Chapter 9
Multi-Variable Regression

Abstract In many situations a variable of interest depends on several other variables. Such multi-variable data is common across the sciences and in many other fields such as economics and business. Multi-variable analysis can be performed in a simple and effective way when the relationship that links the variable of interest to the other quantities is linear. In this chapter we study the method of multi-variable regression and show how it is related to the multiple regression described in Chap. 8 which applies to the traditional two-variable dataset. This chapter also presents methods for hypothesis testing on the multi-variable regression and its parameters.

9.1 Multi-Variable Datasets

Two-dimensional dataset studied so far include an independent variable (X) and a dependent variable (Y) and the data take the form of a collection of (x_i, $y_i \pm \sigma_i$), where $i = 1, \dots, N$ and N indicates the total number of measurements. In Chap. 8 we have developed a method to fit such two-dimensional data. In that case, the linear regression formula takes the form of $y(x) = a + bx$, where a and b are the parameters of the linear regression.

Datasets that have measurements for three or more variables are referred to as multi-variable datasets. An example of multi-variable dataset is presented in Sect. 9.2, which reports measurements of different characteristics of irises performed by Fisher and Anderson in 1936 [14]. Each of those measurement comprises four quantities: the sepal length, sepal width, petal length, and petal width of 50 irises. For several multi-variable datasets such as that of Fisher and Anderson it is often unclear which variable is the dependent one. It typically depends on what question we want to address with the data: if we want to determine the sepal length of an iris flower based on the sepal width, petal length, and petal width, then the sepal length becomes the dependent variable and the remaining three are the independent variables.

Using multi-variable datasets to predict or forecast the behavior of one quantity based on several other variables is a fundamental topic in data analysis. It is common throughout the sciences and especially used in such fields as economics or behavioral sciences, where a number of possible factors can be used to predict one quantity of interest. An example is to predict the score on a college-admission

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts
in Physics, DOI 10.1007/978-1-4939-6572-4_9

test based on factors such as the grade-point average during the sophomore and the junior year, a measure of the motivation of the student and their economic status. Another example is to predict the price of a stock based, e.g., on the overall index of the stock exchange, a consumer's index for goods in the relevant class and the rate of treasury bonds. To address any such questions clearly requires a multi-variable dataset that has several measurements for all quantities of interest.

In this chapter we develop a method to determine the relationship between one of the quantities of a multi-dimensional datasets based on the others, assuming a linear relationship among the variables. This method will also let us study whether one or more of the quantities are in fact not useful in predicting the variable of interest. For example, we may find that the treasury bond rates are irrelevant in predicting the stock value of a given corporation and therefore we can focus only on those variables that are useful in predicting its stock price.

9.2 A Classic Experiment: The R.A. Fisher and E. Anderson Measurements of Iris Characteristics

R.A. Fisher is one of the fathers of modern statistics. In 1936 he published the paper *The Use of Multiple Measurements in Taxonomic Problems* reporting measurements of several characteristics of three species of the iris plant [14].

Figure 9.1 reproduces the original measurements, performed by E. Anderson, of the petal length and the sepal length of 150 iris plants of the species *Iris setosa*, *Iris versicolor*, and *Iris virginica*. The measurements are in millimeters (mm). Fisher's aim was to find a linear combination of the four characteristics that would be best suited to identify one species from the others. It is already clear from the data in Fig. 9.1 that one of the quantites (e.g., the sepal length) may be used as a discriminator among the three species. R.A. Fisher used this dataset to find a linear combination of the four quantities that would improve the classification of irises.

The dataset is a classic example of a multi-variate dataset, in which several variables are measured simultaneously and independently. In addition to Fisher's original purpose, these data can also be used to determine whether one of the characteristics, e.g., the sepal length, can be efficiently predicted based on any (or all) of the other characteristics. For example, one could expect that the length of the sepal (which is part of the calyx of the flower) is related linearly to its width, or to the length of the petal. Assuming a linear relationship among the variables, we set

$$SL = a + bSW + cPL + dPW \tag{9.1}$$

where a, b, c, and d are coefficients that we can estimate from the data using the method described in Sect. 9.3.

