Chapter 11

Multiple regression: matrix formulation

In this chapter we use matrices to write regression models. Properties of matrices are reviewed in Appendix A. The economy of notation achieved through using matrices allows us to arrive at some interesting new insights and to derive several of the important properties of regression analysis.

11.1 Random vectors

In this section we discuss vectors and matrices that are made up of random variables rather than just numbers. For simplicity, we focus our discussion on vectors that contain 3 rows, but the results are completely general.

Let y_1, y_2, and y_3 be random variables. From these, we can construct a 3×1 random vector, say

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

The expected value of the random vector is just the vector of expected values of the random variables. For the random variables write $\mathrm{E}(y_i) = \mu_i$, then

$$\mathrm{E}(Y) \equiv \begin{bmatrix} \mathrm{E}(y_1) \\ \mathrm{E}(y_2) \\ \mathrm{E}(y_3) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \equiv \mu.$$

In other words, expectation of a random vector is performed elementwise. In fact, the expected value of any random matrix (a matrix consisting of random variables) is the matrix made up of the expected values of the elements in the random matrix. Thus if w_{ij}, $i = 1,2,3$, $j = 1,2$ is a collection of random variables and we write

$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{33} \end{bmatrix},$$

then

$$\mathrm{E}(W) \equiv \begin{bmatrix} \mathrm{E}(w_{11}) & \mathrm{E}(w_{12}) \\ \mathrm{E}(w_{21}) & \mathrm{E}(w_{22}) \\ \mathrm{E}(w_{31}) & \mathrm{E}(w_{33}) \end{bmatrix}.$$

We also need a concept for random vectors that is analogous to the variance of a random variable. This is the *covariance matrix*, sometimes called the *dispersion matrix*, the *variance matrix*, or the *variance-covariance matrix*. The covariance matrix is simply a matrix consisting of all the variances and covariances associated with the vector Y. Write

$$\mathrm{Var}(y_i) = \mathrm{E}(y_i - \mu_i)^2 \equiv \sigma_{ii}$$

and

$$\mathrm{Cov}(y_i, y_j) = \mathrm{E}[(y_i - \mu_i)(y_j - \mu_j)] \equiv \sigma_{ij}.$$

Two subscripts are used on σ_{ii} to indicate that it is the variance of y_i *rather than* writing $\text{Var}(y_i) = \sigma_i^2$.

The covariance matrix of our 3×1 vector Y is

$$\text{Cov}(Y) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}.$$

When Y is 3×1, the covariance matrix is 3×3. If Y were 20×1, $\text{Cov}(Y)$ would be 20×20. The covariance matrix is always symmetric because $\sigma_{ij} = \sigma_{ji}$ for any i, j. The variances of the individual random variables lie on the diagonal that runs from the top left to the bottom right. The covariances lie off the diagonal.

In general, if Y is an $r \times 1$ random vector and $\text{E}(Y) = \mu$, then $\text{Cov}(Y) = \text{E}[(Y-\mu)(Y-\mu)']$. In other words, $\text{Cov}(Y)$ is the expected value of the random matrix $(Y-\mu)(Y-\mu)'$.

11.2 Matrix formulation of regression models

Simple linear regression in matrix form

The usual model for simple linear regression is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n, \tag{11.2.1}$$

$\text{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$, and $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. In matrix terms this can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times 2} \quad \beta_{2\times 1} + e_{n\times 1}$$

Multiplying and adding the matrices on the right-hand side gives

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix}.$$

These two vectors are equal if and only if the corresponding elements are equal, which occurs if and only if model (11.2.1) holds. The conditions on the ε_is translate into matrix terms as

$$\text{E}(e) = 0$$

where 0 is the $n \times 1$ matrix containing all zeros and

$$\text{Cov}(e) = \sigma^2 I$$

where I is the $n \times n$ identity matrix. By definition, the covariance matrix $\text{Cov}(e)$ has the variances of the ε_is down the diagonal. The variance of each individual ε_i is σ^2, so all the diagonal elements of $\text{Cov}(e)$ are σ^2, just as in $\sigma^2 I$. The covariance matrix $\text{Cov}(e)$ has the covariances of distinct ε_is as its off-diagonal elements. The covariances of distinct ε_is are all 0, so all the off-diagonal elements of $\text{Cov}(e)$ are zero, just as in $\sigma^2 I$.

Table 11.1: *Weights for various heights.*

Ht.	Wt.	Ht.	Wt.
65	120	63	110
65	140	63	135
65	130	63	120
65	135	72	170
66	150	72	185
66	135	72	160

EXAMPLE 11.2.1. Height and weight data are given in Table 11.1 for 12 individuals. In matrix terms, the SLR model for regressing weights (y) on heights (x) is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 65 \\ 1 & 65 \\ 1 & 65 \\ 1 & 65 \\ 1 & 66 \\ 1 & 66 \\ 1 & 63 \\ 1 & 63 \\ 1 & 63 \\ 1 & 72 \\ 1 & 72 \\ 1 & 72 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

The observed data for this example are

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} 120 \\ 140 \\ 130 \\ 135 \\ 150 \\ 135 \\ 110 \\ 135 \\ 120 \\ 170 \\ 185 \\ 160 \end{bmatrix}.$$

We could equally well rearrange the order of the observations to write

$$\begin{bmatrix} y_7 \\ y_8 \\ y_9 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_4 \\ y_6 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 63 \\ 1 & 63 \\ 1 & 63 \\ 1 & 65 \\ 1 & 65 \\ 1 & 65 \\ 1 & 65 \\ 1 & 66 \\ 1 & 66 \\ 1 & 72 \\ 1 & 72 \\ 1 & 72 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

in which the x_i values are ordered from smallest to largest. □

The general linear model

The general linear model is a generalization of the matrix form for the simple linear regression model. The general linear model is

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

Y is an $n \times 1$ vector of observable random variables. X is an $n \times p$ matrix of known constants. β is a $p \times 1$ vector of unknown (regression) parameters. e is an $n \times 1$ vector of unobservable random errors. It will be assumed that $n \geq p$. Regression is any general linear model where the rank of X is p.

EXAMPLE 11.2.2. *Multiple regression*
In non-matrix form, the multiple regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{11.2.2}$$

where

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \; i \neq j.$$

In matrix terms this can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times p} \beta_{p\times 1} + e_{n\times 1}$$

Multiplying and adding the right-hand side gives

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{p-1} x_{1,p-1} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_{p-1} x_{2,p-1} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{p-1} x_{n,p-1} + \varepsilon_n \end{bmatrix},$$

which holds if and only if (11.2.2) holds. The conditions on the ε_is translate into

$$\mathrm{E}(e) = 0,$$

where 0 is the $n \times 1$ matrix consisting of all zeros, and

$$\mathrm{Cov}(e) = \sigma^2 I,$$

where I is the $n \times n$ identity matrix. □

EXAMPLE 11.2.3. In Example 11.2.1 we illustrated the matrix form of a SLR using the data on heights and weights. We now illustrate some of the models from Chapter 8 applied to these data.

