a) $\tilde{\beta}=V\tilde{\gamma}$,

(b) $\mathrm{E}[(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)]=\mathrm{E}[(\tilde{\gamma}-\gamma)'(\tilde{\gamma}-\gamma)]$.

The estimate $\tilde{\beta}$ has expected mean square

$$\begin{aligned}\mathrm{E}[(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)]&=\sigma^2\mathrm{tr}[(X'X+kI)^{-1}X'X(X'X+kI)^{-1}]\\&\quad+\beta'\left\{(X'X+kI)^{-1}X'X-I\right\}'\left\{(X'X+kI)^{-1}X'X-I\right\}\beta.\end{aligned}$$

Writing $X'X = VL^2V'$, $I = VV'$, and in the second term $I = (X'X+kI)^{-1}(X'X+kI)$, so that $(X'X+kI)^{-1}X'X - I = -(X'X+kI)^{-1}kI$, this can be simplified to

$$\mathrm{E}[(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)] = \sigma^2 \sum_{j=1}^{p} \lambda_j^2/(\lambda_j^2+k)^2 + k^2\beta'(X'X+kI)^{-2}\beta.$$

The derivative of this with respect to k at $k=0$ can be shown to be negative. Since $k=0$ is least squares estimation, in terms of mean squared error there exists $k > 0$ that gives better estimates of β than least squares. Unfortunately, the particular values of such k are not known.

Frequently, a *ridge trace* is used to determine k. A ridge trace is a simultaneous plot of the estimated regression coefficients (which are functions of k) against k. The value of k is chosen so that the regression coefficients change little for any larger values of k.

Because the mean squared error, $\mathrm{E}[(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)]$, puts equal weight on each regression coefficient, it is often suggested that ridge regression be used only on model (15.1.6).

The ridge regression technique admits obvious generalizations. One is to use $\tilde{\beta} = (X'X+K)^{-1}X'Y$, where $K = \mathrm{Diag}(k_j)$. The ridge estimates for canonical regression become

$$\tilde{\gamma}_j = \frac{\lambda_j^2}{\lambda_j^2+k_j}\hat{\gamma}_j.$$

The ridge regression estimate (1) can also be arrived at from a Bayesian argument. With $Y = X\beta + e$ and $e \sim N(0,\sigma^2 I)$, incorporating prior information of the form $\beta|\sigma^2 \sim N[0,(\sigma^2/k)I]$ leads to fitting a version of (2.9.3) that has

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ I \end{bmatrix}\beta + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix}, \quad \begin{bmatrix} e \\ \tilde{e} \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{n\times 1} \\ 0_{p\times 1} \end{bmatrix}, \sigma^2 \begin{bmatrix} I_n & 0 \\ 0 & (1/k)I_p \end{bmatrix} \right).$$

It is easily seen that the generalized least squares estimate of β associated with this model is (1). When k is near 0, the prior variance is large, so the prior information is very weak and the posterior mean is very close to the least squares estimate.

Exercise 15.5 Evaluate the data of Exercise 14.6 for collinearity problems and if necessary apply an appropriate procedure.

Exercise 15.6 It could be argued the the canonical model should be standardized before applying ridge regression. Define an appropriate standardization and show that under this standardization

$$\tilde{\gamma}_j = \frac{1}{1+k_j}\hat{\gamma}_j.$$

15.4 More on Mean Squared Error

For the canonical regression model, Goldstein and Smith (1974) have shown that for $0 \le h_j \le 1$, if $\tilde{\gamma}_j$ is defined by

$$\tilde{\gamma}_j = h_j \hat{\gamma}_j$$

and if

$$\frac{\gamma_j^2}{\text{Var}(\hat{\gamma}_j)} < \frac{1+h_j}{1-h_j}, \tag{1}$$

then $\tilde{\gamma}_j$ is a better estimate than $\hat{\gamma}_j$ in that

$$\text{E}(\tilde{\gamma}_j - \gamma_j)^2 \le \text{E}(\hat{\gamma}_j - \gamma_j)^2.$$

In particular, if

$$\gamma_j^2 < \text{Var}(\hat{\gamma}_j) = \sigma^2/\lambda_j^2, \tag{2}$$

then $\tilde{\gamma}_j = 0$ is a better estimate than $\hat{\gamma}_j$. Estimating σ^2 with RMS and γ_j with $\hat{\gamma}_j$ leads to taking $\tilde{\gamma}_j = 0$ if the absolute t statistic for testing $H_0 : \gamma_j = 0$ is less than 1. This is only an approximation to condition (2), so taking $\tilde{\gamma}_j = 0$ only for larger values of the t statistic may well be justified.

