9

Regression Analysis—Inference for Curve- and Surface-Fitting

The two previous chapters began a study of inference methods for multisample studies by considering first those which make no explicit use of structure relating several samples and then discussing some directed at the analysis of factorial structure. The discussion in this chapter will primarily consider inference methods for multisample studies where factors involved are inherently quantitative and it is reasonable to believe that some approximate functional relationship holds between the values of the system/input/independent variables and observed system responses. That is, this chapter introduces and applies inference methods for the curve- and surface-fitting contexts discussed in Sections 4.1 and 4.2.

The chapter begins with a discussion of the simplest situation of this type—namely, where a response variable y is approximately linearly related to a single quantitative input variable x. In this specific context, it is possible to give explicit formulas and illustrate in concrete terms what is possible in the way of inference methods for surface-fitting analyses. The second section then treats the general problem of statistical inferences in multiple regression (curve- and surface-fitting) analyses. In the general case, it is not expedient to produce many computational formulas. So the exposition relies instead on summary measures commonly appearing on multiple regression printouts from statistical packages. A final section further illustrates the broad utility of the multiple regression methods by applying them to "response surface," and then factorial, analyses.

9.1 Inference Methods Related to the Least Squares Fitting of a Line (Simple Linear Regression)

This section considers inference methods that are applicable where a response y is approximately linearly related to an input/system variable x. It begins by introducing the (normal) simple linear regression model and discussing how to estimate response variance in this context. Next there is a look at standardized residuals. Then inference for the rate of change $(\Delta y/\Delta x)$ is considered, along with inference for the average response at a given x. There follows a discussion of prediction and tolerance intervals for responses at a given setting of x. Next is an exposition of ANOVA ideas in the present situation. The section then closes with an illustration of how statistical software expedites the calculations introduced in the section.

9.1.1 The Simple Linear Regression Model, Corresponding Variance Estimate, and Standardized Residuals

Chapter 7 introduced the one-way (equal variances, normal distributions) model as the most common probability basis of inference methods for multisample studies. It was represented in symbols as

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{9.1}$$

where the means $\mu_1, \mu_2, \ldots, \mu_r$ were treated as r unrestricted parameters. In Chapter 8, it was convenient (for example) to rewrite equation (9.1) in two-way contexts as

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} \qquad (= \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}) \tag{9.2}$$

where the μ_{ij} are still unrestricted, and to consider restrictions/simplifications of model (9.2) such as

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk} \tag{9.3}$$

Model (9.3) really differs from model (9.2) or (9.1) only in the fact that it postulates a special form or restriction for the means μ_{ij}. Expression (9.3) says that the means must satisfy a parallelism relationship.

Turning now to the matter of inference based on data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ exhibiting an approximately linear scatterplot, one once again proceeds by imposing a restriction on the one-way model (9.1). In words, the model assumptions will be that there are underlying normal distributions for the response y with a

common variance σ^2 but means $\mu_{y|x}$ that change linearly in x. In symbols, it is typical to write that for $i = 1, 2, \ldots, n$,

The (normal) simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{9.4}$$

where the ϵ_i are (unobservable) iid normal $(0, \sigma^2)$ random variables, the x_i are known constants, and β_0, β_1, and σ^2 are unknown model parameters (fixed constants). Model (9.4) is commonly known as **the (normal) simple linear regression model**. If one thinks of the different values of x in an (x, y) data set as separating it into various samples of y's, expression (9.4) is the specialization of model (9.1) where the (previously unrestricted) means of y satisfy the linear relationship $\mu_{y|x} = \beta_0 + \beta_1 x$. Figure 9.1 is a pictorial representation of the "constant variance, normal, linear (in x) mean" model.

Inferences about quantities involving those x values represented in the data (like the mean response at a single x or the difference between mean responses at two different values of x) will typically be sharper when methods based on model (9.4) can be used in place of the general methods of Chapter 7. And to the extent that model (9.4) describes system behavior for values of x not included in the data, a model like (9.4) provides for inferences involving limited interpolation and extrapolation on x.

Section 4.1 contains an extensive discussion of the use of least squares in the fitting of the approximately linear relation

$$y \approx \beta_0 + \beta_1 x \tag{9.5}$$

to a set of (x, y) data. Rather than redoing that discussion, it is most sensible simply to observe that Section 4.1 can be thought of as an exposition of fitting and the use of residuals in model checking for the simple linear regression model (9.4). In

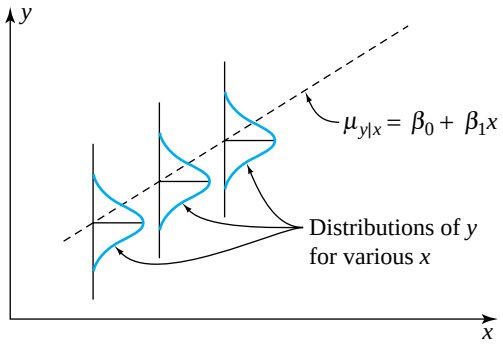

Figure 9.1 Graphical representation of the simple linear regression model

particular, associated with the simple linear regression model are the estimates of β_1 and β_0