The cubic model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \tag{11.2.3}$$

is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 65 & 65^2 & 65^3 \\ 1 & 65 & 65^2 & 65^3 \\ 1 & 65 & 65^2 & 65^3 \\ 1 & 65 & 65^2 & 65^3 \\ 1 & 66 & 66^2 & 66^3 \\ 1 & 66 & 66^2 & 66^3 \\ 1 & 63 & 63^2 & 63^3 \\ 1 & 63 & 63^2 & 63^3 \\ 1 & 63 & 63^2 & 63^3 \\ 1 & 72 & 72^2 & 72^3 \\ 1 & 72 & 72^2 & 72^3 \\ 1 & 72 & 72^2 & 72^3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

Some of the numbers in X are getting quite large, i.e., $65^3 = 274625$. The model has better numerical properties if we compute $\bar{x}_\cdot = 69.4166\overline{6}$ and replace model (11.2.3) with the equivalent model

$$y_i = \gamma_0 + \gamma_1(x_i - \bar{x}_\cdot) + \gamma_2(x_i - \bar{x}_\cdot)^2 + \beta_3(x_i - \bar{x}_\cdot)^3 + \varepsilon_i$$

and its matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & (65-\bar{x}_\cdot) & (65-\bar{x}_\cdot)^2 & (65-\bar{x}_\cdot)^3 \\ 1 & (65-\bar{x}_\cdot) & (65-\bar{x}_\cdot)^2 & (65-\bar{x}_\cdot)^3 \\ 1 & (65-\bar{x}_\cdot) & (65-\bar{x}_\cdot)^2 & (65-\bar{x}_\cdot)^3 \\ 1 & (65-\bar{x}_\cdot) & (65-\bar{x}_\cdot)^2 & (65-\bar{x}_\cdot)^3 \\ 1 & (66-\bar{x}_\cdot) & (66-\bar{x}_\cdot)^2 & (66-\bar{x}_\cdot)^3 \\ 1 & (66-\bar{x}_\cdot) & (66-\bar{x}_\cdot)^2 & (66-\bar{x}_\cdot)^3 \\ 1 & (63-\bar{x}_\cdot) & (63-\bar{x}_\cdot)^2 & (63-\bar{x}_\cdot)^3 \\ 1 & (63-\bar{x}_\cdot) & (63-\bar{x}_\cdot)^2 & (63-\bar{x}_\cdot)^3 \\ 1 & (63-\bar{x}_\cdot) & (63-\bar{x}_\cdot)^2 & (63-\bar{x}_\cdot)^3 \\ 1 & (72-\bar{x}_\cdot) & (72-\bar{x}_\cdot)^2 & (72-\bar{x}_\cdot)^3 \\ 1 & (72-\bar{x}_\cdot) & (72-\bar{x}_\cdot)^2 & (72-\bar{x}_\cdot)^3 \\ 1 & (72-\bar{x}_\cdot) & (72-\bar{x}_\cdot)^2 & (72-\bar{x}_\cdot)^3 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

This third degree polynomial is the largest polynomial that we can fit to these data. Two points determine a line, three points determine a quadratic, and with only four district x values in the data, we cannot fit a model greater than a cubic.

Define $\tilde{x} = (x - 63)/9$ so that

$$(x_1, \ldots, x_{12}) = (65, 65, 65, 65, 66, 66, 63, 63, 63, 72, 72, 72)$$

transforms to

$$(\tilde{x}_1, \ldots, \tilde{x}_{12}) = (2/9, 2/9, 2/9, 2/9, 1/3, 1/3, 0, 0, 0, 1, 1, 1).$$

The basis function model based on cosines

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \cos(\pi \tilde{x}_i) + \beta_3 \cos(\pi 2 \tilde{x}_i) + \varepsilon_i$$

becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 65 & \cos(2\pi/9) & \cos(4\pi/9) \\ 1 & 65 & \cos(2\pi/9) & \cos(4\pi/9) \\ 1 & 65 & \cos(2\pi/9) & \cos(4\pi/9) \\ 1 & 65 & \cos(2\pi/9) & \cos(4\pi/9) \\ 1 & 66 & \cos(\pi/3) & \cos(2\pi/3) \\ 1 & 66 & \cos(\pi/3) & \cos(2\pi/3) \\ 1 & 63 & \cos(0) & \cos(0) \\ 1 & 63 & \cos(0) & \cos(0) \\ 1 & 63 & \cos(0) & \cos(0) \\ 1 & 72 & \cos(\pi) & \cos(2\pi) \\ 1 & 72 & \cos(\pi) & \cos(2\pi) \\ 1 & 72 & \cos(\pi) & \cos(2\pi) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

The "Haar wavelet" model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{[0,.50)}(\tilde{x}_i) + \beta_3 I_{[.5,1]}(\tilde{x}_i) + \varepsilon_i$$

becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 65 & 1 & 0 \\ 1 & 65 & 1 & 0 \\ 1 & 65 & 1 & 0 \\ 1 & 65 & 1 & 0 \\ 1 & 66 & 1 & 0 \\ 1 & 66 & 1 & 0 \\ 1 & 63 & 1 & 0 \\ 1 & 63 & 1 & 0 \\ 1 & 63 & 1 & 0 \\ 1 & 72 & 0 & 1 \\ 1 & 72 & 0 & 1 \\ 1 & 72 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

Notice that the last two columns of the X matrix add up to a column of 1s, like the first column. This causes the rank of the 12×4 model matrix X to be only 3, so the model is not a regression model. Dropping either of the last two columns (or the first column) does not change the model in any meaningful way but makes the model a regression.

If we partition the SLR model into points below 65.5 and above 65.5, the matrix model becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 0 & 0 & 1 & 66 \\ 0 & 0 & 1 & 66 \\ 1 & 63 & 0 & 0 \\ 1 & 63 & 0 & 0 \\ 1 & 63 & 0 & 0 \\ 0 & 0 & 1 & 72 \\ 0 & 0 & 1 & 72 \\ 0 & 0 & 1 & 72 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

Alternatively, we could rewrite the model as

$$\begin{bmatrix} y_7 \\ y_8 \\ y_9 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_4 \\ y_6 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 63 & 0 & 0 \\ 1 & 63 & 0 & 0 \\ 1 & 63 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 0 & 0 & 1 & 66 \\ 0 & 0 & 1 & 66 \\ 0 & 0 & 1 & 72 \\ 0 & 0 & 1 & 72 \\ 0 & 0 & 1 & 72 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

This makes is a bit clearer that we are fitting a SLR to the points with small x values and a separate SLR to cases with large x values. The pattern of 0s in the X matrix ensure that the small x values only involve the intercept and slope parameters β_0 and β_1 for the line on the first partition set and that the large x values only involve the intercept and slope parameters β_2 and β_3 for the line on the second partition set.