For ridge regression, $h_j = \lambda_j^2/(\lambda_j^2 + k_j)$, and condition (1) becomes

$$\gamma_j^2 < \frac{\sigma^2}{\lambda_j^2}\frac{2\lambda_j^2 + k_j}{k_j} = \sigma^2\left[\frac{2}{k_j} + \frac{1}{\lambda_j^2}\right].$$

If λ_j^2 is small, almost any value of k will give an improvement over least squares. If λ_j^2 is large, only very small values of k will give an improvement.

Note that, since the $\hat{\gamma}_j$s are unbiased, the $\tilde{\gamma}_j$s will, in general, be biased estimates.

15.5 Penalized Estimation

In applications of linear model theory to nonparametric regression (cf. Subsection 6.2.1 and Christensen, 2001, Chapter 7) and in applications where p is large relative to n, it is not uncommon to replace least squares estimates with estimates that incorporate a penalty on the regression coefficients. These estimates are determined by adding a nonnegative *penalty function* $p(\beta)$ to the least squares criterion function, i.e., they minimize

$$(Y - X\beta)'(Y - X\beta) + k\,p(\beta), \tag{1}$$

where $k \ge 0$ is a *tuning parameter*. Obviously, if $k = 0$, the estimates are least squares estimates. Typical penalty functions are minimized at the vector $\beta = 0$, so

as k gets large, the penalty function dominates the minimization and the procedure, in some fashion, shrinks the least squares estimates towards 0.

Because of the nature of commonly used penalty functions, it is often suggested that the model matrix X should be standardized as in (15.1.6). If the height of my doghouse is a predictor variable, the appropriate regression coefficient depends a great deal on whether the height is measured in kilometers or microns. For a penalty function to be meaningful, it needs to be defined on an appropriate scale in each dimension.

As in Subsection 6.2.1, when using a basis function approach for nonparametric regression of y on a scalar predictor x, the linear model is

$$y_i = \beta_0 + \beta_1\phi_1(x_i) + \cdots + \beta_{p-1}\phi_{p-1}(x_i) + \varepsilon_i$$

for known functions ϕ_j. The basis functions $\phi_j(x)$ are frequently subjected to some form of standardization when being defined, thus obviating a strong need for further standardization. For example, if $0 \le x_i \le 1$ and $\phi_j(x) = \cos(\pi jx)$, there is little need to standardize X further. When using simple polynomials $\phi_j(x) = x^j$, the model matrix should be standardized. When using the corresponding *Legendre polynomials*, it need not be.

Penalty functions are often used to avoid *overfitting* a model. For example, with $\phi_j(x) = \cos(\pi jx)$, when j is large the cosine functions oscillate very rapidly, leading to *nonsmooth* or *noisy* behavior. Typically, with basis function approaches to nonparametric regression, large j is indicative of more noisy behavior. We want to allow noisy behavior if the data require it, but we prefer smooth functions if they seem reasonable. It therefore makes sense to place larger penalties on the regression coefficients for large j. In other words, for large values of j we shrink the least squares estimate $\hat{\beta}_j$ towards 0 more than when j is small. Such penalty functions make it possible to use numbers of parameters that are similar to the number of observations without overfitting the model. See Christensen (1996, Section 7.11) for some plots of overfitted polynomials.

