CHAPTER 6

Extra Sums of Squares and Tests for Several Parameters Being Zero

The ideas connected with the extra sum of squares principle are extremely important and must be understood fully by regression practitioners.

6.1. THE "EXTRA SUM OF SQUARES" PRINCIPLE

In regression work, the question often arises as to whether or not it was worthwhile to include certain terms in the model. This question can be investigated by considering the extra portion of the regression sum of squares which arises due to the fact that the terms under consideration *were* in the model. The mean square derived from this extra sum of squares can then be compared with the estimate, s^2, of σ^2 to see if it appears significantly large. If it does, the terms should have been included; if it does not, the terms would be judged unnecessary and could be removed.

We have already seen an example of this in the case of fitting a straight line where SS $(b_1|b_0)$ represented the extra sum of squares due to including the term $\beta_1 X$ in the model. We now state the procedure more generally. Suppose the functions Z_1, Z_2, $\ldots$, Z_{p-1} are known functions of the basic variables X_1, X_2, $\ldots$, and suppose that values of the X's and the corresponding response Y are available. Consider the two models below.

1. $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_{p-1} Z_{p-1} + \epsilon.$

Suppose we obtain the following least squares estimates: $b_0(1), b_1(1), b_2(1), \ldots , b_{p-1}(1)$ and suppose that $\text{SS}(b_0(1), b_1(1), b_2(1), \ldots , b_{p-1}(1)) = S_1$, and there is no lack of fit. Let the estimate of σ^2 be s^2, obtained from the residual of Model 1.

2. $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_{q-1} Z_{q-1} + \epsilon(q < p).$

The Z's in this Model 2 are the same functions as in Model 1 when subscripts are the same. There are, however, fewer terms in this second model.

Suppose we now obtain the following least squares estimates: $b_0(2)$, $b_1(2)$, $b_2(2)$, $\ldots$, $b_{q-1}(2)$. *Note:* These may or may not be the same as $b_0(1)$, $b_1(1)$, $\ldots$, $b_{q-1}(1)$ above. If they are identical then $b_i(1)$ and $b_j(1)$ are orthogonal linear functions for $1 \leq i \leq q - 1$, $q \leq j \leq p - 1$. This happens when, in Model 1, the first q columns of the $\mathbf{X}$ matrix are all orthogonal to the last $p - q$ columns. This can happen in planned experiments. It rarely happens otherwise. See Appendix 6A.

Suppose that $\mathrm{SS}(b_0(2), b_1(2), b_2(2), \ldots, b_{q-1}(2)) = S_2$, for this second model. Then $S_1 - S_2$ is the *extra sum of squares*, due to the inclusion of the terms $\beta_q Z_q + \cdots + \beta_{p-1} Z_{p-1}$ in Model 1. Since S_1 has p degrees of freedom and S_2 has q degrees of freedom, $S_1 - S_2$ has $(p - q)$ degrees of freedom. It can be shown that, if $\beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$, then $E\{(S_1 - S_2)/(p - q)\} = \sigma^2$. In addition, if the errors are normally distributed, $(S_1 - S_2)$ will then be distributed as $\sigma^2\chi^2_{p-q}$ independently of s^2. This means we can compare $(S_1 - S_2)/(p - q)$ with s^2 by an $F(p - q, \nu)$ test, where ν is the number of degrees of freedom on which s^2 is based, to test the hypothesis $H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$.

We can write $S_1 - S_2$ conveniently as $\mathrm{SS}(b_q, \ldots, b_{p-1}|b_0, b_1, \ldots, b_{q-1})$ where we must keep in mind that two models are actually involved since the notation does not show it. This is read as *the sum of squares of* $b_q, \ldots, b_{p-1}$ *given* $b_0, b_1 \ldots, b_{q-1}$. By continued application of this principle we can obtain, successively, for any regression model, $\mathrm{SS}(b_0)$, $\mathrm{SS}(b_1|b_0)$, $\mathrm{SS}(b_2|b_0, b_1), \ldots, \mathrm{SS}(b_{p-1}|b_0, b_1, \ldots, b_{p-2})$, if we wish. All these sums of squares are distributed independently of s^2 and equal their mean squares since each has one degree of freedom. The mean squares can be compared with s^2 by a series of F-tests. This is useful when the terms of the model have a logical "order of entry," as would be the case, for example, if $Z_j = X^j$. A judgment can then be made about how many terms should be in the model.

Polynomial Models

When the terms in the model occur in natural groupings, such as happens, for example, in polynomial models with (1) β_0, (2) first-order terms, and (3) second-order terms, we can construct alternative extra sums of squares, for example, $\mathrm{SS}(b_0)$, SS(first-order b's$|b_0$), SS(second-order b's$|b_0$, first-order b's), and compare *these* with s^2. The extra sum of squares principle can be used in many ways therefore to achieve whatever breakup of the regression sum of squares seems reasonable for the problem at hand.

Other Points

The number of degrees of freedom for each sum of squares will be the number of parameters before the vertical division line (except when the estimates are linearly dependent; this happens when $\mathbf{X}'\mathbf{X}$ is singular and the normal equations are linearly dependent and will not usually concern us. The number of degrees of freedom is then the maximum number of linearly independent estimates in the set being considered). These extra SS are distributed independently of s^2. The corresponding mean squares, which equal (sum of squares)/(degrees of freedom), can be divided by s^2 to provide an F-ratio for testing the hypothesis that the true values of the coefficients whose estimates gave rise to the extra sum of squares are zero.

The expected value of an extra sum of squares is evaluated in Appendix 6B.

The extra sum of squares principle is actually a special case of testing a general linear hypothesis. In the more general treatment the extra sum of squares is calculated from the residual sums of squares and not the regression sum of squares. Since the total sum of squares $\mathbf{Y}'\mathbf{Y}$ is the same for both regression calculations, we would obtain the same result numerically whether we used the difference of regression or residual sums of squares. See Eq. (6.1.8).

Two Alternative Forms of the Extra SS

We can decide to remove the correction factor $n\overline{Y}^2$ or not remove it, before we take the difference between two sums of squares to get an extra sum of squares. For example, suppose our initial model (Model 1) is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon, \tag{6.1.1}$$

and we want the extra SS for b_3, b_4, and b_5 given b_0, b_1, and b_2. The reduced model (Model 2) is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{6.1.2}$$

For Model 1, the regression SS is

$$\text{SS}(b_0, b_1, b_2, b_3, b_4, b_5) = S_1, \tag{6.1.3}$$

the correction factor is $n\overline{Y}^2$, and the total SS is $\mathbf{Y}'\mathbf{Y}$. Thus the residual SS is $\mathbf{Y}'\mathbf{Y} - S_1$.

For model 2, the regression SS is

$$\text{SS}(b_0, b_1, b_2) = S_2, \tag{6.1.4}$$

where we *no longer show* that the Model 2 b's could be different from the Model 1 b's of same subscript, though in general they are. (One has to get used to this notation and to realize that it conceals a possible confusion!) The correction factor is $n\overline{Y}^2$ and the total sum of squares is $\mathbf{Y}'\mathbf{Y}$. Thus the residual SS is $\mathbf{Y}'\mathbf{Y} - S_2$.

We now require

$$\text{SS}(b_3, b_4, b_5 | b_0, b_1, b_2) = S_1 - S_2. \tag{6.1.5}$$

[*Note:* A reordering of b's before the vertical bar and/or a reordering after the vertical bar does not change the meaning of (6.1.5).] We can rewrite this as

$$S_1 - S_2 = (S_1 - n\overline{Y}^2) - (S_2 - n\overline{Y}^2) \tag{6.1.6}$$

when it becomes a difference between sums of squares corrected for b_0, that is,

$$\text{SS}(b_1, b_2, b_3, b_4, b_5 | b_0) - \text{SS}(b_1, b_2 | b_0). \tag{6.1.7}$$

There is yet a third way to get this extra sum of squares. We can rewrite the $S_1 - S_2$ as

$$S_1 - S_2 = (\mathbf{Y}'\mathbf{Y} - S_2) - (\mathbf{Y}'\mathbf{Y} - S_1) \tag{6.1.8}$$

when it becomes a difference of residual SS but in reversed order, because the regression with the larger regression SS (S_1) must have the smaller residual SS; and vice versa for S_2. Of the three calculations, the best is the one you prefer! (For a specific matrix formula for the extra sum of squares in general, see Section 10.4.)