Fitting this model can also be accomplished by fitting the model

$$\begin{bmatrix} y_7 \\ y_8 \\ y_9 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_4 \\ y_6 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 63 & 0 & 0 \\ 1 & 63 & 0 & 0 \\ 1 & 63 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 65 & 0 & 0 \\ 1 & 66 & 1 & 66 \\ 1 & 66 & 1 & 66 \\ 1 & 72 & 1 & 72 \\ 1 & 72 & 1 & 72 \\ 1 & 72 & 1 & 72 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma_0 \\ \gamma_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

Here we have changed the first two columns to make them agree with the SLR of Example 11.2.1. However, notice that if we subtract the third column from the first column we get the first column of the previous version. Similarly,if we subtract the fourth column from the second column we get the second column of the previous version. This model has intercept and slope parameters β_0 and β_1 for the first partition and intercept and slope parameters $(\beta_0 + \gamma_0)$ and $(\beta_1 + \gamma_1)$ for the second partition.

Because of the particular structure of these data with 12 observations but only four distinct values of x, except for the Haar wavelet model, all of these models are equivalent to one another and

all of them are equivalent to a model with the matrix formulation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}.$$

The models are equivalent in that they all give the same fitted values, residuals, and degrees of freedom for error. We will see in the next chapter that this last matrix model has the form of a one-way analysis of variance model. □

Other models to be discussed later such as analysis of variance and analysis of covariance models can also be written as general linear models.

11.3 Least squares estimation of regression parameters

The regression estimates given by standard computer programs are least squares estimates. For simple linear regression, the least squares estimates are the values of β_0 and β_1 that minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{11.3.1}$$

For multiple regression, the least squares estimates of the β_js minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_{p-1} x_{i,p-1})^2.$$

In matrix terms these can both be written as minimizing

$$(Y - X\beta)'(Y - X\beta). \tag{11.3.2}$$

The form in (11.3.2) is just the sum of the squares of the elements in the vector $(Y - X\beta)$. See also Exercise 11.7.1.

We now give the general form for the least squares estimate of β in regression problems.

Proposition 11.3.1. If $r(X) = p$, then $\hat{\beta} = (X'X)^{-1} X'Y$ is the least squares estimate of β.

PROOF: *The proof is optional material.*

Note that $(X'X)^{-1}$ exists only because in a regression problem the rank of X is p. The proof stems from rewriting the function to be minimized.

$$\begin{aligned} (Y - X\beta)'(Y - X\beta) &= \left(Y - X\hat{\beta} + X\hat{\beta} - X\beta\right)' \left(Y - X\hat{\beta} + X\hat{\beta} - X\beta\right) \\ &= \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) + \left(Y - X\hat{\beta}\right)' \left(X\hat{\beta} - X\beta\right) \\ &\quad + \left(X\hat{\beta} - X\beta\right)' \left(Y - X\hat{\beta}\right) + \left(X\hat{\beta} - X\beta\right)' \left(X\hat{\beta} - X\beta\right). \end{aligned} \tag{11.3.3}$$

Consider one of the two cross-product terms from the last expression, say $\left(X\hat{\beta}-X\beta\right)'\left(Y-X\hat{\beta}\right)$. Using the definition of $\hat{\beta}$ given in the proposition,

$$\begin{aligned}\left(X\hat{\beta}-X\beta\right)'\left(Y-X\hat{\beta}\right) &= \left[X\left(\hat{\beta}-\beta\right)\right]'\left(Y-X\hat{\beta}\right)\\ &= \left(\hat{\beta}-\beta\right)'X'\left(Y-X\left(X'X\right)^{-1}X'Y\right)\\ &= \left(\hat{\beta}-\beta\right)'X'\left(I-X\left(X'X\right)^{-1}X'\right)Y\end{aligned}$$

but

$$X'\left(I-X\left(X'X\right)^{-1}X'\right)=X'-\left(X'X\right)\left(X'X\right)^{-1}X'=X'-X'=0.$$

Thus

$$\left(X\hat{\beta}-X\beta\right)'\left(Y-X\hat{\beta}\right)=0$$

and similarly

$$\left(Y-X\hat{\beta}\right)'\left(X\hat{\beta}-X\beta\right)=0.$$

Eliminating the two middle terms in (11.3.3) gives

$$(Y-X\beta)'(Y-X\beta)=\left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)+\left(X\hat{\beta}-X\beta\right)'\left(X\hat{\beta}-X\beta\right).$$

This form is easily minimized. The first of the terms on the right-hand side does not depend on β, so the β that minimizes $(Y-X\beta)'(Y-X\beta)$ is the β that minimizes the second term $\left(X\hat{\beta}-X\beta\right)'\left(X\hat{\beta}-X\beta\right)$. The second term is non-negative because it is the sum of squares of the elements in the vector $X\hat{\beta}-X\beta$ and it is minimized by making it zero. This is accomplished by choosing $\beta=\hat{\beta}$. □

EXAMPLE 11.3.2. *Simple linear regression*

We now show that Proposition 11.3.1 gives the usual estimates for simple linear regression. Readers should refamiliarize themselves with the results in Section 6.10. They should also be warned that the algebra in the first half of the example is a bit more sophisticated than that used elsewhere in this book.

Assume the model

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i \quad i=1,\ldots,n.$$

and write

$$X=\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}$$

so

$$X'X=\begin{bmatrix}n & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{bmatrix}.$$

Inverting this matrix gives

$$\left(X'X\right)^{-1}=\frac{1}{n\sum_{i=1}^n x_i^2-\left(\sum_{i=1}^n x_i\right)^2}\begin{bmatrix}\sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i\\ -\sum_{i=1}^n x_i & n\end{bmatrix}.$$

The denominator in this term can be simplified by observing that

$$n\sum_{i=1}^n x_i^2-\left(\sum_{i=1}^n x_i\right)^2=n\left(\sum_{i=1}^n x_i^2-n\bar{x}_\cdot^2\right)=n\sum_{i=1}^n\left(x_i-\bar{x}_\cdot\right)^2.$$

Note also that

$$X'Y = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}.$$

Finally, we get

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1} X'Y \\
&= \frac{1}{n\sum_{i=1}^n (x_i - \bar{x}.)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ -\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n\sum_{i=1}^n x_i y_i \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x}.)^2} \begin{bmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}. \sum_{i=1}^n x_i y_i \\ (\sum_{i=1}^n x_i y_i) - n\bar{x}.\bar{y} \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x}.)^2} \begin{bmatrix} \bar{y}\sum_{i=1}^n x_i^2 - n\bar{x}_\cdot^2 \bar{y} - \left\{ \bar{x}. \left(\sum_{i=1}^n x_i y_i\right) - \left(n\bar{x}_\cdot^2 \bar{y}\right) \right\} \\ \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x}.)^2} \begin{bmatrix} \bar{y}\left(\sum_{i=1}^n x_i^2 - n\bar{x}_\cdot^2\right) - \bar{x}. \left(\sum_{i=1}^n x_i y_i - n\bar{x}.\bar{y}\right) \\ \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix} \\
&= \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x}. \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}.
\end{aligned}$$