Classical ridge regression provides one application of penalty functions. It amounts to using the penalty function

$$p_R(\beta) = \sum_{j=0}^{p-1} \beta_j^2 = \beta'\beta.$$

Note that this application does not penalize coefficients differently based on j. It is easy to see that the function $(Y - X\beta)'(Y - X\beta) + k\beta'\beta$ is the least squares criterion function for the model

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ \sqrt{k}I \end{bmatrix} \beta + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix}.$$

Fitting this augmented model by least squares yields the ridge regression estimate $\hat{\beta}_R = (X'X + kI)^{-1}X'Y$.

Generalized ridge regression takes the form of a penalty

$$p_{GR}(\beta) = \beta' Q \beta,$$

where Q is a nonnegative definite matrix. Most often Q is diagonal so that $p_{GR}(\beta) = \sum_{j=0}^{p-1} q_{jj}\beta_j^2$. We can minimize

$$(Y - X\beta)'(Y - X\beta) + k\beta' Q \beta \tag{2}$$

using the least squares fit to

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ \sqrt{k}\tilde{Q} \end{bmatrix} \beta + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix},$$

where $Q = \tilde{Q}'\tilde{Q}$. Alternatively, when Q is nonsingular, finding the generalized least squares estimate for the model

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ I \end{bmatrix} \beta + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix}, \quad \mathrm{E}\begin{bmatrix} e \\ \tilde{e} \end{bmatrix} = \begin{bmatrix} 0_{n\times 1} \\ 0_{p\times 1} \end{bmatrix}, \quad \mathrm{Cov}\begin{bmatrix} e \\ \tilde{e} \end{bmatrix} = \sigma^2 \begin{bmatrix} I_n & 0 \\ 0 & (1/k)Q^{-1} \end{bmatrix}$$

also leads to minimizing the criterion function (2) and gives the generalized ridge estimate $\hat{\beta}_{GR} = (X'X + kQ)^{-1}X'Y$.

Green and Silverman (1994, Section 3.6) discuss different choices for Q. Those choices generally follow the pattern of more shrinkage for β_js that incorporate noisier behavior into the model. In particular, they often determine $Q = D(q_{jj})$ with q_{jj} increasing in j. In Section 3, our generalized ridge estimates used $k_j \equiv kq_{jj}$.

Currently, a very popular penalty function is Tibshirani's (1996) *lasso (least absolute shrinkage and selection operator)*,

$$p_L(\beta) = \sum_{j=0}^{p-1} |\beta_j|. \tag{3}$$

Because this penalty function is not a quadratic form in β, unlike ridge regression the estimate cannot be obtained by fitting an augmented linear model using least squares. Lasso estimates can be computed efficiently for a variety of values k using a modification of the LARS algorithm of Efron et al. (2004).

As the actual name (not the acronym) suggests, one of the benefits of the lasso penalty is that it automates variable selection. Rather than gradually shrinking all regression coefficients towards 0 like ridge regression, lasso can make some of the regression coefficients collapse to 0.

Just as, in Section 3, we used canonical regression to explore ridge estimation, we can also use canonical regression to explicate the behavior of lasso regression. Applying the lasso criterion to canonical regression we need to minimize

$$\sum_{j=1}^{p} \left[(y_{*j} - \lambda_j \gamma_j)^2 + k|\gamma_j| \right].$$

Because of the simple structure of canonical regression, the lasso criterion acts independently on each coefficient. Without loss of generality, assume $y_{*j} > 0$. Clearly, $\gamma_j < 0$ will not minimize the criterion, because $\gamma_j = 0$ will be better. We therefore want to minimize $(y_{*j} - \lambda_j\gamma_j)^2 + k\gamma_j$ for $\gamma_j \geq 0$.

With $\hat{\gamma}_j = y_{*j}/\lambda_j$ being the least squares estimate, a little bit of work shows that the derivative of the criterion function with respect to γ_j is zero at $\gamma_j = \hat{\gamma}_j - k/2\lambda_j^2$. However, if $\hat{\gamma}_j < k/2\lambda_j^2$, the critical point is outside the domain of the function, so the minimum must occur at the boundary. Therefore, the lasso estimate is

$$\hat{\gamma}_{Lj} = \begin{cases} \hat{\gamma}_j - k/2\lambda_j^2, & \text{if } \hat{\gamma}_j \geq k/2\lambda_j^2 \\ 0, & \text{if } |\hat{\gamma}_j| < k/2\lambda_j^2 \\ \hat{\gamma}_j + k/2\lambda_j^2, & \text{if } \hat{\gamma}_j \leq -k/2\lambda_j^2 \end{cases}.$$

Clearly, if the least squares estimate is too small, the lasso estimate is zero and the variable is effectively removed from the model.