Sequential Sums of Squares

When we call the regression option in a programming system, we tell the computer a certain order for our X's. Sometimes this order has a meaning for us, sometimes it is just the order in which we wrote down the data. Let us suppose we fitted model

(6.1.1) and loaded the X's in the order shown there. Then we would (in some programs, e.g., MINITAB) or could (in others) see a printout of extra SS of form

$$\begin{aligned}
\mathrm{SS}_1 &= \mathrm{SS}(b_1|b_0),\\
\mathrm{SS}_2 &= \mathrm{SS}(b_2|b_1, b_0),\\
\mathrm{SS}_3 &= \mathrm{SS}(b_3|b_2, b_1, b_0),\\
\mathrm{SS}_4 &= \mathrm{SS}(b_4|b_3, b_2, b_1, b_0),\\
\mathrm{SS}_5 &= \mathrm{SS}(b_5|b_4, b_3, b_2, b_1, b_0),
\end{aligned} \tag{6.1.9}$$

often called the sequential sum of squares printout. If, as in our example above, we wanted the extra SS (6.1.5) or (6.1.6) or (6.1.8) we could get it by summing $\mathrm{SS}_5 + \mathrm{SS}_4 + \mathrm{SS}_3$ in (6.1.9). In any printout like this where subscripts 1 and 2 come first and second (in either order 12 or 21) and subscripts 3, 4, and 5 come third, fourth, and fifth (in order 345 or 354 or 435 or 453 or 534 or 543), a similar calculation will give the correct answer. We cannot get $\mathrm{SS}(b_1, b_4, b_5|b_0, b_2, b_3)$ from (6.1.9), however. The breakdown given does not allow this, except in special cases where there is enough pairwise orthogonality among the columns of the $\mathbf{X}$ matrix to make the calculation correct. In general, it will not work and there is little point discussing the exceptions in much more detail than is given in Appendix 6A. It is usually easier to rerun the regression with another ordering of the predictor variables.

We follow our example one more step. Suppose we wish to test $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ in (6.1.1) versus H_1: not so. (There are many ways H_0 would not be true, so this is the easiest way to state the alternative hypothesis.) Our F-test would be carried out on

$$F = \{(S_1 - S_2)/(6 - 3)\}/s^2,$$

where s^2 is the residual mean square from the larger of the two models, namely, from the fit of (6.1.1). The degrees of freedom would be $6 - 3 = 3$ for the numerator and $(n - 6)$ for the denominator, where n is the number of observations.

Special Problems with Polynomial Models

In models where the X's are individual predictor variables, it may not make any difference which β's are set equal to zero in H_0 and so tested via an extra SS test. In the case of a polynomial model, however, certain tests do not make practical sense. For example, suppose that in (6.1.1), we had

$$X_3 = X_1^2, \qquad X_4 = X_2^2, \qquad X_5 = X_1 X_2. \tag{6.1.10}$$

Testing $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ *is* sensible because it answers the question: "Do we need the quadratic curvature in the model?" The question of whether $H_0: \beta_1 = \beta_2 = 0$ is true is, in general, *not* a good one, as it is asking if the stationary point of the surface lies at the origin, a very rare event that depends on the coding of the factors as well as the shape of the surface. We discuss this issue in more detail in Chapter 12 and there suggest some rules that may be useful.

Partial Sums of Squares

We have seen how to obtain extra sums of squares for one or more estimated coefficients given other coefficients by considering two models, one of which includes the coefficients in question and one of which does not.

If we have several terms in a regression model we can think of them as "entering" the equation in any desired sequence. If we find

$$\text{SS}(b_i|b_0, b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_k), \qquad i = 1, 2, \ldots, k, \tag{6.1.11}$$

we shall have a one degree of freedom sum of squares, which measures the contribution to the regression sum of squares of each coefficient b_i given that all the terms that did not involve β_i were already in the model. In other words, we shall have a measure of the value of *adding a β_i term to the model* that originally did not include such a term. Another way of saying that is that we have a measure of the value of β_i *as though it were added to the model last.* The corresponding mean square, equal to the sum of squares since it has one degree of freedom, can be compared by an F-test to s^2 as described. This particular type of F-test is often called a *partial F-test* for β_i. If the extra term under consideration is $\beta_t X_t$, say, we can talk (loosely) about a partial F-test on the variable X_t, even though we are aware that the test actually is on the coefficient β_t.

When a suitable model is being "built" the partial F-test is a useful criterion for adding or removing terms from the model. The effect of an X-variable (X_q, say) in determining a response may be large when the regression equation includes only X_q. However, when the same variable is entered into the equation after other variables, it may affect the response very little, due to the fact that X_q is highly correlated with variables already in the regression equation. The partial F-test can be made for all regression coefficients as though each corresponding variable were the last to enter the equation—to see the relative effects of each variable in excess of the others. This information can be combined with other information if a choice of variables need be made. Suppose, for example, either X_1 or X_2 alone could be used to provide a regression equation for a response Y. Suppose use of X_1 provided smaller predictive errors than use of X_2. Then if predictive accuracy were desired, X_1 would probably be used in future work. If, however, X_2 were a variable through which the response level could be controlled (whereas X_1 was a measured but noncontrolling variable) and if control were important rather than prediction, then it might be preferable to use X_2 rather than X_1 as a predictor variable for future work.

When $t = F^{1/2}$

The partial F-statistic with 1 and ν degrees of freedom for testing $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$ is exactly equal to the square of the t-statistic with ν degrees of freedom obtained via $t = b_j/\{\text{se}(b_j)\}$, where $\text{se}(b_j)$, the standard error of b_j, is the square root of the appropriate diagonal term of $(\mathbf{X}'\mathbf{X})^{-1}s^2$ and s^2 is based on ν df. (This is a distributional fact that we do not prove.) The test can be made in either F or t form with the same results. Examination of the tables of percentage points will show that $F(1, \nu, 1 - \alpha) = \{t(\nu, 1 - \alpha/2)\}^2$ for any values of ν and α. (As always, round-off errors will sometimes prevent the numerical relationship from being exact to the number of figures quoted.)

6.2. TWO PREDICTOR VARIABLES: EXAMPLE

We now look at a two-predictor example using part of the steam data. This enables us to illustrate some of the details mentioned in Section 6.1, as well as the matrix algebra of Chapter 5. Consider the first-order linear model of form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{6.2.1}$$

We shall continue with the example used in Chapter 1 (the data for which are given in Appendix 1A) and will now add variable number 6 to the problem. So that we are clear about which variables are being considered in the model, we shall use the original variable subscripts. Thus our model will be written

$$Y = \beta_0 X_0 + \beta_8 X_8 + \beta_6 X_6 + \epsilon, \tag{6.2.2}$$

where Y = response or number of pounds of steam used per month, coded,
X_0 = dummy variable, whose value is always unity,
X_8 = average atmospheric temperature in the month (in °F),
X_6 = number of operating days in the month.

The following matrices can be constructed. (The complete figures for the vector $\mathbf{Y}$ and the second and third columns of matrix $\mathbf{X}$ appear in Appendix 1A and are also given in Table 6.1.)

$$\mathbf{Y} = \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ 8.4 \\ \vdots \\ 10.36 \\ 11.08 \end{bmatrix}, \quad \mathbf{X} = \begin{array}{c} \begin{array}{ccc} X_0 & X_8 & X_6 \end{array} \\ \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ 1 & 58.8 & 20 \\ \vdots & \vdots & \vdots \\ 1 & 33.4 & 20 \\ 1 & 28.6 & 22 \end{bmatrix} \end{array}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_8 \\ \beta_6 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_{24} \\ \epsilon_{25} \end{bmatrix}$$

where $\mathbf{Y}$ is a (25×1) vector,
$\mathbf{X}$ is a (25×3) matrix,
$\boldsymbol{\beta}$ is a (3×1) vector,
$\boldsymbol{\epsilon}$ is a (25×1) vector.

Using the results of Chapter 5, the least squares estimates of β_0, β_8, and β_6 are given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

where **b** is the vector of estimates of the elements of $\boldsymbol{\beta}$, provided that $\mathbf{X}'\mathbf{X}$ is nonsingular. Thus

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & 30.8 & \cdots & 28.6 \\ 20 & 20 & 23 & \cdots & 22 \end{bmatrix} \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ \vdots & \vdots & \vdots \\ 1 & 28.6 & 22 \end{bmatrix} \right\}^{-1}$$

$$\times \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & 30.8 & \cdots & 28.6 \\ 20 & 20 & 23 & \cdots & 22 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \vdots \\ 11.08 \end{bmatrix}.$$

Note the sizes of the matrices in the above statement:

$$[3 \times 1] = \{[3 \times 25][25 \times 3]\}^{-1}[3 \times 25][25 \times 1].$$

Multiplying the matrices within the large braces, we have

$$\underset{[3 \times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}} = \underset{[3 \times 3]^{-1}}{\begin{bmatrix} 25.00 & 1315.00 & 506.00 \\ 1315.00 & 76323.42 & 26353.30 \\ 506.00 & 26353.30 & 10460.00 \end{bmatrix}^{-1}}$$

$$\times \underset{[3 \times 25]}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & \cdots & 28.6 \\ 20 & 20 & \cdots & 22 \end{bmatrix}} \underset{[25 \times 1]}{\begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \vdots \\ 11.08 \end{bmatrix}}.$$

Then,

$$\underset{[3 \times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}} = \underset{[3 \times 3]^{-1}}{\begin{bmatrix} 25.00 & 1315.00 & 506.00 \\ 1315.00 & 76323.42 & 26353.30 \\ 506.00 & 26353.30 & 10460.00 \end{bmatrix}^{-1}} \underset{[3 \times 1]}{\begin{bmatrix} 235.6000 \\ 11821.4320 \\ 4831.8600 \end{bmatrix}}.$$

Next, the inverse of the [3 × 3] matrix is obtained to give

$$\underset{[3\times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}} = \underset{[3\times 3]}{\begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ & 0.146207\times 10^{-3} & 0.175467\times 10^{-3} \\ \text{(Symmetric)} & & 0.478599\times 10^{-2} \end{bmatrix}} \times \underset{[3\times 1]}{\begin{bmatrix} 235.6000 \\ 11821.4320 \\ 4831.8600 \end{bmatrix}}.$$

The inverse calculation can be checked by multiplying $(\mathbf{X}'\mathbf{X})^{-1}$ by the original $(\mathbf{X}'\mathbf{X})$ to give a 3 × 3 unit matrix. Note that, since the inverse (like the original matrix) is symmetric, only an upper triangular portion of it is recorded. Performing the matrix multiplication gives

$$\underset{[3\times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}} = \underset{[3\times 1]}{\begin{bmatrix} 9.1266 \\ -0.0724 \\ 0.2029 \end{bmatrix}}.$$

Thus the fitted least squares equation is

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6.$$

Actually, when these matrix calculations are performed by a computer routine, they are not carried through in precisely this way. One reason for this is that large rounding errors may occur when this sequence is followed.