As usual, the alternative regression model

$$y_i = \beta_{*0} + \beta_1 (x_i - \bar{x}.) + \varepsilon_i \quad i = 1, \ldots, n$$

is easier to work with. Write the model in matrix form as

$$Y = Z\beta_* + e$$

where

$$Z = \begin{bmatrix} 1 & (x_1 - \bar{x}.) \\ 1 & (x_2 - \bar{x}.) \\ \vdots & \vdots \\ 1 & (x_n - \bar{x}.) \end{bmatrix}$$

and

$$\beta_* = \begin{bmatrix} \beta_{*0} \\ \beta_1 \end{bmatrix}.$$

We need to compute $\hat{\beta}_* = (Z'Z)^{-1} Z'Y$. Observe that

$$Z'Z = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix},$$

$$(Z'Z)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & 1 \Big/ \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix},$$

$$Z'Y = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n (x_i - \bar{x}.) y_i \end{bmatrix},$$

and

$$\hat{\beta}_* = (Z'Z)^{-1} Z'Y = \begin{bmatrix} \bar{y} \\ \sum_{i=1}^n (x_i - \bar{x}.) y_i \Big/ \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{*0} \\ \hat{\beta}_1 \end{bmatrix}.$$

These are the usual estimates. □

Recall that least squares estimates have a number of other properties. If the errors are independent with mean zero, constant variance, and are normally distributed, the least squares estimates are maximum likelihood estimates and minimum variance unbiased estimates. If the errors are merely uncorrelated with mean zero and constant variance, the least squares estimates are best (minimum variance) linear unbiased estimates.

In multiple regression, simple algebraic expressions for the parameter estimates are not possible. The only nice equations for the estimates are the matrix equations.

We now find expected values and covariance matrices for the data Y and the least squares estimate $\hat{\beta}$. Two simple rules about expectations and covariance matrices can take one a long way in the theory of regression. These are matrix analogues of Proposition 1.2.11. In fact, to prove these matrix results, one really only needs Proposition 1.2.11, cf. Exercise 11.7.3.

Proposition 11.3.3. Let A be a fixed $r \times n$ matrix, let c be a fixed $r \times 1$ vector, and let Y be an $n \times 1$ random vector, then

1. $\mathrm{E}(AY + c) = A\,\mathrm{E}(Y) + c$

2. $\mathrm{Cov}(AY + c) = A\mathrm{Cov}(Y)A'$.

Applying these results allows us to find the expected value and covariance matrix for Y in a linear model. The linear model has $Y = X\beta + e$ where $X\beta$ is a fixed vector (even though β is unknown), $\mathrm{E}(e) = 0$, and $\mathrm{Cov}(e) = \sigma^2 I$. Applying the proposition gives

$$\mathrm{E}(Y) = \mathrm{E}(X\beta + e) = X\beta + \mathrm{E}(e) = X\beta + 0 = X\beta$$

and

$$\mathrm{Cov}(Y) = \mathrm{Cov}(e) = \sigma^2 I.$$

We can also find the expected value and covariance matrix of the least squares estimate $\hat{\beta}$. In particular, we show that $\hat{\beta}$ is an *unbiased* estimate of β by showing

$$\mathrm{E}\left(\hat{\beta}\right) = \mathrm{E}\left((X'X)^{-1}X'Y\right) = (X'X)^{-1}X'\mathrm{E}(Y) = (X'X)^{-1}X'X\beta = \beta.$$

To find variances and standard errors we need $\mathrm{Cov}\left(\hat{\beta}\right)$. To obtain this matrix, we use the rules in Proposition A.7.1. In particular, recall that the inverse of a symmetric matrix is symmetric and that $X'X$ is symmetric.

$$\begin{aligned}
\mathrm{Cov}\left(\hat{\beta}\right) &= \mathrm{Cov}\left[(X'X)^{-1}X'Y\right] \\
&= \left[(X'X)^{-1}X'\right]\mathrm{Cov}(Y)\left[(X'X)^{-1}X'\right]' \\
&= \left[(X'X)^{-1}X'\right]\mathrm{Cov}(Y)X\left[(X'X)^{-1}\right]' \\
&= (X'X)^{-1}X'\mathrm{Cov}(Y)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}.
\end{aligned}$$

EXAMPLE 11.3.2 CONTINUED. For simple linear regression the covariance matrix becomes

$$\mathrm{Cov}\left(\hat{\beta}\right) = \sigma^2(X'X)^{-1}$$

$$
\begin{aligned}
&= \sigma^2 \frac{1}{n\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \\
&= \sigma^2 \frac{1}{n\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 - n\bar{x}_\cdot^2 + n\bar{x}_\cdot^2 & -n\bar{x}_\cdot \\ -n\bar{x}_\cdot & n \end{bmatrix} \\
&= \sigma^2 \frac{1}{n\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} \begin{bmatrix} \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2 + n\bar{x}_\cdot^2 & -n\bar{x}_\cdot \\ -n\bar{x}_\cdot & n \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}_\cdot^2}{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} & \frac{-\bar{x}_\cdot}{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} \\ \frac{-\bar{x}_\cdot}{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x}_\cdot)^2} \end{bmatrix},
\end{aligned}
$$

which agrees with results given earlier for simple linear regression.

11.4 Inferential procedures

We begin by examining the analysis of variance table for the regression model (11.2.2). We then discuss tests, confidence intervals, and prediction intervals.

There are two frequently used forms of the ANOVA table:

Source	df	SS	MS
β_0	1	$n\bar{y}^2 \equiv C$	$n\bar{y}^2$
Regression	$p-1$	$\hat{\beta}'X'X\hat{\beta} - C$	$SSReg/(p-1)$
Error	$n-p$	$Y'Y - C - SSReg$	$SSE/(n-p)$
Total	n	$Y'Y$	

and the more often used form

Source	df	SS	MS
Regression	$p-1$	$\hat{\beta}'X'X\hat{\beta} - C$	$SSReg/(p-1)$
Error	$n-p$	$Y'Y - C - SSReg$	$SSE/(n-p)$
Total	$n-1$	$Y'Y - C$	

Note that $Y'Y = \sum_{i=1}^n y_i^2$, $C = n\bar{y}^2 = \left(\sum_{i=1}^n y_i\right)^2/n$, and $\hat{\beta}'X'X\hat{\beta} = \hat{\beta}'X'Y$. The difference between the two tables is that the first includes a line for the intercept or grand mean while in the second the total has been corrected for the grand mean.