The lasso penalty (3) treats every coefficient the same. An obvious modification of lasso to penalize coefficients at different rates has

$$p_{GL}(\beta) = \sum_{j=0}^{p-1} q_{jj}|\beta_j|$$

with q_{jj} often increasing in j.

Exercise 15.7 Using the standardization for the canonical model of Exercise 15.6, find the lasso estimates.

15.5.1 Bayesian Connections

Another way to address the issue of penalizing regression coefficients is through the Bayesian methods illustrated in Section 2.9. The likelihood function for normal data is

$$L(\beta, \sigma^2) = (2\pi)^{-n/2}[\det(\sigma^2 I)]^{-1/2} \exp\left[-(Y - X\beta)'(Y - X\beta)/2\sigma^2\right].$$

We take a prior density of the form $\pi(\beta, \sigma^2) \equiv \pi_1(\beta|\sigma^2)\pi_2(\sigma^2)$ where the conditional density of β given σ^2 is written as

$$\pi_1(\beta|\sigma^2) = h(\sigma^2) \exp\left[-k\, p(\beta)/2\sigma^2\right],$$

with $p(\beta)$ once again being the penalty function. The posterior is proportional to the likelihood times the prior, so it has the form

$$\pi(\beta, \sigma^2|Y) \propto \left(\sigma^2\right)^{-n/2} \pi_2(\sigma^2)h(\sigma^2) \times$$

$$\exp\left\{-\frac{1}{2\sigma^2}\left[(Y-X\beta)'(Y-X\beta)+k\,p(\beta)\right]\right\}.$$

The value $\hat{\beta}$ that minimizes (1) is also the posterior mode, regardless of the value of σ^2. A further prior can be placed on k.

Ridge regression amounts to placing a normal prior on β and using the one number that is the posterior mean, median, and mode as an estimate of β. In particular, the generalized ridge estimate devolves from the prior distribution

$$\beta|\sigma^2 \sim N\left(0, \frac{\sigma^2}{k}Q^{-1}\right).$$

When Q is diagonal, large penalties clearly correspond to small prior variances, i.e., strong prior beliefs that β_j is near the prior mean of 0.

Lasso regression can be constructed as the posterior mode of β by putting a Laplace (double exponential) prior on β. Given the discontinuous nature of the lasso minimization problem, it is not surprising that technical difficulties can arise. Park and Casella (2008) provide a good discussion, but use a slightly different prior.

Another interesting Bayesian method for avoiding overfitting is *thresholding*, see Smith and Kohn (1996), Clyde and George (2004), or Christensen et al. (2010, Section 15.2). The idea is to put positive prior probability on each regression coefficient being 0, so there will be positive, although perhaps very small, posterior probability of it being 0. For example, with a basis function model a form of generalized ridge regression corresponds to independent normal priors on β_j with mean 0 and a variance σ^2/kq_{jj} decreasing in j. Instead, we might write

$$\beta_j = \delta_j \beta_j^*$$

with $\beta_j^* \sim N(0, \sigma^2/k)$ independent of $\delta_j \sim \text{Bern}(2^{-j})$, i.e., $\Pr[\delta_j = 1] = 2^{-j}$. This obviously makes it harder, but not impossible, for β_j to be nonzero as j increases.

15.6 Orthogonal Regression

Suppose we have bivariate data (x_i, y_i) and want to fit a line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Rather than using least squares (that minimizes vertical distances to the line) we will do orthogonal regression that minimizes perpendicular distances to the line. We will run this line through the point $(\bar{x}_\cdot, \bar{y}_\cdot)$ so we need only worry about the slope of the line.