(continued)

Throughout this chapter we use these data to study the linear regression of (9.1) for the species *Iris setosa*. We will find that the most important variable needed to predict the sepal length is the sepal width, while the measurements of characteristics of petals are not very important in predicting the sepal length.

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	3.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Fig. 9.1 Measurements of three iris species from the 1936 R.A. Fisher paper [14]. Measurements are in mm

9.3 The Multi-Variable Linear Regression

Consider a dataset of N measurements of $m + 1$ variables which we call $Y, X_1, \ldots,$ X_m. We can use the index i to indicate the measurement, $i = 1, \ldots, N$, and the index k for the variables X_k, $k = 1, \ldots, m$. Each set of measurements is therefore indicated as $(y_i \pm \sigma_i, x_{1i}, \ldots, x_{mi})$.

We write the variable Y as a linear function of the m variables X_i,

$$y(x) = a_0 + a_1x_1 + \cdots + a_mx_m = a_0 + \sum_{k=1}^{m} a_kx_k. \tag{9.2}$$

The goal is to find the values for the $m + 1$ coefficients a_k, $k = 0, \ldots, m$ that minimize the χ^2 function

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2. \tag{9.3}$$

The quantity $y(x_i) = a_0 + a_1x_{1i} + \cdots + a_mx_{mi}$ is the value of $y(x)$ calculated for the i-th set of measurements of the X_k's. The coefficient a_0 is an overall offset, equivalent to the constant a for the two-dimensional linear regression function $y = a + bx$.

This form for the χ^2 function is the same as that used for the multiple linear regression of Sect. 8.4. The only change is that the measurements x_{ki} take the place of the functions $f_k(x_i)$. The quantity σ_i is interpreted as the error in the variable Y, which is the dependent quantity in this regression. As in the case of the two-variable dataset, we ignore the errors in the variables X_k (see Chap. 12 for an extension of the two-variable dataset regression with errors in both variables). When the multi-variable dataset has no errors, or if we choose to ignore the errors in the Y variable as well, we can omit the σ_i term in (9.3). This corresponds to assuming a uniform error for all measurements.

The similarity in form between the χ^2 functions to minimize for the present multi-variable linear regression and the multiple regression of Sect. 8.4 means that we have already at hand a solution for the coefficients of the regression and their errors. We need to make the following substitutions:

$$\begin{cases} f_1(x) = 1 \equiv x_0 \text{ (thus } x_{0i}\text{'s are not needed)} \\ f_{k+1}(x) = x_k, \; k = 1, \ldots, m. \end{cases} \tag{9.4}$$

and use the solution from Sect. 8.4 with $m + 1$ terms. The best-fit parameters a_k can be found via the matrix equation

$$a = \beta A^{-1}, \tag{9.5}$$

where the row vectors $\boldsymbol{\beta}$ and $\boldsymbol{a}$ and the $(m+1) \times (m+1)$ symmetric matrix $\boldsymbol{A}$ are given by

$$\begin{cases} \boldsymbol{\beta} & = (\beta_0, \beta_1, \ldots, \beta_m) \qquad \text{in which } \beta_k = \sum\limits_{i=1}^{N} x_{ki} y_i / \sigma_i^2 \\ \boldsymbol{a} & = (a_0, a_1, \ldots, a_m) \\ A_{lk} & = \sum\limits_{i=1}^{N} \dfrac{x_{li} x_{ki}}{\sigma_i^2} \qquad (l, k \text{ component of } \boldsymbol{A}). \end{cases}$$

The errors and covariances among parameters are likewise given by the error matrix $\epsilon = A^{-1}$. Assuming a constant value for the variance σ^2 (i.e., uniform measurement errors), the matrix A and the vector β can be written in extended form as

$$A = \frac{1}{\sigma^2} \begin{bmatrix} N & \sum x_{1i} & \ldots & \sum x_{mi} \\ \sum x_{1i} & \sum x_{1i}^2 & \ldots & \sum x_{1i} x_{mi} \\ \ldots & & & \\ \sum x_{mi} & \sum x_{mi} x_{1i} & \ldots & \sum x_{mi}^2 \end{bmatrix} \tag{9.6}$$

$$\beta = \frac{1}{\sigma^2} \left(\sum y_i, \sum x_{1i} y_i, \ldots, \sum x_{mi} y_i\right) \tag{9.7}$$

where all sums are over the N measurements. An estimate for the variance σ^2 is given by

$$s^2 = \frac{1}{N - m - 1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{9.8}$$

where $\hat{y}_i = a_0 + a_1 x_{1i} + \cdots + a_m x_{mi}$ is calculated for the best-fit values of the coefficients a_k.