The coefficient of determination can be computed as

$$R^2 = \frac{SSReg}{Y'Y - C}.$$

This is the ratio of the variability explained by the predictor variables to the total variability of the data. Note that $(Y'Y - C)/(n-1) = s_y^2$, the sample variance of the ys without adjusting for any structure except the existence of a possibly nonzero mean.

EXAMPLE 11.4.1. *Simple linear regression*

For simple linear regression, we know that

$$SSReg = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2 = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2 \hat{\beta}_1$$

We will examine the alternative model

$$y_i = \beta_{*0} + \beta_1 (x_i - \bar{x}_\cdot) + \varepsilon_i.$$

Note that $C = n\hat{\beta}_{*0}^2$, so the general form for *SSReg* reduces to the simple linear regression form because

$$\begin{aligned} SSReg &= \hat{\beta}_*' Z' Z \hat{\beta}_* - C \\ &= \begin{bmatrix} \hat{\beta}_{*0} \\ \hat{\beta}_1 \end{bmatrix}' \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_{*0} \\ \hat{\beta}_1 \end{bmatrix} - C \\ &= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_\cdot)^2 . \end{aligned}$$

The same result can be obtained from $\hat{\beta}' X' X \hat{\beta} - C$ but the algebra is more tedious. □

To obtain tests and confidence regions we need to make additional distributional assumptions. In particular, we assume that the y_is have independent normal distributions. Equivalently, we take

$$\varepsilon_1, \ldots, \varepsilon_n \text{ indep. } N(0, \sigma^2).$$

To test the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

use the analysis of variance table test statistic

$$F = \frac{MSReg}{MSE}.$$

Under H_0,

$$F \sim F(p-1, n-p).$$

We can also perform a variety of t tests for individual regression parameters β_k. The procedures fit into the general techniques of Chapter 3 based on identifying 1) the parameter, 2) the estimate, 3) the standard error of the estimate, and 4) the distribution of $(Est - Par)/SE(Est)$. The parameter of interest is β_k. Having previously established that

$$\mathrm{E} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix},$$

it follows that for any $k = 0, \ldots, p-1$,

$$\mathrm{E}\left(\hat{\beta}_k\right) = \beta_k.$$

This shows that $\hat{\beta}_k$ is an unbiased estimate of β_k. Before obtaining the standard error of $\hat{\beta}_k$, it is necessary to identify its variance. The covariance matrix of $\hat{\beta}$ is $\sigma^2 (X'X)^{-1}$, so the variance of $\hat{\beta}_k$ is the $(k+1)$st diagonal element of $\sigma^2 (X'X)^{-1}$. The $(k+1)$st diagonal element is appropriate because the first diagonal element is the variance of $\hat{\beta}_0$ not $\hat{\beta}_1$. If we let a_k be the $(k+1)$st diagonal element of $(X'X)^{-1}$ and estimate σ^2 with MSE, we get a standard error for $\hat{\beta}_k$ of

$$\mathrm{SE}\left(\hat{\beta}_k\right) = \sqrt{MSE}\sqrt{a_k}.$$

Under normal errors, the appropriate reference distribution is

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}(\hat{\beta}_k)} \sim t(n-p).$$

Standard techniques now provide tests and confidence intervals. For example, a 95% confidence interval for β_k has endpoints

$$\hat{\beta}_k \pm t(.975, n-p)\,\mathrm{SE}(\hat{\beta}_k)$$

where $t(.975, n-p)$ is the 97.5th percentile of a t distribution with $n-p$ degrees of freedom.

A $(1-\alpha)100\%$ simultaneous confidence region for $\beta_0, \beta_1, \ldots, \beta_{p-1}$ consists of all the β vectors that satisfy

$$\frac{\left(\hat{\beta}-\beta\right)' X'X \left(\hat{\beta}-\beta\right) \Big/ p}{MSE} \leq F(1-\alpha, p, n-p).$$

This region also determines joint $(1-\alpha)100\%$ confidence intervals for the individual β_ks with limits

$$\hat{\beta}_k \pm \sqrt{pF(1-\alpha, p, n-p)}\,\mathrm{SE}(\hat{\beta}_k).$$

These intervals are an application of Scheffé's method of multiple comparisons, cf., Section 13.3.

We can also use the Bonferroni method to obtain joint $(1-\alpha)100\%$ confidence intervals with limits

$$\hat{\beta}_k \pm t\left(1-\frac{\alpha}{2p}, n-p\right)\mathrm{SE}(\hat{\beta}_k).$$

Finally, we consider estimation of the point on the surface that corresponds to a given set of predictor variables and the prediction of a new observation with a given set of predictor variables. Let the predictor variables be $x_1, x_2, \ldots, x_{p-1}$. Combine these into the row vector

$$x' = (1, x_1, x_2, \ldots, x_{p-1}).$$

The point on the surface that we are trying to estimate is the parameter $x'\beta = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j$. The least squares estimate is $x'\hat{\beta}$ which can be thought of as a 1×1 matrix. The variance of the estimate is

$$\mathrm{Var}\left(x'\hat{\beta}\right) = \mathrm{Cov}\left(x'\hat{\beta}\right) = x'\mathrm{Cov}\left(\hat{\beta}\right)x = \sigma^2 x' \left(X'X\right)^{-1} x,$$

so the standard error is

$$\mathrm{SE}\left(x'\hat{\beta}\right) = \sqrt{MSE}\sqrt{x' \left(X'X\right)^{-1} x} \equiv \mathrm{SE}(Surface).$$

This is the standard error of the estimated regression surface. The appropriate reference distribution is

$$\frac{x'\hat{\beta} - x'\beta}{\mathrm{SE}\left(x'\hat{\beta}\right)} \sim t(n-p)$$

and a $(1-\alpha)100\%$ confidence interval has endpoints

$$x'\hat{\beta} \pm t\left(1-\frac{\alpha}{2}, n-p\right)\mathrm{SE}(x'\hat{\beta}).$$

When predicting a new observation, the point prediction is just the estimate of the point on the surface but the standard error must incorporate the additional variability associated with a new observation. The original observations were assumed to be independent with variance σ^2. It is reasonable to assume that a new observation is independent of the previous observations and has the same variance. Thus, in the prediction we have to account for the variance of the new observation, which is σ^2, plus the variance of the estimate $x'\hat{\beta}$, which is $\sigma^2 x' \left(X'X\right)^{-1} x$. This leads to a variance for the prediction of $\sigma^2 + \sigma^2 x' \left(X'X\right)^{-1} x$ and a standard error of

$$\sqrt{MSE + MSE\, x' \left(X'X\right)^{-1} x} = \sqrt{MSE\left[1 + x' \left(X'X\right)^{-1} x\right]} \equiv \mathrm{SE}(Prediction).$$

Note that

$$\text{SE}(Prediction) = \sqrt{MSE + [\text{SE}(Surface)]^2}.$$

The $(1-\alpha)100\%$ prediction interval has endpoints

$$x'\hat{\beta} \pm t\left(1-\frac{\alpha}{2}, n-p\right)\sqrt{MSE\left[1+x'\left(X'X\right)^{-1}x\right]}.$$

Results of this section constitute the theory behind most of the applications in Sections 9.1 and 9.2.