Transform $(x_i, y_i)'$ into $v_i \equiv (x_i - \bar{x}_\cdot, y_i - \bar{y}_\cdot)'$. We want to find a vector a, or more properly a one-dimensional column space $C(a)$, that minimizes the squared perpendicular distances between the data v_i and the regression line that consists of all multiples of the vector a. We need to take v_i and project it onto the line, that is, project it into $C(a)$. The projection is $M_a v_i$. The squared perpendicular distance between the data and the line is $\|v_i - M_a v_i\|^2 = \|(I - M_a)v_i\|^2 = v_i'(I - M_a)v_i$. It will

become important later to recognize this as the squared length of the perpendicular projection of v_i onto the orthogonal complement of the regression surface vector space. In any case, we need to pick the line so as to minimize the sum of all these squared distances, i.e., so that $\sum_{i=1}^n v_i'(I-M_a)v_i$ is minimized. However, if a minimizes $\sum_{i=1}^n v_i'(I-M_a)v_i$, it maximizes $\sum_{i=1}^n v_i'M_a v_i$. Note also that

$$\begin{aligned}\max_a \sum_{i=1}^n v_i'M_a v_i &= \max_a \sum_{i=1}^n v_i'a(a'a)^{-1}a'v_i \\ &= \max_a \frac{1}{a'a}\sum_{i=1}^n a'v_i v_i'a \\ &= \max_a \frac{1}{a'a}a'\left(\sum_{i=1}^n v_i v_i'\right)a \\ &= \max_a \frac{n-1}{a'a}a'Sa,\end{aligned}$$

where S is the sample covariance matrix of the data.

It is enough to find $\hat{a}$ such that

$$\frac{\hat{a}'S\hat{a}}{\hat{a}'\hat{a}} = \max_a \frac{a'Sa}{a'a}.$$

It is well-known (see Christensen, 2001, Proposition 2.3.4) that this max is achieved by eigenvectors associated with the largest eigenvalue of S. In particular, if we pick a maximizing eigenvector $\hat{a}$ to be $\hat{a}' = (1, \hat{\beta})$, then $\hat{\beta}$ is the orthogonal regression slope estimate. The estimated line becomes

$$\hat{y} = \bar{y}_\cdot + \hat{\beta}(x - \bar{x}_\cdot).$$

Technically, this occurs because for the fitted values to fall on the regression line, they must determine a multiple of the eigenvector, i.e., $\hat{y}$ is *defined* so that

$$\begin{bmatrix} x-\bar{x}_\cdot \\ \hat{y}-\bar{y}_\cdot \end{bmatrix} \equiv (x-\bar{x}_\cdot)\begin{bmatrix} 1 \\ \hat{\beta} \end{bmatrix}.$$

Normally, we would find the eigenvector corresponding to the largest eigenvalue computationally, but our problem can be solved analytically. To find the maximizing eigenvector we need to solve the matrix equation

$$\begin{bmatrix} s_x^2-\lambda & s_{xy} \\ s_{xy} & s_y^2-\lambda \end{bmatrix}\begin{bmatrix} 1 \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

This simplifies to the set of equations

$$\lambda = s_x^2 + s_{xy}\hat{\beta} \tag{1}$$

and

$$s_{xy} + (s_y^2 - \lambda)\hat{\beta} = 0. \tag{2}$$

Substituting λ from (1) into (2),

$$s_{xy} + (s_y^2 - s_x^2)\hat{\beta} - s_{xy}\hat{\beta}^2 = 0. \tag{3}$$

Applying the quadratic formula gives

$$\hat{\beta} = \frac{-(s_y^2 - s_x^2) \pm \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2}}{-2s_{xy}} = \frac{(s_y^2 - s_x^2) \pm \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2}}{2s_{xy}}.$$

Substituting $\hat{\beta}$ back into (1), the larger of the two values of λ corresponds to

$$\hat{\beta} = \frac{(s_y^2 - s_x^2) + \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2}}{2s_{xy}},$$

which is our slope estimate.

If you find the eigenvalues and eigenvectors computationally, remember that eigenvectors for a given eigenvalue (essentially) form a vector space. (Eigenvectors are not allowed to be 0.) Thus, a reported eigenvector $(a_1, a_2)'$ for the largest eigenvalue also determines the eigenvector we want, $(1, a_2/a_1)'$.