Substitution into the fitted equation of the data values of X_8 and X_6 leads to the fitted values $\hat{Y}_i$ and residuals $Y_i - \hat{Y}_i$ given in Table 6.1. A plot of the observations Y_i and the fitted values $\hat{Y}_i$ is shown in Figure 6.1.

How Useful Is the Fitted Equation?

The analysis of variance table takes the following form:

ANOVA

Source of Variation	df	SS	MS	F
Regression $\|b_0$	2	54.1871	27.0936	61.8999
Residual	22	9.6287	0.4377	
Total (corrected)	24	63.8158		
Mean (b_0)	1	2220.2944		
Total (uncorrected)	25	2284.1102		

Provided that further examination of the model and residuals shows no flaw, the least squares equation

TABLE 6.1. Steam Data, Fitted Values and Residuals

Observation Number	X_8	X_6	Y	$\hat{Y}$	Residual
1	35.3	20	10.98	10.63	0.35
2	29.7	20	11.13	11.03	0.10
3	30.8	23	12.51	11.56	0.95
4	58.8	20	8.40	8.93	−0.53
5	61.4	21	9.27	8.94	0.33
6	71.3	22	8.73	8.43	0.30
7	74.4	11	6.36	5.97	0.39
8	76.7	23	8.50	8.24	0.26
9	70.7	21	7.82	8.27	−0.45
10	57.5	20	9.14	9.02	0.12
11	46.4	20	8.24	9.82	−1.58
12	28.9	21	12.19	11.29	0.90
13	28.1	21	11.88	11.35	0.53
14	39.1	19	9.57	10.15	−0.58
15	46.8	23	10.94	10.40	0.54
16	48.5	20	9.58	9.67	−0.09
17	59.3	22	10.09	9.30	0.79
18	70.0	22	8.11	8.52	−0.41
19	70.0	11	6.83	6.29	0.54
20	74.5	23	8.88	8.40	0.48
21	72.1	20	7.68	7.96	−0.28
22	58.1	21	8.47	9.18	−0.71
23	44.6	20	8.86	9.96	−1.10
24	33.4	20	10.36	10.77	−0.41
25	28.6	22	11.08	11.52	−0.44
			235.60		$\Sigma(Y_i - \hat{Y}_i) = 0$
			$\bar{Y} = 9.424$		$\Sigma(Y_i - \hat{Y}_i)^2 = 9.6432$

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6$$

is a significant explanation of the data. The calculated $F = 61.90$ exceeds the tabulated $F(2, 22, 0.95) = 3.44$ by a healthy margin. In fact, the tail area beyond 61.90 is only $p = 0.00007$ for the $F(2, 22)$ distribution.

What Has Been Accomplished by the Addition of a Second Predictor Variable (Namely, X_6)?

There are several useful criteria that can be applied to answer this question, and we now discuss them.

R^2

The square of the multiple correlation coefficient R^2 is defined as

$$R^2 = \frac{\text{Sum of squares due to regression } |b_0}{\text{Total (corrected) sum of squares}}.$$

It is often stated as a percentage, $100R^2$. The larger it is, the better the fitted equation explains the variation in the data. We can compare the value of R^2 at each stage of the regression problem:

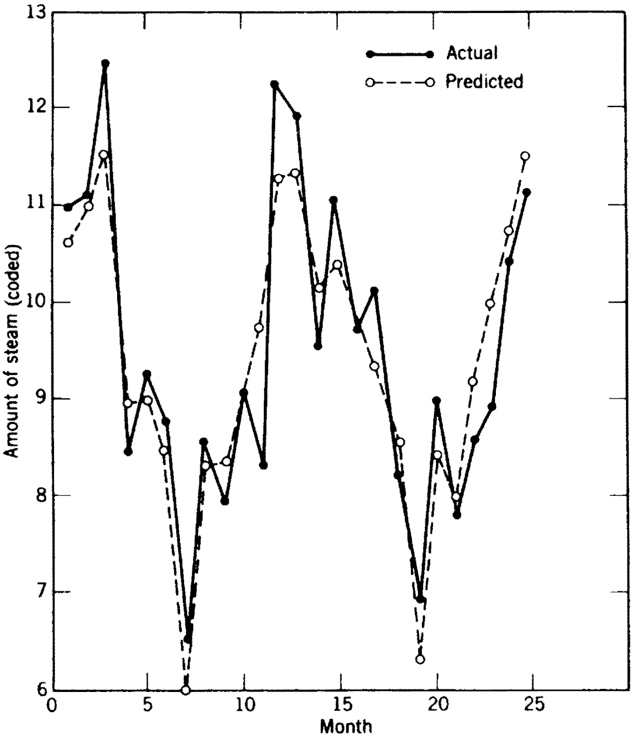

Figure 6.1. Plot of Y_i and $\hat{Y}_i$ values by month, for steam data.

Step 1.	$Y = f(X_8)$.	
	Regression equation	$100\,R^2$
	$\hat{Y} = 13.6230 - 0.0798X_8$	71.44% (see Section 1.3)
Step 2.	$Y = f(X_8, X_6)$.	
	Regression equation	$100\,R^2$
	$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6$	84.89%

Thus we see a substantial increase in R^2.

The addition of a new predictor variable to a regression will generally increase R^2. (More exactly, it cannot decrease it and will leave it the same only if the new predictor is a linear combination of the predictors already in the equation.) Moreover, the addition of more and more predictors will give the highest feasible value of R^2 when the number of data sites equals the number of parameters. The pure error can never be explained by any fitted model, however, as already mentioned.

The increase in R^2 from one equation to the other could be tested but it is pointless to do so, because the R^2 statistic is related to the F-test for regression given b_0, while the increase in R^2 is related to an extra SS F-test. Thus all desired R^2 tests are conducted via F-tests.

The Standard Error *s*

The residual mean square s^2 is an estimate of $\sigma^2_{Y \cdot X}$, the variance about the regression. Before and after adding a variable to the model, we can check

$$s = \sqrt{\text{Residual mean square}}.$$

Examination of this statistic indicates that the smaller it is the better, that is, the more precise will be the predictions. Of course, s can be reduced to the pure error value (or to zero if there is no pure error) by including as many parameters as there are data sites. Apart from an approach to such an extreme, reduction of s is desirable. In our example at Step 1,

$$s = \sqrt{0.7926} = 0.89.$$

At Step 2,

$$s = \sqrt{0.4377} = 0.66.$$

Thus the addition of X_6 has decreased s and improved the precision of estimation.

The value of s is not always decreased by adding a predictor variable. This is because the reduction in the residual sum of squares may be less than the original residual mean square. Since one degree of freedom is removed from the residual degrees of freedom as well, the resulting mean square may get larger.

$s/\overline{Y}$

A useful way of looking at the decrease in s is to consider it in relation to the response. In our example, at Step 1, s as a percentage of mean $\overline{Y}$ is

$$0.89/9.424 = 9.44\%.$$

At Step 2, s as a percentage of mean $\overline{Y}$ is

$$0.66/9.424 = 7.00\%$$

The addition of X_6 has reduced the standard error of estimate down to about 7% of the mean response. Whether this level of precision is satisfactory or not is a matter for the experimenter to decide, on the basis of prior knowledge and personal feelings.

Extra SS F-Test Criterion

This method consists of breaking down the sum of squares due to regression given b_0 into two sequential pieces as follows:

ANOVA

Source of Variation	df	SS	MS	F
Regression $\|b_0$	2	54.1871	27.0936	61.8999
Due to $b_8\|b_0$	1	45.5924	45.5924	104.1636
Due to $b_6\|b_8, b_0$	1	8.5947	8.5947	19.6361
Residual	22	9.6287	0.4377	
Total (corrected)	24	63.8158		

The F-value of 19.64 exceeds $F(1, 22, 0.95) = 4.30$ by a factor of more than four, indicating a statistically significant contribution by the addition of X_6 to the equation.