11.5 Residuals, standardized residuals, and leverage

Let $x_i' = (1, x_{i1}, \ldots, x_{i,p-1})$ be the ith row of X, then the ith fitted value is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1} = x_i'\hat{\beta}$$

and the corresponding residual is

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - x_i'\hat{\beta}.$$

The vector of predicted values is

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_1'\hat{\beta} \\ \vdots \\ x_n'\hat{\beta} \end{bmatrix} = X\hat{\beta}.$$

The vector of residuals is

$$\begin{aligned} \hat{e} &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= \left(I - X(X'X)^{-1}X'\right)Y \\ &= (I-M)Y \end{aligned}$$

where

$$M \equiv X(X'X)^{-1}X'.$$

M is called the perpendicular projection operator (matrix) onto $C(X)$, the column space of X. M is the key item in the analysis of the general linear model, cf. Christensen (2011). Note that M is symmetric, i.e., $M = M'$, and idempotent, i.e., $MM = M$, so it is a perpendicular projection operator as discussed in Appendix A. Using these facts, observe that

$$\begin{aligned} SSE &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\ &= \hat{e}'\hat{e} \\ &= [(I-M)Y]'[(I-M)Y] \\ &= Y'(I-M'-M+M'M)Y \\ &= Y'(I-M)Y. \end{aligned}$$

Another common way of writing SSE is

$$SSE = \left[Y - X\hat{\beta}\right]'\left[Y - X\hat{\beta}\right].$$

Having identified M, we can define the standardized residuals. First we find the covariance matrix of the residual vector $\hat{e}$:

$$\begin{aligned}\mathrm{Cov}(\hat{e}) &= \mathrm{Cov}([I-M]Y)\\ &= [I-M]\mathrm{Cov}(Y)[I-M]'\\ &= [I-M]\sigma^2 I[I-M]'\\ &= \sigma^2(I-M-M'+MM')\\ &= \sigma^2(I-M).\end{aligned}$$

The last equality follows from $M = M'$ and $MM = M$. Typically, the covariance matrix is not diagonal, so the residuals are not uncorrelated.

The variance of a particular residual $\hat{\varepsilon}_i$ is σ^2 times the ith diagonal element of $(I-M)$. The ith diagonal element of $(I-M)$ is the ith diagonal element of I, 1, minus the ith diagonal element of M, say, m_{ii}. Thus

$$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1-m_{ii})$$

and the standard error of $\hat{\varepsilon}_i$ is

$$\mathrm{SE}(\hat{\varepsilon}_i) = \sqrt{MSE(1-m_{ii})}.$$

The ith standardized residual is defined as

$$r_i \equiv \frac{\hat{\varepsilon}_i}{\sqrt{MSE(1-m_{ii})}}.$$

The leverage of the ith case is defined to be m_{ii}, the ith diagonal element of M. Some people like to think of M as the 'hat' matrix because it transforms Y into $\hat{Y}$, i.e., $\hat{Y} = X\hat{\beta} = MY$. More common than the name 'hat matrix' is the consequent use of the notation h_i for the ith leverage. This notation was used in Chapter 7 but the reader should realize that $h_i \equiv m_{ii}$. In any case, the leverage can be interpreted as a measure of how unusual x_i' is relative to the other rows of the X matrix, cf. Christensen (2011, section 13.1).

11.6 Principal components regression

In Section 9.7 we dealt with the issue of collinearity. Four points were emphasized as the effects of collinearity.

1. The estimate of any parameter, say $\hat{\beta}_2$, depends on *all* the variables that are included in the model.
2. The sum of squares for any variable, say x_2, depends on *all* the other variables that are included in the model. For example, none of $SSR(x_2)$, $SSR(x_2|x_1)$, and $SSR(x_2|x_3,x_4)$ would typically be equal.
3. In a model such as $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, small t statistics for both $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ are not sufficient to conclude that an appropriate model is $y_i = \beta_0 + \beta_3 x_{i3} + \varepsilon_i$. To arrive at a reduced model, one must compare the reduced model to the full model.
4. A moderate amount of collinearity has little effect on predictions and therefore little effect on SSE, R^2, and the explanatory power of the model. Collinearity increases the variance of the $\hat{\beta}_j$s, making the estimates of the parameters less reliable. Depending on circumstances, sometimes a large amount of collinearity can have an effect on predictions. Just by chance one may get a better fit to the data than can be justified scientifically.

At its worst, collinearity involves near redundancies among the predictor variables. An exact redundancy among the predictor variables occurs when we can find a $p \times 1$ vector $d \neq 0$ so that $Xd = 0$. When this happens the rank of X is not p, so we cannot find $(X'X)^{-1}$ and we cannot find the estimates of β in Proposition 11.3.1. *Near* redundancies occur when we can find a vector d

Table 11.2: *Eigen analysis of the correlation matrix.*

Eigenvalue	2.8368	1.3951	0.4966	0.2025	0.0689
Proportion	0.567	0.279	0.099	0.041	0.014
Cumulative	0.567	0.846	0.946	0.986	1.000

that is not too small, say with $d'd = 1$, having $Xd \doteq 0$. Principal components (PC) regression is a method designed to identify near redundancies among the predictor variables. Having identified near redundancies, they can be eliminated if we so choose. In Section 10.7 we mentioned that having small collinearity requires more than having small correlations among all the predictor variables, it requires all partial correlations among the predictor variables to be small as well. For this reason, eliminating near redundancies cannot always be accomplished by simply dropping well chosen predictor variables from the model.

The basic idea of principal components is to find new variables that are linear combinations of the x_js and that are *best able to (linearly) predict the entire set of* x_j*s*, see Christensen (2001, Chapter 3). Thus the first principal component variable is the one linear combination of the x_js that is best able to predict all of the x_js. The second principal component variable is the linear combination of the x_js that is best able to predict all the x_js among those linear combinations having a sample correlation of 0 with the first principal component variable. The third principal component variable is the best predictor that has sample correlations of 0 with the first two principal component variables. The remaining principal components are defined similarly. With $p-1$ predictor variables, there are $p-1$ principal component variables. The full collection of principal component variables always predicts the full collection of x_js perfectly. The last few principal component variables are least able to predict the original x_j variables, so they are the least useful. They are also the aspects of the predictor variables that are most redundant, see Christensen (2011, Section 14.5). The best (linear) predictors used in defining principal components can be based on either the covariances between the x_js or the correlations between the x_js. Unless the x_js are measured on the same scale (with similarly sized measurements), it is generally best to use principal components defined using the correlations.