An alternative notation for finding the coefficients a_k makes use of the following definitions:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix}; X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1m} \\ 1 & x_{21} & \ldots & x_{2m} \\ \ldots & & & \\ 1 & x_{N1} & \ldots & x_{Nm} \end{bmatrix} \text{ and } a = \begin{bmatrix} a_0 \\ a_1 \\ \ldots \\ a_m \end{bmatrix} \tag{9.9}$$

where X is called the *design matrix* and we have arranged the Y measurements and the vector of coefficients in column vectors. With this notation, the least-squares approach gives the following solution for the coefficients [41]:

$$a = (X^T X)^{-1} X^T Y \tag{9.10}$$

It is easy to show that (9.5) and (9.10) are equivalent (see Problem 9.3). Using this notation, the error matrix is given by

$$\epsilon = s^2 (X^T X)^{-1}. \tag{9.11}$$

We therefore have two equivalent methods to calculate the coefficients of the multiple regression and their errors. The latter form (9.9) may be convenient if the data are already tabulated according to the form of matrix A and therefore a can be found using the matrix algebra of (9.10). The drawback is that the design matrix can be of very large size, $N \times (m + 1)$, where N is the number of measurements. The form of (9.5) is more compact, since the matrix A is $(m + 1) \times (m + 1)$, and the summation over the N measurements must be performed beforehand to obtain A.

9.4 Tests for Significance of the Multiple Regression Coefficients

The multi-variable linear regression model of (9.2) is specified by the $m + 1$ coefficients a_k. After determining their best-fit values and errors, it is necessary to establish whether the model is an accurate representation of the data and whether there are any independent variables X_i that do not provide significant contribution to the prediction of the Y variable. Both tasks can be performed using hypothesis testing on the relevant statistic. We discuss these tests of significance using the Fisher's data of Sect. 9.2

9.4.1 T-Test for the Significance of Model Components

It is necessary to test the significance of each of the $m + 1$ parameters of the multi-variable linear regression. The null hypothesis is that their true value is zero, i.e., the corresponding variable is not needed in the model. For this purpose, we show that the ratio of the parameter's best-fit value a_k and its standard deviation s_k,

$$t_k = \frac{a_k}{s_k} \tag{9.12}$$

is distributed like a Student's t distribution with $N - m - 1$ degrees of freedom.

Proof Following the derivation provided in Sect. 7.5.1 for the sample mean, we can write

$$t_k = \frac{(a_k - \mu_k)/\sigma_k}{s_k/\sigma_k} \tag{9.13}$$

where $\mu_k = 0$ is the null hypothesis and σ_k^2 is the unknown parent variance for the parameter. Recall that the sample variance of the parameter s_k^2 is obtained as a product of the diagonal term in the error matrix and the estimate of the data variance s^2. Accordingly we set

$$\frac{(N-m-1)s_k^2}{\sigma_k^2} \sim \frac{\sum(y_i - \hat{y}_i)^2}{\sigma_k^2} \sim \chi^2(N-m-1), \tag{9.14}$$

i.e., the denominator of t_k can be written as a function of a variable that is χ^2-distributed. It is also clear that, under the null hypothesis, $\mu_k = 0$ is the parent value of a_k, and therefore the numerator of t_k is distributed like a standard normal distribution.

It follows that t_k is distributed like a t distribution,

$$t_k \sim \frac{N(0,1)}{\sqrt{\chi^2(N-m-1)/(N-m-1)}} \sim t(N-m-1) \tag{9.15}$$

according to the definition of the t distribution of (7.33). □

To test for the significance of coefficient a_k we therefore use the critical value for the t distribution for the appropriate number of degrees of freedom and the desired confidence level.

Example 9.1 (Multi-Variable Linear Regression on Iris setosa Data) The data of Fig. 9.1 for the *Iris setosa* species are fit to the linear model of (9.1), where the sepal length is used as the Y variable and the remaining three variables are the independent variables. Using (9.5) and the inverse of matrix A for the errors, we find the results shown in Table 9.1, including the t scores for the four parameters of the multiple regression.