We can, of course, also consider what would have been the effect of adding the variables in the reverse order, X_6 followed by X_8. It is still the same amount $\text{SS}(b_8, b_6|b_0) = 54.1871$, which is split up, but the split is now different (and will be different in general except when the conditions of Appendix 6A prevail). We obtain:

ANOVA

Source of Variation	df	SS	MS	F
Regression $\|b_0$	2	54.1871	27.0936	61.8999
Due to $b_6\|b_0$	1	18.3424	18.3424	41.9063
Due to $b_8\|b_6, b_0$	1	35.8447	35.8447	81.8933
Residual	22	9.6287	0.4377	
Total (corrected)	24	63.8158		

To get the entry for SS($b_6|b_0$) we have to fit $\hat{Y} = 3.561 + 0.2897X_6$ and evaluate the expression $S_{6Y}^2/S_{66} = 18.3424$. Comparing the sums of squares, we find:

Contribution of	X_8 in First		X_6 in First
X_8	45.59		35.84
X_6	8.59		18.34
Totals	54.18	=	54.18

In this example, each variable picks up more of the variation when it gets into the equation first than it does when it gets in second. This is also reflected in the corresponding F-values. However, X_8 is still the more important variable in both cases, since its contribution in reducing the residual sum of squares is the larger, regardless of the order of introduction of the variables. Behavior like this is common, but is not guaranteed. See Appendix 6B.

Standard Error of b_i

Using the result given in Section 5.2, the variance–covariance matrix of **b** is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

Thus variance of $b_i = V(b_i) = c_{ii}\sigma^2$, where c_{ii} is the diagonal element in $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to the ith variable.

The covariance of b_i, $b_j = c_{ij}\sigma^2$, where c_{ij} is the off-diagonal element in $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to the intersection of the ith row and jth column, or jth row and ith column, since $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric.

Thus the standard deviation of b_i is $\sigma\sqrt{c_{ii}}$ and we replace σ by s to obtain the standard error of b_i. For example, the standard error of b_8 is obtained as follows:

$$\begin{aligned}\text{est. var}(b_8) &= s^2c_{88}\\ &= (0.4377)(0.146207 \times 10^{-3})\\ &= 0.639948 \times 10^{-4}.\end{aligned}$$

Then $\text{se}(b_8) = \sqrt{\text{est. var}(b_8)} = \sqrt{0.639948 \times 10^{-4}} = 0.008000$.

Note that the t-statistic, $t = b_8/\text{se}(b_8) = -0.0724/0.008 = -9.05$ so that $t^2 = 81.9025$. In theory, this is *identical* to the partial F-value $F_{8|6,0} = 81.8933$ in the foregoing table. As usual, rounding errors have crept into the calculations. The parallel calculation for b_6 is $t = 0.2029/\{(0.4377)(0.478599 \times 10^{-2})\}^{1/2} = 0.2029/0.0458 = 4.433 = (19.6515)^{1/2}$, whereas $F_{6|8,0} = 19.6361$.

Correlations Between Parameter Estimates

We can convert the $(\mathbf{X}'\mathbf{X})^{-1}s^2$ matrix or, more simply, the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix, because s^2 cancels out, into a correlation matrix by dividing each row *and* each column by the square root of the appropriate diagonal entry. Using the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix of the steam data, for example, we divide the first row *and* the first column by the square root of the first diagonal entry, that is by, $(2.778747)^{1/2}$ and so on, moving down the diagonal. We obtain, to three decimal places,

$$\begin{array}{c} \\ 0 \\ 8 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & \quad 8 & \quad 6 \end{array} \\ \left[\begin{array}{ccc} 1.000 & -0.558 & -0.920 \\ & 1.000 & 0.210 \\ \text{(Symmetric)} & & 1.000 \end{array}\right] \end{array}.$$

The correlations between estimates are $\text{Corr}(b_0, b_8) = -0.558$, $\text{Corr}(b_0, b_6) = -0.920$, and $\text{Corr}(b_8, b_6) = 0.210$.

Confidence Limits for the True Mean Value of *Y*, Given a Specific Set of *X*s

The predicted value $\hat{Y} = b_0 + b_1X_1 + \cdots + b_{p-1}X_{p-1}$ is an estimate of

$$E(Y) = \beta_0 + \beta_1X_1 + \cdots + \beta_{p-1}X_{p-1}.$$

The variance of $\hat{Y}$, $V[b_0 + b_1X_1 + \cdots + b_{p-1}X_{p-1}]$, is

$$V(b_0) + X_1^2V(b_1) + \cdots + X_{p-1}^2V(b_{p-1})$$
$$+ 2X_1\,\text{cov}(b_0, b_1) + \cdots + 2X_{p-2}X_{p-1}\,\text{cov}(b_{p-2}, b_{p-1}).$$

This expression can be written very conveniently in matrix notation as follows, where $\text{C} = (\mathbf{X}'\mathbf{X})^{-1}$.

$$V(\hat{Y}) = \sigma^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)$$
$$= \sigma^2[1 \quad X_1 \quad \cdots \quad X_{p-1}] \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0,p-1} \\ c_{10} & c_{11} & \cdots & c_{1,p-1} \\ \vdots & & \ddots & \vdots \\ c_{p-1,0} & & & c_{p-1,p-1} \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix}.$$

Thus the $1 - \alpha$ confidence limits on the true mean value of Y at $\mathbf{X}_0$ are given by

$$\hat{Y} \pm t\{(n-p), 1 - \tfrac{1}{2}\alpha\} \cdot s\sqrt{\mathbf{X}_0'\mathbf{C}\mathbf{X}_0}.$$

For example, the variance of $\hat{Y}$ for the point in the X-space ($X_8 = 32$, $X_6 = 22$) is obtained as follows:

$$\text{est. var}(\hat{Y}) = s^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)$$
$$= (0.4377)(1, 32, 22)$$
$$\times \begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ -0.011242 & 0.146207 \times 10^{-3} & 0.175467 \times 10^{-3} \\ -0.106098 & 0.175467 \times 10^{-3} & 0.478599 \times 10^{-2} \end{bmatrix} \begin{bmatrix} 1 \\ 32 \\ 22 \end{bmatrix}$$
$$= (0.4377)(0.104140) = 0.045582.$$

The 95% confidence limits on the true mean value of Y at $X_8 = 32$, $X_6 = 22$ are given by

$$\begin{aligned}\hat{Y} \pm t(22, 0.975) \cdot s\sqrt{\mathbf{X}_0'\mathbf{C}\mathbf{X}_0} &= 11.2736 \pm (2.074)(0.213499)\\ &= 11.2736 \pm 0.4418\\ &= 10.8318, 11.7154.\end{aligned}$$

These limits are interpreted as follows. Suppose repeated samples of Y's are taken of the same size each time and at the same fixed values of (X_8, X_6) as were used to determine the fitted equation obtained above. Then of all the 95% confidence intervals constructed for the mean value of Y for $X_8 = 32$, $X_6 = 22$, 95% of these intervals will contain the true mean value of Y at $X_8 = 32$, $X_6 = 22$. From a practical point of view we can say that there is a 0.95 probability that the statement, the true mean value of Y at $X_8 = 32$, $X_6 = 22$ lies between 10.8318 and 11.7154, is correct.

Confidence Limits for the Mean of *g* Observations Given a Specific Set of *X*'s

These limits are calculated from

$$\hat{Y} \pm t(\nu, 1 - \tfrac{1}{2}\alpha) \cdot s\sqrt{1/g + \mathbf{X}_0'\mathbf{C}\mathbf{X}_0}.$$

For example, the 95% confidence limits for an individual observation for the point $(X_8 = 32, X_6 = 22)$ are

$$\begin{aligned}\hat{Y} \pm t(22, 0.975) \cdot s\sqrt{1 + \mathbf{X}_0'\mathbf{C}\mathbf{X}_0} &= 11.2736 \pm (2.074)(0.661589)\sqrt{1 + 0.10413981}\\ &= 11.2736 \pm (2.074)(0.661589)(1.050781)\\ &= 11.2736 \pm 1.4418\\ &= 9.8318, 12.7154.\end{aligned}$$

Note: To obtain simultaneous confidence surfaces appropriate for the whole regression function over its entire range, it would be necessary to replace $t(\nu, 1 - \tfrac{1}{2}\alpha)$ by $\{pF(p, n - p, 1 - \alpha)\}^{1/2}$, where p is the total number of parameters in the model including β_0. (Currently, $\nu = n - p$. In the example, $n = 25, p = 3$.) See, for example, Miller (1981).

Examining the Residuals

The residuals shown in Table 6.1 could be examined to see if they provide any indication that the model is inadequate. We leave this as an exercise, except for the following comments:

1. Residual versus $\hat{Y}$ plot (Figure 6.2). No unusual behavior is indicated.
2. The runs test and the Durbin-Watson test indicated no evidence of time-dependent nonrandomness. (See also Exercise A, in "Exercises for Chapter 7.")