For *The Coleman Report* data, a matrix of sample correlations between the x_js was given in Example 9.7.1. Principal components are derived from the eigenvalues and eigenvectors of this matrix, cf., Section A.8. (Alternatively, one could use eigenvalues and eigenvectors of the matrix of sample covariances.) An eigenvector corresponding to the largest eigenvalue determines the first principal component variable.

The eigenvalues are given in Table 11.2 along with proportions and cumulative proportions. The proportions in Table 11.2 are simply the eigenvalues divided by the sum of the eigenvalues. The cumulative proportions are the sum of the first group of eigenvalues divided by the sum of all the eigenvalues. In this example, the sum of the eigenvalues is

$$5 = 2.8368 + 1.3951 + 0.4966 + 0.2025 + 0.0689.$$

The sum of the eigenvalues must equal the sum of the diagonal elements of the original matrix. The sum of the diagonal elements of a correlation matrix is the number of variables in the matrix. The third eigenvalue in Table 11.2 is .4966. The proportion is $.4966/5 = .099$. The cumulative proportion is $(2.8368 + 1.3951 + .4966)/5 = .946$. With an eigenvalue proportion of 9.9%, the third principal component variable accounts for 9.9% of the variance associated with predicting the x_js. Taken together, the first three principal components account for 94.6% of the variance associated with predicting the x_js because the third cumulative eigenvalue proportion is .946.

For the school data, the principal component (PC) variables are determined by the coefficients in Table 11.3. The first principal component variable is

Table 11.3: *Principal component variable coefficients.*

Variable	PC1	PC2	PC3	PC4	PC5
x_1	−0.229	−0.651	0.723	0.018	−0.024
x_2	−0.555	0.216	0.051	−0.334	0.729
x_3	−0.545	0.099	−0.106	0.823	−0.060
x_4	−0.170	−0.701	−0.680	−0.110	0.075
x_5	−0.559	0.169	−0.037	−0.445	−0.678

Table 11.4: *Table of coefficients: Principal component regression.*

Predictor	$\hat{\gamma}$	SE($\hat{\gamma}$)	t	P
Constant	35.0825	0.4638	75.64	0.000
PC1	−2.9419	0.2825	−10.41	0.000
PC2	0.0827	0.4029	0.21	0.840
PC3	−2.0457	0.6753	−3.03	0.009
PC4	4.380	1.057	4.14	0.001
PC5	1.433	1.812	0.79	0.442

$$\begin{aligned}\text{PC1}_i = &-0.229(x_{i1}-\bar{x}_{\cdot 1})/s_1 - 0.555(x_{i2}-\bar{x}_{\cdot 2})/s_2 \\ &- 0.545(x_{i3}-\bar{x}_{\cdot 3})/s_3 - 0.170(x_{i4}-\bar{x}_{\cdot 5})/s_4 - 0.559(x_{i5}-\bar{x}_{\cdot 5})/s_5 \quad (11.6.1)\end{aligned}$$

for $i = 1,\ldots,20$ where s_1 is the sample standard deviation of the x_{i1}s, etc. The columns of coefficients given in Table 11.3 are actually eigenvectors for the correlation matrix of the x_js. The PC1 coefficients are an eigenvector corresponding to the largest eigenvalue, the PC2 coefficients are an eigenvector corresponding to the second largest eigenvalue, etc.

We can now perform a regression on the new principal component variables. The table of coefficients is given in Table 11.4. The analysis of variance is given in Table 11.5. The value of R^2 is .906. The analysis of variance table and R^2 are identical to those for the original predictor variables given in Section 9.1. The plot of standardized residuals versus predicted values from the principal component regression is given in Figure 11.1. This is identical to the plot given in Figure 10.2 for the original variables. All of the predicted values and all of the standardized residuals are identical.

Since Table 11.5 and Figure 11.1 are unchanged, any usefulness associated with principal component regression must come from Table 11.4. The principal component variables display no collinearity. Thus, contrary to the warnings given earlier about the effects of collinearity, we can make final conclusions about the importance of variables directly from Table 11.4. We do not have to worry about fitting one model after another or about which variables are included in which models. From examining Table 11.4, it is clear that the important variables are PC1, PC3, and PC4. We can construct a reduced model with these three; the estimated regression surface is simply

$$\hat{y} = 35.0825 - 2.9419(\text{PC1}) - 2.0457(\text{PC3}) + 4.380(\text{PC4}), \quad (11.6.2)$$

where we merely used the estimated regression coefficients from Table 11.4. Refitting the reduced model is unnecessary because there is no collinearity.

Table 11.5: *Analysis of variance: Principal component regression.*

Source	df	SS	MS	F	P
Regression	5	582.69	116.54	27.08	0.000
Error	14	60.24	4.30		
Total	19	642.92			

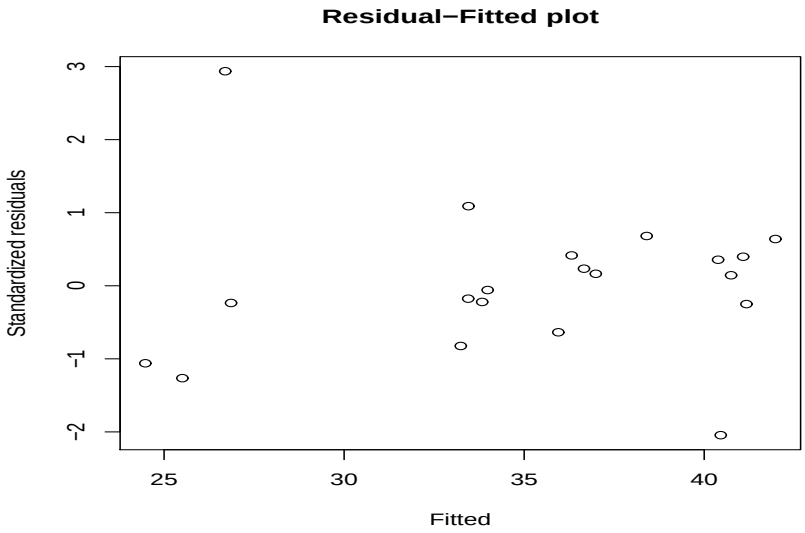

Figure 11.1: *Standardized residuals versus predicted values for principal component regression.*

To get predictions for a new set of x_js, just compute the corresponding PC1, PC3, and PC4 variables using formulae similar to those in equation (11.6.1) and make the predictions using the fitted model in equation (11.6.2). When using equations like (11.6.1) to obtain new values of the principal component variables, continue to use the $\bar{x}_{\cdot j}$s and s_js computed from only the original observations.