For each parameter is reported the probability to exceed the absolute value of the measured t according to a t distribution with $f = 46$ degrees of freedom, where $f = N - m - 1$ with $N = 50$ measurements and $m = 3$ independent variable. It is clear that the parameters a_2 and a_3, corresponding to the petal length and width, are not significant because of the large probability p to exceed their value under the null hypothesis. Accordingly, it would be meaningful to repeat the linear regression using only the sepal width as an estimator for the sepal length. ◇

Table 9.1 Multiple regression parameter for the *Iris setosa* data

Parameter	Best-fit value	Error	t score	p value
a_0	2.352	0.393	5.99	< 0.001
a_1	0.655	0.092	7.08	< 0.001
a_2	0.238	0.208	1.14	0.26
a_3	0.252	0.347	0.73	0.47

9.4.2 F-Test for Goodness of Fit

The purpose of the multi-variable linear model is to provide a fit to the data that is more accurate than a simple constant predictor, i.e., the average of the Y measurements. In other words, we want to establish whether any of the parameters $a_1, \ldots, a_m$ provides a significant improvement over the constant model with $a_1 = a_2 = \ldots = a_m = 0$.

For this purpose we write the total variance of the data as follows:

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \tag{9.16}$$

where $\hat{y}_i = y(x_i)$ is evaluated for the best-fit values of the parameters a_k. This equation can be shown to hold because the following property applies,

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0 \tag{9.17}$$

(see Problem 9.7). The parent variance σ^2 of the data is unknown and it is not required for this test. We therefore ignore it for the considerations that follow by setting $\sigma^2 = 1$. The three terms in (9.16) are interpreted as follows. The left-hand side term is the total variance of the data and it is distributed like

$$S^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2 \sim \chi^2(N-1). \tag{9.18}$$

The total variance S^2 can be interpreted as the variance obtained using a model with $a_1 = \ldots = a_m = 0$, i.e., a constant model equal to the average of the Y measurements.

The first term on the right-hand side is the *residual variance* after the data are fit to the linear model and it follows the usual χ^2 distribution

$$S_r^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \sim \chi^2(N-m-1) \tag{9.19}$$

because of the $m + 1$ parameters used in the fit. This is the usual variance obtained using the full model in which at least some of the a_k parameters are not equal to zero.

Finally, the second term on the right-hand side can be interpreted as the variance *explained* by the best-fit model and it is distributed like

$$S_e^2 = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \sim \chi^2(m). \tag{9.20}$$

The distribution of the last term can be explained by the independence between the two variables on the right-hand side of the equation and the distribution of the left-hand side term, following a derivation similar to that of Sect. 7.3. Such derivation is not discussed in this book.

The variances described above can be used to define the variable

$$F = \frac{S_e^2/m}{S_r^2/(N-m-1)} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2/m}{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2/(N-m-1)}, \tag{9.21}$$

which is distributed as an F variable with m, $N-m-1$ degrees of freedom under the null hypothesis that $a_1 = \ldots = a_m = 0$. The meaning of this variable is the ratio between the variance explained by the fit and the residual variance, each normalized by the respective degrees of freedom. A large value of this ratio is desirable, since it means that the model does a good job at explaining the variability of the data.

The measurement of F that results from the fit of a dataset to the multi-variable linear model can therefore be used to test the null hypothesis. If the measurement exceeds the critical value of the F distribution for the desired confidence level, the null hypothesis must be rejected and the linear model is considered acceptable.