6.3. SUM OF SQUARES OF A SET OF LINEAR FUNCTIONS OF *Y*'s

In some applications, for example, two-level factorial designs, the items of interest are *contrasts*, that is, linear combinations of Y's whose coefficients add to zero, so

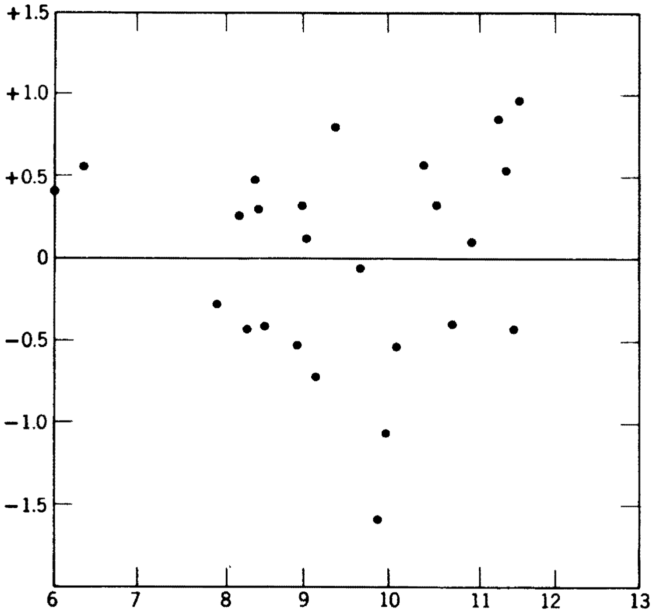

Figure 6.2. Residual versus $\hat{Y}$ plot.

that part of the data is "contrasted" with another part. It is often puzzling in such circumstances to know what is the sum of squares attributable to such a contrast. In regression work, the parameter estimates are also linear combinations of the observations, although not contrasts. We give here a rule that is foolproof for getting an appropriate sum of squares of any set of linear functions $\mathbf{C}'\mathbf{Y}$, say, where $\mathbf{C}'$ is an $m \times n$ matrix. The correct answer emerges even if the set contains duplicated linear functions or linear functions that are linear combinations of other linear functions!

Let $\mathbf{C}'\mathbf{Y}$ be a set of linear functions of the observations $\mathbf{Y}$. Then the sum of squares due to $\mathbf{C}'\mathbf{Y}$ is defined as follows:

$$\mathrm{SS}(\mathbf{C}'\mathbf{Y}) = \mathbf{z}'\mathbf{C}'\mathbf{Y} = \mathbf{Y}'\mathbf{C}\mathbf{z}, \tag{6.3.1}$$

where $\mathbf{z}$ is *any* solution of the equations

$$\mathbf{C}'\mathbf{C}\mathbf{z} = \mathbf{C}'\mathbf{Y}. \tag{6.3.2}$$

Sometimes $\mathbf{z}$ is unique, sometimes not. Nevertheless, the resulting sum of squares is always unique. In the above, $\mathbf{C}'\mathbf{Y}$ is $m \times 1$, $\mathbf{C}'$ is $m \times n$, $\mathbf{Y}$ is $n \times 1$, and $\mathbf{z}$ is $m \times 1$.

Special Case m = 1. Let $\mathbf{C}' = \mathbf{c}'$, a $1 \times n$ row vector. Then

$$\mathrm{SS}(\mathbf{c}'\mathbf{Y}) = (\mathbf{c}'\mathbf{Y})^2/\mathbf{c}'\mathbf{c}. \tag{6.3.3}$$

General Nonadditivity of SS. Suppose that

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C}_1' \\ \\ \mathbf{C}_2' \end{bmatrix} \tag{6.3.4}$$

where $\mathbf{C}_1'$ is $m_1 \times n$ and $\mathbf{C}_2'$ is $m_2 \times n$, where $m_1 + m_2 = m$. Then

$$\mathrm{SS}(\mathbf{C}'\mathbf{Y}) = \mathrm{SS}(\mathbf{C}_1'\mathbf{Y}) + \mathrm{SS}(\mathbf{C}_2'\mathbf{Y}) \tag{6.3.5}$$

if and only if $\mathbf{C}_1'\mathbf{C}_2 = \mathbf{0}$, that is, if and only if all the rows of $\mathbf{C}_1'$ are orthogonal to all the rows of $\mathbf{C}_2'$. (This can also be expressed as "all the columns of $\mathbf{C}_1$ are orthogonal to all the columns of $\mathbf{C}_2$.")

Example 1. Find the sum of squares due to the least squares estimates $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in the nonsingular case.

Here $\mathbf{C}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so that $\mathbf{z}$ is any solution to

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

which is clearly given (uniquely) by $\mathbf{z} = \mathbf{X}'\mathbf{Y}$. Thus

$$\text{SS}(\mathbf{b}) = (\mathbf{Y}'\mathbf{C})\mathbf{z} = \mathbf{b}'\mathbf{X}'\mathbf{Y},$$

the familiar formula.

Example 2. Find SS($\overline{Y}$). We see that $\overline{Y} = \Sigma Y_i/n = \mathbf{c}'\mathbf{Y}$, where

$$\mathbf{c}' = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

Thus $\mathbf{c}'\mathbf{c} = 1/n$ and $\text{SS}(\overline{Y}) = \overline{Y}^2/(1/n) = n\overline{Y}^2$.

Example 3. For a straight line fit, find SS(b_1). We write $b_1 = \mathbf{c}'\mathbf{Y}$, where the ith element of $\mathbf{c}'$ is $(X_i - \overline{X})/S_{XX}$. Thus $\mathbf{c}'\mathbf{c} = 1/S_{XX}$ and $\text{SS}(b_1) = S_{XX}b_1^2 = S_{XY}^2/S_{XX}$.

Example 4. For a straight line fit, find SS(b_0). Now $b_0 = \overline{Y} - b_1\overline{X}$ so that the ith element of $\mathbf{c}'$ is

$$c_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{S_{XX}}$$

and

$$\mathbf{c}'\mathbf{c} = \sum_{i=1}^{n}\left\{\frac{1}{n^2} - \frac{2\overline{X}(X_i - \overline{X})}{nS_{XX}} + \frac{\overline{X}^2(X_i - \overline{X})^2}{S_{XX}^2}\right\},$$
$$= \frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}$$

and so

$$\text{SS}(b_0) = b_0^2\bigg/\left\{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right\}$$

Note: This is not the usual SS(b_0) in regression tables, because the usual $\text{SS}(b_0) = n\overline{Y}^2$ is really SS($\overline{Y}$), that is, the SS of the b_0 we would get from the model $\mathbf{Y} = \beta_0 + \epsilon$. The SS($b_0$) of this example is

$$\text{SS}(b_0|b_1) \text{ for the model } Y = \beta_0 + \beta_1 X + \epsilon.$$

Examples 2, 3, and 4 provide important clues as to what formula (6.3.1) produces. It is always the *extra sum of squares*. For the general case, we write

$$\mathbf{Y} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\epsilon},$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}(\mathbf{X}_1, \mathbf{X}_2) = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'X_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix},$$

say. We write the inverse as

$$\begin{bmatrix}\mathbf{C}^{11} & \mathbf{C}^{12}\\ \mathbf{C}^{21} & \mathbf{C}^{22}\end{bmatrix}$$

so that

$$\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix} = \begin{bmatrix}\mathbf{C}^{11}\mathbf{X}_1' + \mathbf{C}^{12}\mathbf{X}_2'\\ \mathbf{C}^{21}\mathbf{X}_1' + \mathbf{C}^{22}\mathbf{X}_2'\end{bmatrix}\mathbf{Y} = \begin{bmatrix}\mathbf{D}_1'\mathbf{Y}\\ \mathbf{D}_2'\mathbf{Y}\end{bmatrix}, \text{say.}$$

If formulas (6.3.1) and (6.3.2) are applied to $\mathbf{b}_2$ the result will be the extra $\text{SS}(\mathbf{b}_2|\mathbf{b}_1)$. We omit the proof, which is intricate but not difficult. It requires use of the second inverse in Result 7, Appendix 5A, and involves showing that $\mathbf{D}_2'\mathbf{D}_2 = \mathbf{Q}^{-1}$, where $\mathbf{Q} = \mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2 = \mathbf{X}_2'\mathbf{Z}$, say, where $\mathbf{Z}$ is the residual matrix when $\mathbf{X}_2$ is regressed on $\mathbf{X}_1$. Thus $\mathbf{Z}$ is "the part of $\mathbf{X}_2$ orthogonal to $\mathbf{X}_1$," and it is easy to see that $\mathbf{X}_1'\mathbf{Z} = 0$. It then follows that $\mathbf{z} = \mathbf{Q}\mathbf{D}_2'\mathbf{Y}$ and that $\mathbf{z}'\mathbf{D}_2\mathbf{Y} = \mathbf{Y}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. This is one form of the extra sum of squares.

APPENDIX 6A. ORTHOGONAL COLUMNS IN THE X MATRIX

Suppose we have a regression problem involving parameters β_0, β_1, and β_2. Using the extra sum of squares principle we can calculate a number of quantities such as:

$\text{SS}(b_2)$	from the model	$Y = \beta_2 X_2 + \epsilon$
$\text{SS}(b_2\|b_0)$	from the model	$Y = \beta_0 + \beta_2 X_2 + \epsilon$
$\text{SS}(b_2\|b_0, b_1)$	from the model	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

These will usually have completely different numerical values except when the "β_2" column of the $\mathbf{X}$ matrix is orthogonal to the "β_0" and the "β_1" columns. When this happens we can unambiguously talk about "$\text{SS}(b_2)$." We now examine this situation in more detail.