As an alternative to this prediction procedure, we could use the definitions of the principal component variables, e.g., equation (11.6.1), and substitute for PC1, PC3, and PC4 in equation (11.6.2) to obtain estimated coefficients on the original x_j variables.

$$\begin{aligned}
\hat{y} &= 35.0825 + [-2.9419, -2.0457, 4.380] \begin{bmatrix} \text{PC1} \\ \text{PC3} \\ \text{PC4} \end{bmatrix} \\
&= 35.0825 + [-2.9419, -2.0457, 4.380] \times \\
&\quad \begin{bmatrix} -0.229 & -0.555 & -0.545 & -0.170 & -0.559 \\ 0.723 & 0.051 & -0.106 & -0.680 & -0.037 \\ 0.018 & -0.334 & 0.823 & -0.110 & -0.445 \end{bmatrix} \begin{bmatrix} (x_1 - \bar{x}_{\cdot 1})/s_1 \\ (x_2 - \bar{x}_{\cdot 2})/s_2 \\ (x_3 - \bar{x}_{\cdot 3})/s_3 \\ (x_4 - \bar{x}_{\cdot 4})/s_4 \\ (x_5 - \bar{x}_{\cdot 5})/s_5 \end{bmatrix} \\
&= 35.0825 + [-0.72651, 0.06550, 5.42492, 1.40940, -0.22889] \times \\
&\quad \begin{bmatrix} (x_1 - 2.731)/0.454 \\ (x_2 - 40.91)/25.90 \\ (x_3 - 3.14)/9.63 \\ (x_4 - 25.069)/1.314 \\ (x_5 - 6.255)/0.654 \end{bmatrix}.
\end{aligned}$$

Obviously this can be simplified into a form $\hat{y} = 35.0825 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_3 x_3 + \tilde{\beta}_4 x_4 + \tilde{\beta}_5 x_5$, which in turn simplifies the process of making predictions and provides new estimated regression coefficients for the x_js that correspond to the fitted principal component model. These PC regression estimates of the original β_js can be compared to the least squares estimates. Many computer programs

for performing PC regression report these estimates of the β_js and their corresponding standard errors.

It was mentioned earlier that collinearity tends to increase the variance of regression coefficients. The fact that the later principal component variables are more nearly redundant is reflected in Table 11.4 by the fact that the standard errors for their estimated regression coefficients increase (excluding the intercept).

One rationale for using PC regression is that you just don't believe in using nearly redundant variables. The exact nature of such variables can be changed radically by small errors in the x_js. For this reason, one might choose to ignore PC5 because of its small eigenvalue proportion, regardless of any importance it may display in Table 11.4. If the t statistic for PC5 appeared to be significant, it could be written off as a chance occurrence or, perhaps more to the point, as something that is unlikely to be reproducible. If you don't believe redundant variables, i.e., if you don't believe that they are themselves reproducible, any predictive ability due to such variables will not be reproducible either.

When considering PC5, the case is pretty clear. PC5 accounts for only about 1.5% of the variability involved in predicting the x_js. It is a very poorly defined aspect of the predictor variables x_j and, anyway, it is not a significant predictor of y. The case is less clear when considering PC4. This variable has a significant effect for explaining y, but it accounts for only 4% of the variability in predicting the x_js, so PC4 is reasonably redundant within the x_js. If this variable is measuring some reproducible aspect of the original x_j data, it should be included in the regression. If it is not reproducible, it should not be included. From examining the PC4 coefficients in Table 11.3, we see that PC4 is roughly the average of the percent white collar fathers x_2 and the mothers' education x_5 contrasted with the socio- economic variable x_3. (Actually, this comparison is between the variables after they have been adjusted for their means and standard deviation as in equation (11.6.1).) If PC4 strikes the investigator as a meaningful, reproducible variable, it should be included in the regression.

In our discussion, we have used PC regression both to eliminate questionable aspects of the predictor variables and as a method for selecting a reduced model. We dropped PC5 primarily because it was poorly defined. We dropped PC2 solely because it was not a significant predictor. Some people might argue against this second use of PC regression and choose to take a model based on PC1, PC2, PC3, and possibly PC4.

On occasion, PC regression is based on the sample covariance matrix of the x_js rather than the sample correlation matrix. Again, eigenvalues and eigenvectors are used, but in using relationships like equation (11.6.1), the s_js are deleted. The eigenvalues and eigenvectors for the covariance matrix typically differ from those for the correlation matrix. The relationship between estimated principal component regression coefficients and original least squares regression coefficient estimates is somewhat simpler when using the covariance matrix.

It should be noted that PC regression is just as sensitive to violations of the assumptions as regular multiple regression. Outliers and high leverage points can be very influential in determining the results of the procedure. Tests and confidence intervals rely on the independence, homoscedasticity, and normality assumptions. Recall that in the full principal components regression model, the residuals and predicted values are identical to those from the regression on the original predictor variables. Moreover, highly influential points in the original predictor variables typically have a large influence on the coefficients in the principal component variables.

Minitab commands

Minitab commands for the principal components regression analysis are given below. The basic command is 'pca.' The 'scores' subcommand places the principal component variables into columns c12 through c16. The 'coef' subcommand places the eigenvectors into columns c22 through c26. If one wishes to define principal components using the covariances rather than the correlations, simply include a pca subcommand with the word 'covariance.'

```
MTB > pca c2-c6;
SUBC> scores c12-c16;
SUBC> coef c22-c26.
MTB > regress c8 on 5 c12-c16 c17 c18
MTB > plot c17 c18
```

11.7 Exercises

EXERCISE 11.7.1. Show that the form (11.3.2) simplifies to the form (11.3.1) for simple linear regression.

EXERCISE 11.7.2. Show that $\mathrm{Cov}(Y) = \mathrm{E}[(Y-\mu)(Y-\mu)']$.

EXERCISE 11.7.3. Use Proposition 1.2.11 to show that $\mathrm{E}(AY+c) = A\,\mathrm{E}(Y)+c$ and $\mathrm{Cov}(AY+c) = A\mathrm{Cov}(Y)A'$.

EXERCISE 11.7.4. Using eigenvalues, discuss the level of collinearity in:

(a) the Younger data from Exercise 9.12.1,

(b) the Prater data from Exercise 9.12.3,

(c) the Chapman data of Exercise 9.12.4,

(d) the pollution data from Exercise 9.12.5,

(e) the body fat data of Exercise 9.12.6.

EXERCISE 11.7.5. Do a principal components regression for the Younger data from Exercise 9.12.1.

EXERCISE 11.7.6. Do a principal components regression for the Prater data from Exercise 9.12.3.

EXERCISE 11.7.7. Do a principal components regression for the Chapman data of Exercise 9.12.4.

EXERCISE 11.7.8. Do a principal components regression on for the pollution data of Exercise 9.12.5.

EXERCISE 11.7.9. Do a principal components regression on for the body fat data of Exercise 9.12.6.