Example 9.2 (F-Test of Iris setosa Data) The variances for the *Iris setosa* data are shown in Table 9.2. The variable F is

$$F = \frac{S_e^2/3}{S_r^2/46} = \frac{3.50/2}{2.59/46} = 20.76. \tag{9.22}$$

The 99 % ($p = 0.01$) critical value for an F distribution with 3, 46 degrees of freedom is 4.24. Therefore the null hypothesis that the linear model does *not* provide a significant improvement must be rejected. In practice, this means that the linear

Table 9.2 Variances and F-test results for the *Iris setosa* data

Variances	Value	d.o.f	F-test	Value	p value
S^2	6.09	$N-1$ =49			
S_r^2	2.59	$N-m-1$=46			
S_e^2	3.50	$m=3$			
			$\dfrac{S_e^2/m}{S_r^2/(N-m-1)}$	20.76	1.2×10^{-8}

model is warranted. The probability to exceed the measured value of 20.7 for the test statistic is 1.2×10^{-8}, i.e., very small. ◇

9.4.3 The Coefficient of Determination

The ratio of the explained variance S_e^2 to the total variance S^2, defined as

$$R^2 = \frac{S_e^2}{S^2} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{9.23}$$

is a common measure of the ability of the linear model to describe the data. This ratio is called the *coefficient of (multiple) determination* and it is $0 \leq R^2 \leq 1$. A value close to 1 indicates that the model describes the data with little additional variance left unexplained.

It is possible to relate the coefficient R^2 to the F-test variable defined in (9.21) and obtain an equivalent test for the multi-variable regression based on R^2 instead of the F variable (see, e.g., [41] and [29]). Since the two quantities are related, it is sufficient to test the overall multiple regression using the F test provided in the previous section. The advantage of reporting explicitly a value for R^2 is that we can identify in a simple way the amount of variance that remains in the data after performing the multiple regression.

Example 9.3 (R^2 Value for the Iris setosa Data) We can use the data in Table 9.2 to calculate a coefficient of multiple determination $R^2 = 0.575$. This number means that 57.5 % of the total data variance is explained by the best-fit regression model. ◇

In the case of the simple linear regression with just one independent variable, $y = a + bx$, the coefficient of determination is the same as the coefficient of linear correlation r defined earlier in (2.19) (see Problem 9.4). In this case it is possible to test the significance of the linear model using either the correlation coefficient r or the F test. The two tests will be equivalent.

Example 9.4 (Linear Fit to the Iris setosa Data Using a Single Independent Variable) In a previous example we have shown that the coefficients of multiple regression for the variables Petal Length and Petal Width were not statistically significant, according to the t test.

Excluding these two columns of data, a fit to the function $y = a + bx$, where Y is the Sepal Length and X the Sepal Width, can be shown to return the values $a = 2.64 \pm 0.31$ and $b = 0.69 \pm 0.09$ with a correlation coefficient of $r = 0.7425$ or a value of $F = 58.99$ for 1, 49 degrees of freedom (see Problem 9.5). The value of $r^2 = 0.551$ is very similar to that obtained from the full fit using the additional two variables. The fact that the reduction in r^2 is minimal between the $m = 3$ and the $m = 1$ case is an indication that the Sepal Length can be predicted with nearly the same precision using just the Sepal Width as an indicator. ◇

Summary of Key Concepts for this Chapter

□ *Multi-variable dataset* Simultaneous measurements of several (> 2) variables, usually without reference to a specific independent variable.

□ *Multi-variable linear regression*: Extension of the (multiple) linear regression to the case of multi-variable data. Best-fit coefficients are given by the matrix equation

$$a = (X^T X)^{-1} X^T Y.$$

□ *Coefficient of determination*: The ratio between the explained variance and total variance $R^2 = S_e^2/S^2 \leq 1$.

Problems

9.1 Calculate the best-fit parameters and uncertainties for the multi-variable regression of the *Iris setosa* data of Fig. 9.1.

9.2 Use an F test to determine whether the multi-variable regression of the *Iris setosa* data is justified or not.

9.3 Prove that (9.5) and (9.10) are equivalent. Take into consideration that in (9.5) the vectors a and β are row vectors. You may re-write (9.5) using column vectors.

9.4 Prove that the coefficient of determination R^2 for the simple linear regression $y = a + bx$ is equivalent to the sample correlation coefficient of (2.20).

9.5 Fit the *Iris setosa* data using the function $y = a + bx$, where Y is the Sepal Length and X the Sepal Width. For this fit, you will ignore the data associated with the petal. Determine the best-fit parameters of the linear model and their errors.

9.6 Using the results of Problem 9.5, determine whether there is sufficient evidence for the use of the simple $y = a + bx$ model for the data. Use a confidence level of 99 % to draw your conclusions.

9.7 Prove (9.17).