Suppose in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ we divide the matrix $\mathbf{X}$ up into t sets of columns denoted in matrix form by

$$\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t\}.$$

A corresponding division can be made in $\boldsymbol{\beta}$ so that

$$\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_t\end{bmatrix},$$

where the number of columns in $\mathbf{X}_i$ is equal to the number of rows in $\boldsymbol{\beta}_i$, $i = 1, 2, \ldots, t$. The model can then be written

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \cdots + \mathbf{X}_t\boldsymbol{\beta}_t.$$

Suppose that

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_t \end{bmatrix}$$

is the vector estimate of $\boldsymbol{\beta}$ for this model (and given data) obtained from the normal equations

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

Result. If the columns of $\mathbf{X}_i$ are orthogonal to the columns of $\mathbf{X}_j$ for all $i, j = 1, 2, \ldots, t (i \neq j)$, that is, if $\mathbf{X}_i'\mathbf{X}_j = \mathbf{0}$, it is true that

$$\begin{aligned} \mathrm{SS}(\mathbf{b}) &= \mathrm{SS}(\mathbf{b}_1) + \mathrm{SS}(\mathbf{b}_2) + \cdots + \mathrm{SS}(\mathbf{b}_t) \\ &= \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} + \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y} + \cdots + \mathbf{b}_t'\mathbf{X}_t'\mathbf{Y} \end{aligned}$$

and $\mathbf{b}_i$ is the least square estimate of $\boldsymbol{\beta}_i$, and $\mathrm{SS}(\mathbf{b}_i) = \mathbf{b}_i'\mathbf{X}_i'\mathbf{Y}$ *whether any of the other terms are in the model or not.* Thus

$$\mathrm{SS}(\mathbf{b}_i) = \mathrm{SS}(\mathbf{b}_i | \text{any set of } \mathbf{b}_j, j \neq i).$$

(Note that it is *not* necessary for the columns of $\mathbf{X}_i$ to be orthogonal to *each other*—only for the $\mathbf{X}_i$ columns all to be orthogonal to all other columns of $\mathbf{X}$.)

We consider the case $t = 2$. Here

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2),$$

where $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2'\mathbf{X}_1 = \mathbf{0}$. (This means that all the columns in $\mathbf{X}_1$ are orthogonal to all the columns in $\mathbf{X}_2$.) We can write the model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon,$$

where $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ is split into the two sets of coefficients, which correspond to the $\mathbf{X}_1$ and $\mathbf{X}_2$ sets of columns. The normal equations are $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$; that is,

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{bmatrix},$$

where a split in $\mathbf{b}$ corresponding to that in $\boldsymbol{\beta}$ has been made. Since the off-diagonal terms $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, $\mathbf{X}_2'\mathbf{X}_1 = \mathbf{0}$, the normal equations can be split into the two sets of equations

$$\mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1 = \mathbf{X}_1'\mathbf{Y}; \qquad \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{Y}$$

with solutions

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}; \qquad \mathbf{b}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y,$$

assuming that the matrices shown inverted are nonsingular. Thus $\mathbf{b}_1$ is the least squares estimate of $\boldsymbol{\beta}_1$ whether $\boldsymbol{\beta}_2$ is in the model or not, and vice versa. Now

$$\mathrm{SS}(\mathbf{b}_1) = \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} \quad \text{and} \quad \mathrm{SS}(\mathbf{b}_2) = \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y}.$$

Thus

$$\begin{aligned}\mathrm{SS}(\mathbf{b}_1, \mathbf{b}_2) &= \mathbf{b}'\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{b}_1', \mathbf{b}_2')(\mathbf{X}_1, \mathbf{X}_2)'\mathbf{Y} \\ &= (\mathbf{b}_1', \mathbf{b}_2')\begin{pmatrix}\mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y}\end{pmatrix} \\ &= \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} + \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y} \\ &= \mathrm{SS}(\mathbf{b}_1) + \mathrm{SS}(\mathbf{b}_2).\end{aligned}$$

It follows that

$$\mathrm{SS}(\mathbf{b}_1|\mathbf{b}_2) = \mathrm{SS}(\mathbf{b}_1, \mathbf{b}_2) - \mathrm{SS}(\mathbf{b}_2) = \mathrm{SS}(\mathbf{b}_1).$$

Similarly,

$$\mathrm{SS}(\mathbf{b}_2|\mathbf{b}_1) = \mathrm{SS}(\mathbf{b}_2)$$

and this depends only on the orthogonality of $\mathbf{X}_1$ and $\mathbf{X}_2$. The extension to cases where $t > 2$ is immediate.

APPENDIX 6B. TWO PREDICTORS: SEQUENTIAL SUMS OF SQUARES

We saw in the example of Section 6.2 that the SS(b_8, $b_6|b_0$) splits differently according to the sequence of entry.

SS Contribution of	First Predictor in Is	
Predictor Below When:	X_8	X_6
X_8	45.59	35.84
X_6	8.59	18.34
Totals	54.18	54.18

This type of split-up where one variable (here X_8) has the larger SS whether it enters first or second happens quite frequently but other cases can occur, some of them quite strange. Schey (1993) has discussed and illustrated this point using seven "contrived examples."

Table 6B.1 shows the seven data sets and Table 6B.2 shows the disposition of the regression sums of squares using the same formation as above. (Our recomputations vary slightly from Schey's numbers, but the point is not affected.) The model fitted is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

TABLE 6B.1. Seven Example Sets of Data Devised by Schey (1993)

Set	X_1	X_2	Y	Set	X_1	X_2	Y
1	1.80	12.80	3.71	5	4.66	0.56	6.23
	8.90	12.21	12.29		8.05	−3.94	5.66
	4.76	15.46	−0.79		9.61	−2.86	4.16
	1.86	8.93	0.90		8.59	−1.63	3.43
	2.69	3.92	3.04		3.47	0.22	−2.46
	1.36	1.64	−4.00		4.04	−1.13	−0.41
	7.84	16.07	10.33		0.63	2.25	0.15
	2.79	9.56	4.80		5.91	−0.90	1.18
	3.94	15.55	6.03		8.18	−1.87	8.44
	4.31	3.14	0.86		4.38	−1.10	1.55
2	1.83	−4.20	−0.53	6	0.23	2.12	1.01
	7.17	3.55	7.20		0.15	5.41	−3.73
	6.18	1.57	5.55		5.03	1.90	11.00
	9.44	−2.64	3.02		7.99	−3.36	11.17
	0.86	12.27	−0.56		1.08	2.68	0.32
	0.34	−5.94	−6.36		3.24	4.03	6.65
	5.02	2.35	8.64		9.41	−1.84	7.52
	9.98	5.47	12.24		6.34	−1.62	0.86
	2.00	12.21	5.07		8.17	−4.15	1.55
	7.92	8.14	9.86		5.00	1.48	6.24
3	9.09	−3.97	31.60	7	0.26	2.36	−1.81
	4.09	37.43	21.21		3.74	0.45	1.72
	0.97	−17.97	−4.36		8.46	−1.43	8.13
	1.45	24.94	10.18		9.27	−4.54	1.62
	0.73	7.61	−0.54		0.67	2.45	−1.06
	3.31	−6.93	9.53		9.51	−5.07	7.89
	7.97	19.45	28.53		8.91	−5.49	1.97
	9.81	−41.36	16.77		3.97	3.31	6.69
	0.79	−8.44	−5.94		5.77	−2.69	1.02
	2.15	3.56	2.03		3.50	3.95	8.43
4	3.82	13.36	11.77				
	8.06	20.41	10.79				
	0.01	−1.42	1.27				
	3.27	3.92	1.55				
	5.10	7.49	3.19				
	8.22	12.69	−1.79				
	5.49	15.16	7.86				
	2.98	15.02	9.52				
	8.97	19.46	3.75				
	9.48	12.14	−3.53				

which can be fitted in the form

$$Y - \overline{Y} = \beta_1(X_1 - \overline{X}_1) + \beta_2(X_2 - \overline{X}_2) + \epsilon$$

with identical results. If the columns $X_1 - \overline{X}_1$ and $X_2 - \overline{X}_2$ were orthogonal, the entry order would not matter, as in Appendix 6A.

We see the following:

TABLE 6B.2. Sequential Sums of Squares in Seven Examples Devised by Schey (1993)

Data Set Number	SS Contribution of Predictor Below When:	First Predictor in Is	
		X_1	X_2
1	X_1	134.10	71.89
	X_2	9.62	71.83
2	X_1	169.12	169.09
	X_2	56.34	56.37
3	X_1	1162.89	1446.14
	X_2	387.22	103.14
4	X_1	14.70	164.41
	X_2	204.51	54.80
5	X_1	43.94	23.60
	X_2	3.16	23.50
6	X_1	73.87	74.21
	X_2	24.85	24.51
7	X_1	28.83	115.16
	X_2	86.33	0.00

1. Case 1 shows X_1 very important when it enters first, and X_1 and X_2 are equally important in the reverse order.
2. Case 2 has the vectors of $(X_1 - \overline{X}_1)$ and $(X_2 - \overline{X}_2)$ orthogonal, so the entry order is irrelevant, as in Appendix 6A.
3. Each variable contributes more when it comes in second than it does when it comes in first! When $\text{SS}(b_2|b_1, b_0) > \text{SS}(b_2|b_0)$, X_1 is said to be a suppressor variable. Here, both X_1 and X_2 are suppressor variables.
4. This case is similar to the third, but it differs in certain geometrical aspects discussed in the reference, aspects which we have omitted.
5. Similar to 1, but different geometry.
6. A remarkable case because the vectors of $(X_1 - \overline{X}_1)$ and $(X_2 - \overline{X}_2)$ are not orthogonal, but the sums of squares are essentially unchanged. So orthogonality is sufficient, but not necessary, for this SS behavior.
7. An extreme case that is unlikely to arise from real data. Variable X_2 contributes nothing when it goes in first, but accounts for $\text{SS}(b_2|b_1, b_0) = 86.33$ when it goes in second!