It turns out that we can also get least squares estimates by modifying these matrix equations. Consider

$$\begin{bmatrix} s_x^2 + [s_y^2 - s_{xy}^2/s_x^2] - \lambda & s_{xy} \\ s_{xy} & s_y^2 - \lambda \end{bmatrix} \begin{bmatrix} 1 \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

or equivalently,

$$\begin{bmatrix} s_x^2 + [s_y^2 - s_{xy}^2/s_x^2] & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix} \begin{bmatrix} 1 \\ \hat{\beta} \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ \hat{\beta} \end{bmatrix}.$$

Rather than solving this for $\hat{\beta}$, simply observe that one solution is $\lambda = s_x^2 + s_y^2$ and $(1, \hat{\beta})' = (1, s_{xy}/s_x^2)'$. Note that $[s_y^2 - s_{xy}^2/s_x^2] = [(n-2)/(n-1)]MSE$ where MSE is the mean squared error from the least squares fit of $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

To generalize the orthogonal regression procedure to multiple regression we need a more oblique approach. With data (x_i', y_i) consisting of p-dimensional vectors, a linear regression surface corresponds to a $(p-1)$-dimensional hyperplane so that if we specify a $(p-1)$ vector of predictor variables x, we know the fitted value $\hat{y}$ because it is the point corresponding to x on the hyperplane. By considering data $v_i' \equiv (x_i' - \bar{x}_\cdot', y_i - \bar{y}_\cdot)$, the regression surface goes through the origin and becomes a $(p-1)$-dimensional vector space. Rather than specify the $(p-1)$-dimensional space, it is easier to find the orthogonal complement which is one-dimensional. Writing the one-dimensional space as $C(a)$ for a p vector a, the regression surface will be $C(a)^\perp$. The squared distances from the v_is to the $(p-1)$-dimensional regression space are now $v_i' M_a v_i$, so we want to minimize $\sum_{i=1}^n v_i' M_a v_i$. Similar to our earlier argument,

$$\min_a \sum_{i=1}^{n} v_i' M_a v_i = \min_a \frac{n-1}{a'a} a' S a,$$

with S being the sample covariance matrix of the complete data. However, now we obtain our fitted values differently. The minimum is achieved by choosing the eigenvector $\hat{a} = (\hat{\beta}', -1)'$ corresponding to the smallest eigenvalue. Our fitted values $\hat{y}$ now must determine vectors that are orthogonal to this eigenvector, so they satisfy

$$\begin{bmatrix} \hat{\beta}' & -1 \end{bmatrix} \begin{bmatrix} x - \bar{x}. \\ \hat{y} - \bar{y}. \end{bmatrix} = 0$$

or

$$\hat{y} = \bar{y}. + \hat{\beta}'(x - \bar{x}.).$$

As illustrated earlier for $p = 2$, any eigenvector (for the smallest eigenvalue) reported by a computer program is easily rescaled to $(\hat{\beta}', -1)'$

Finally, this approach better give the same answers for simple linear regression that we got from our first procedure. It is not difficult to see that

$$\begin{bmatrix} s_x^2 - \lambda & s_{xy} \\ s_{xy} & s_y^2 - \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

once again leads to equation (3) but now

$$\hat{\beta} = \frac{(s_y^2 - s_x^2) + \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2}}{2s_{xy}}$$

corresponds to the smallest eigenvalue. If you think about it, with distinct eigenvalues, the eigenvector $(1, \hat{\beta})'$ corresponding to the largest eigenvalue must be orthogonal to any eigenvector for the smallest eigenvalue, and $(\hat{\beta}, -1)'$ is orthogonal to $(1, \hat{\beta})'$.

Like least squares, this procedure is a geometric justification for an estimate, not a statistical justification. In Chapter 2 we showed that least squares estimates have statistical optimality properties like being BLUEs, MVUEs, and MLEs. The ideas used here are similar to those needed for looking at the separating hyperplanes used in support vector machines, see Moguerza and Muñoz (2006) or Zhu (2008).