References

Freund (1988); Hamilton (1987a, b); Mitra (1988); Schey (1993).

EXERCISES FOR CHAPTERS 5 AND 6

A. Consider the data in the following table:

X_0	X_1	X_2	Y
1	1	8	6
1	4	2	8
1	9	−8	1
1	11	−10	0
1	3	6	5
1	8	−6	3
1	5	0	2
1	10	−12	−4
1	2	4	10
1	7	−2	−3
1	6	−4	5

Requirements

1. Using least squares procedures, estimate the β's in the model:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

2. Write out the analysis of variance table.
3. Using $\alpha = 0.05$, test to determine if the overall regression is statistically significant.
4. Calculate the square of the multiple correlation coefficient, namely, R^2. What portion of the total variation about $\overline{Y}$ is explained by the two variables?
5. The inverse of the **X′X** matrix for this problem is as follows:

$$\begin{bmatrix} 4.3705 & -0.8495 & -0.4086 \\ -0.8495 & 0.1690 & 0.0822 \\ -0.4086 & 0.0822 & 0.0422 \end{bmatrix}.$$

Using the results of the analysis of variance table with this matrix, calculate estimates of the following:

a. Variance of b_1.
b. Variance of b_2.
c. The variance of the predicted value of Y for the point $X_1 = 3$, $X_2 = 5$.

6. How useful is the regression using X_1 alone? What does X_2 contribute, given that X_1 is already in the regression?
7. How useful is the regression using X_2 alone? What does X_1 contribute, given that X_2 is already in the regression?
8. What are your conclusions?

B. The table below gives 12 sets of observations on three variables X, Y, and Z. Find the regression plane of X on Y and Z—that is, the linear combination of Y and Z that best predicts the value of X when only Y and Z are given. By constructing an analysis of variance table for X, or otherwise, test whether it is advantageous to include both Y and Z in the prediction formula.

X	Y	Z
1.52	98	77
1.41	76	139
1.16	58	179
1.45	94	95
1.24	73	142
1.21	57	186
1.63	97	82
1.38	91	100
1.37	79	125

X	Y	Z
1.36	92	96
1.40	92	99
1.03	54	190

Source: Cambridge Diploma, 1949.

C. The data below are selected from a much larger body of data referring to candidates for the General Certificate of Education who were being considered for a special award. Here, Y denotes the candidate's total mark, out of 1000, in the G.C.E. examination. Of this mark the subjects selected by the candidate account for a maximum of 800; the remainder, with a maximum of 200, is the mark in the compulsory papers—"General" and "Use of English"—this mark is shown as X_1. X_2 denotes the candidate's mark, out of 100, in the compulsory School Certificate English Language paper taken on a previous occasion.

Compute the multiple regression of Y on X_1 and X_2, and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory papers may be used to predict aggregate performance in the G.C.E. examination, and on whether previous performance in School Certificate English Language has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

Candidate	Y	X_1	X_2	Candidate	Y	X_1	X_2
1	476	111	68	9	645	117	59
2	457	92	46	10	556	94	97
3	540	90	50	11	634	130	57
4	551	107	59	12	637	118	51
5	575	98	50	13	390	91	44
6	698	150	66	14	562	118	61
7	545	118	54	15	560	109	66
8	574	110	51				

Source: Cambridge Diploma, 1953. (Exercises B and C are published with permission of Cambridge University Press.)

D. Eight runs were made at various conditions of saturation (X_1) and transisomers (X_2). The response, SCI, is listed below as Y for the corresponding levels of X_1 and X_2.

Y	X_1	X_2
66.0	38	47.5
43.0	41	21.3
36.0	34	36.5
23.0	35	18.0
22.0	31	29.5
14.0	34	14.2
12.0	29	21.0
7.6	32	10.0

1. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
2. Is the overall regression significant? (Use $\alpha = 0.05$.)
3. How much of the variation in Y about $\bar{Y}$ is explained by X_1 and X_2?

E. The effect of sealer plate temperature and sealer plate clearance in a soap wrapping machine affects the percentage of wrapped bars that pass inspection. Some data on these variables were collected and are shown as follows:

Sealer Plate Clearance, X_1	Sealer Plate Temperature, X_2	% Sealed Properly, Y
130	190	35.0
174	176	81.7
134	205	42.5
191	210	98.3
165	230	52.7
194	192	82.0
143	220	34.5
186	235	95.4
139	240	56.7
188	230	84.4
175	200	94.3
156	218	44.3
190	220	83.3
178	210	91.4
132	208	43.5
148	225	51.7

Requirements

1. Assume a linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ and determine least squares estimates of β_0, β_1, and β_2.
2. Is the overall regression significant? (Use $\alpha = 0.05$.)
3. Is one of the two variables more useful than the other in predicting the percentage sealed properly?
4. What recommendations would you make concerning the operation of the wrapping machine?

F. Using the 17 observations given below:

1. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
2. Test for lack of fit, using pure error.
3. Examine the residuals.
4. Assess the value of including each of the variables X_1 and X_2 in the regression model.

X_1	X_2	Y
17	42	90
19	45	71,76
20	29	63, 63, 80, 80
21	93	80, 64, 82, 66
25	34	75, 82
27	98	99
28	9	73
30	73	67, 74

G. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to the data below. Check for lack of fit via both pure error and the examination of residuals. Assess the value of including each of the predictors X_1 and X_2 in the regression model.

X_1	X_2	Y
2.6	3.9	83
2.8	4.2	64
2.8	4.2	69
2.9	2.6	56
2.9	2.6	56
2.9	2.6	73

X_1	X_2	Y
2.9	2.6	73
3.0	9.0	57
3.0	9.0	59
3.0	9.0	73
3.0	9.0	75
3.4	3.1	68
3.4	3.1	75
3.6	9.5	92
3.7	0.6	66
3.9	7.0	60
3.9	7.0	67

H. (Please refer to Exercise LL in "Exercises for Chapters 1–3" first.) After he had analyzed the original data, the manager rechecked the records to try to find additional information that would improve his model. He drew out the facts that the numbers of men working at any one time were, for the listing of days shown originally, as follows:

$$Z = 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 3, 6, 6.$$

Fit a planar model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$ to the entire data set via least squares, check for lack of fit, and (if there is no lack of fit) test for overall regression. Also test $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$ using the extra sum of squares principle. What conclusions do you draw from your analysis?

Useful Facts.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 13 & 65 & 17 \\ 65 & 437 & 155 \\ 17 & 155 & 83 \end{bmatrix}^{-1} = \frac{1}{24{,}780}\begin{bmatrix} 12{,}246 & -2{,}760 & 2{,}646 \\ -2{,}760 & 790 & -910 \\ 2{,}646 & -910 & 1{,}456 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 2{,}990 \\ 19{,}120 \\ 6{,}050 \end{bmatrix}, \qquad \mathbf{b} = \frac{1}{24{,}780}\begin{bmatrix} -147{,}360 \\ 1{,}346{,}900 \\ -678{,}860 \end{bmatrix} = \begin{bmatrix} -5.947 \\ 54.354 \\ -27.396 \end{bmatrix}.$$

I. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to the data below, provide an analysis of variance table, and perform the partial F-tests to test $H_0: \beta_i = 0$ versus $H_1: \beta_i \neq 0$ for $i = 1, 2$, given that the other variable is already in the model. Comment on the relative contributions of the variables X_1 and X_2 depending on whether they enter the model first or second.

X_1	X_2	Y
−5	5	11
−4	4	11
−1	1	8
2	−3	2
2	−2	5
3	−2	5
3	−3	4

J. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to the data below. After testing for lack of fit, find the appropriate extra SS F-statistic for testing $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$, and find its degrees of freedom. Relate this F-statistic numerically to the t-statistic typically used to test the same hypothesis.

X_1	X_2	Y
−1	−1	8
−1	1	13
−1	1	12
−1	1	11
1	−1	9
1	−1	8
1	−1	7
1	1	13
0	0	11
0	0	13

K. The questions below relate to fitting the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to the following data:

	X_1	X_2	Y
	−1	−1	7.2
	−1	0	8.1
	0	0	9.8
	1	0	12.3
	1	1	12.9
Sum	0	0	50.3
Sum of squares	4	2	531.19

1. Write down the normal equation $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}$ in matrix format.
2. Obtain the solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ using matrix manipulations.
3. Find $\text{SS}(b_0, b_1, b_2)$ via matrix manipulations.
4. Find the residual sum of squares, and obtain s^2.
5. Evaluate $\text{se}(b_i)$, $i = 0, 1, 2$.
6. Find $\hat{Y}_0$ at the point $(X_{10}, X_{20}) = (0.5, 0)$.
7. Obtain $\text{se}(\hat{Y}_0)$.
8. Find $\text{SS}(b_2|b_1, b_0)$.
9. You are now told that the 9.8 value at (0, 0) in the data is the average of four observations so that the variance of this Y is $\sigma^2/4$ and not σ^2. However, all observations are still independent. Your informant says that you should have used weighted least squares $\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, where $\mathbf{V}$ is a diagonal matrix here, to get your estimates. Provide the new (weighted least squares) values of b_0, b_1, and b_2. (See Section 9.2.)

L. For the experimental situation described in Exercise DD of "Exercises for Chapters 1–3," suppose that data for the concentratioin of chemical A had been recorded for each run. It is suggested that the variation found in this factor might be causing the large variation in response discovered previously. The readings of the concentration factor C (in percent) are:

Batch	1	2	3	4	5	6	7	8	9
C	6	6	8	7	9	8	5	9	11

where the batch numbers are the same as in Exercise DD.

1. Plot the residuals obtained in Exercise DD versus C. Notice anything?

2.* Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to the data, where $X_1 = (T - 300)/10$ and $X_2 = C - 8$.

3. Perform an analysis of variance and test:
   a. The lack of fit.
   b. The significance of the effect of including β_1 and β_2 in the model, rather than only β_0.
   c. The significance of including β_2 in the model, rather than just β_0 and β_1.

4. What percentage of the total variation (corrected) has been "taken up" by including the concentration effect in our model?

5. What is the standard error of $\tilde{b}_1$? Of $\tilde{b}_2$? ($\tilde{b}_1$ and $\tilde{b}_2$ are the regression coefficients in the expression for $\hat{Y}$ in terms of the original variables T and C.)

6. Write the fitted value and the residual for each batch. Notice anything?

7. What is the estimated variance of the predicted value $\hat{Y}$ at the point $T = 315$, $C = 8$?

M. Using the data given in Table 1.1, find a joint 90% confidence region for (β_0, β_1). $[F(2, 23, 0.90) = 2.55.]$

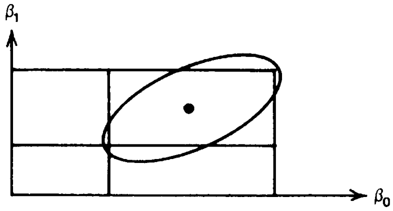

Figure M.1.

On an accurate figure (Figure M1 shows the appropriate format) draw the following:

1. The estimated point (b_0, b_1).
2. The 90% confidence contour for (β_0, β_1).
3. The 95% confidence intervals for β_0 and β_1, separately, and the rectangle these jointly indicate.

Comment briefly on your results.

Hints: (a) Equation (5.3.7), written out, is a quadratic equation in β_0 and β_1. To get the ellipse, set a value of β_0 and solve the resulting quadratic for two values of β_1 to get an upper and lower point on the ellipse (see Figure M2). Imaginary roots mean that your selected β_0 value is outside the ellipse (see Figure M3). Repeat for several values of β_0 and join up the points. It is easiest to use the computer for the calculations.

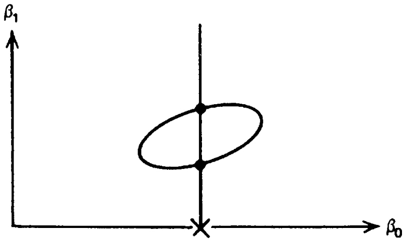

Figure M.2.

*If you employ the coding system $X_1 = (T - 300)/10$, $X_2 = C - 8$, you may use the following hint:

$$\begin{bmatrix} 9 & -2 & -3 \\ -2 & 46 & 13 \\ -3 & 13 & 29 \end{bmatrix}^{-1} = \frac{1}{10111}\begin{bmatrix} 1165 & 19 & 112 \\ 19 & 252 & -111 \\ 112 & -111 & 410 \end{bmatrix}.$$

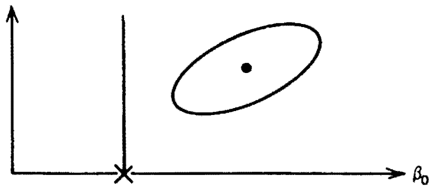

Figure M.3.

(b) Because the two intervals are 95% ones, the rectangle is a sort of (0.95)(0.95) = 0.9025 or $90\frac{1}{4}$% joint region—an incorrect one, we know, but that is the probability level. Thus we compare it with a 90% true region, the nearest we can easily look up from the F-tables.

N. Show that the square of the multiple correlation coefficient R^2 is equal to the square of the correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$

O. Consider the formal regression of the residuals e_i onto a quadratic function $\alpha_0 + \alpha_1\hat{Y}_i + \alpha_2\hat{Y}_i^2$ of the fitted values $\hat{Y}_i$, by least squares. Show that all three estimated coefficients depend on $T_{12} = \Sigma e_i\hat{Y}_i^2$. What does this imply?

P. We fit a straight line model to a set of data using the formulas $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, $\hat{\mathbf{Y}} = \mathbf{Xb}$ with the usual definitions. We define $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Show that

$$\begin{aligned}\text{SS(due to regression)} &= \mathbf{Y}'\mathbf{HY}\\ &= \hat{\mathbf{Y}}'\hat{\mathbf{Y}}\\ &= \hat{\mathbf{Y}}'\mathbf{H}^3\mathbf{Y}.\end{aligned}$$

Q. Show that $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

R. Show that, for any linear model

$$\sum_{i=1}^{n} V(\hat{Y}_i)/n = \text{trace}\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\sigma^2/n = p\sigma^2/n.$$

S. See Exercise Y in "Exercises for Chapters 1–3." Would that result extend if there were more X's? (Yes.)

T. Suppose $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a regression model containing a β_0 term in the first position, and $\mathbf{1} = (1, 1, \ldots, 1)'$ is an $n \times 1$ vector of ones. Show that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1} = (1, 0, \ldots, 0)'$ and hence that $\mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1} = n$. (*Hint:* $\mathbf{X}'\mathbf{1}$ is the first column of $\mathbf{X}'\mathbf{X}$.) These results can be useful in regression matrix manipulations. For connected reading, see letters in The *American Statistician*, April 1972, 47–48.

U. By noting that $\mathbf{X}_0 = (1, \overline{X}_1, \overline{X}_2, \ldots)'$ can be written as $\mathbf{X}'\mathbf{1}/n$, and applying the result in Exercise T above, show that $V(\hat{Y})$ at the point $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)$ is σ^2/n.

V. Look again at the (X, Y_2) data of Exercise E in "Exercises for Chapters 1–3." Fit the quadratic model

$$Y = \beta_0 + \beta_1 X + \beta_{11}X^2 + \epsilon$$

to these data and provide the usual subsidiary analyses. Draw the fitted curve on a plot of the data and estimate the abscissa value at the minimum point of the curve. [*Hint:* It is at $-b_1/(2b_{11})$.] What does your conclusion mean?

[*Note:* You may have to code the data if your intended regression is frustrated by the computer. If that happens, why does it happen? A suggested coding is $U = (X - 0.048)/0.048$. You do not have to use this if you do not wish to!]

W. Look at Appendix 6B. Perform at least one of the regressions yourself and check the results against those given.

X. Show that, in the general linear regression situation with a β_0 term in the model:

1. The correlation between the vectors $\mathbf{e}$ and $\mathbf{Y}$ is $(1 - R^2)^{1/2}$. The implication of this result is that it is a mistake to attempt to find defective regressions by a plot of residuals e_i versus observations Y_i as this will always show a slope.

2. Show that this slope is $1 - R^2$.
3. Show, further, that the correlation between $\mathbf{e}$ and $\hat{\mathbf{Y}}$ is zero.

Y. Four levels, coded as -3, -1, 1, and 3, were chosen for each of two variables X_1 and X_2, to provide a total of 16 experimental conditions when all possible combinations (X_1, X_2) were taken. It was decided to use the resulting 16 observations to fit a regression equation including a constant term, all possible first-order, second-order, third-order, and fourth-order terms in X_1 and X_2. The data were fed into a computer routine, which usually obtains a vector estimate

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The computer refused to obtain the estimates. Why?

The experimeter, who had meanwhile examined the data, decided at this stage to ignore the levels of variable X_2 and fit a fourth-order model in X_1 only to the *same* observations. The computer again refused to obtain the estimates. Why?

Z. The cloud point of a liquid is a measure of the degree of crystallization in a stock that can be measured by the refractive index. It has been suggested that the percentage of I-8 in the base stock is an excellent predictor of cloud point using the second-order model:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon.$$

The following data were collected on stocks with known percentage of I-8.

% I-8, X	Cloud Point, Y	% I-8, X	Cloud Point, Y
0	22.1	2	26.1
1	24.5	4	28.5
2	26.0	6	30.3
3	26.8	8	31.5
4	28.2	10	33.1
5	28.9	0	22.8
6	30.0	3	27.3
7	30.4	6	29.8
8	31.4	9	31.8
0	21.9		

Requirements

1. Determine the best-fitting second-order model.
2. Using $\alpha = 0.05$, check the overall regression.
3. Test for lack of fit.
4. Would the first-order model, $Y = \beta_0 + \beta_1 X + \epsilon$, have been sufficient? Use the residuals from this simpler model to support your conclusions.
5. Comment on the use of the fitted second-order model as a predictive equation.

AA. A certain experiment gives observations $(Y_1, Y_2, Y_3, Y_4) = (4, 2, 1, 5)$. What is the sum of squares of the set of linear functions $L_1 = Y_1 + 2Y_2 + 2Y_3 + Y_4$ and $L_2 = Y_1 - Y_2 - Y_3 + Y_4$? What is the sum of squares of the set of linear functions L_1, L_2, and L_3, where $L_3 = -3Y_1 + 3Y_4$.