Estimator of β_1, the slope

$$b_1 = \frac{\sum(x - x)(y - y)}{\sum(x - x)^2} \tag{9.6}$$

and

Estimator of β_0, the intercept

$$b_0 = \bar{y} - b_1\bar{x} \tag{9.7}$$

and corresponding fitted values

Fitted values for simple linear regression

$$\hat{y}_i = b_0 + b_1 x_i \tag{9.8}$$

and residuals

Residuals for simple linear regression

$$e_i = y_i - \hat{y}_i \tag{9.9}$$

Further, the residuals (9.9) can be used to make up an estimate of σ^2. As always, a sum of squared residuals is divided by an appropriate number of degrees of freedom. That is, there is the following definition of a **(simple linear regression or) line-fitting sample variance**.

Definition 1

For a set of data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where least squares fitting of a line produces fitted values (9.8) and residuals (9.9),

$$s_{\text{LF}}^2 = \frac{1}{n-2}\sum(y - \hat{y})^2 = \frac{1}{n-2}\sum e^2 \tag{9.10}$$

will be called a **line-fitting sample variance**. Associated with it are $\nu = n - 2$ degrees of freedom and an estimated standard deviation of response, $s_{\text{LF}} = \sqrt{s_{\text{LF}}^2}$.

s_{LF}^2 estimates the level of basic background variation, σ^2, whenever the model (9.4) is an adequate description of the system under study. When it is not, s_{LF} will tend to overestimate σ. So comparing s_{LF} to s_{P} is another way of investigating the appropriateness of model (9.4). (s_{LF} much larger than s_{P} suggests the linear regression model is a poor one.)

Example 1
(Example 1, Chapter 4, revisited—page 124)

Inference in the Ceramic Powder Pressing Study

The main example in this section will be the pressure/density study of Benson, Locher, and Watkins (used extensively in Section 4.1 to illustrate the descriptive analysis of (x, y) data). Table 9.1 lists again those $n = 15$ data pairs (x, y) (first presented in Table 4.1) representing

$$x = \text{the pressure setting used (psi)}$$

$$y = \text{the density obtained (g/cc)}$$

in the dry pressing of a ceramic compound into cylinders, and Figure 9.2 is a scatterplot of the data.

Recall further from the calculation of R^2 in Example 1 of Chapter 4 that the data of Table 4.1 produce fitted values in Table 4.2 and then

$$\sum (y - \hat{y})^2 = .005153$$

So for the pressure/density data, one has (via formula (9.10)) that

$$s_{\text{LF}}^2 = \frac{1}{15 - 2}(.005153) = .000396 \text{ (g/cc)}^2$$

so

$$s_{\text{LF}} = \sqrt{.000396} = .0199 \text{ g/cc}$$

If one accepts the appropriateness of model (9.4) in this powder pressing example, for any fixed pressure the standard deviation of densities associated with many cylinders made at that pressure would be approximately .02 g/cc.

Table 9.1
Pressing Pressures and Resultant Specimen Densities

x, Pressure (psi)	y, Density (g/cc)	x, Pressure (psi)	y, Density (g/cc)
2,000	2.486	6,000	2.653
2,000	2.479	8,000	2.724
2,000	2.472	8,000	2.774
4,000	2.558	8,000	2.808
4,000	2.570	10,000	2.861
4,000	2.580	10,000	2.879
6,000	2.646	10,000	2.858
6,000	2.657		

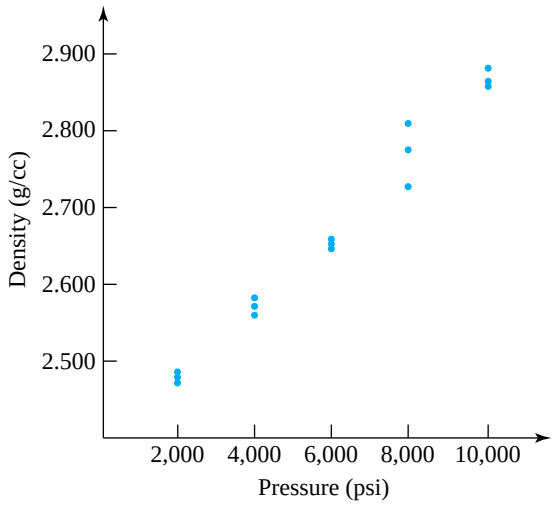

Figure 9.2 Scatterplot of density versus pressing pressure

Table 9.2
Sample Means and Standard Deviations of Densities for Five Different Pressing Pressures

x, Pressure (psi)	$\bar{y}$, Sample Mean	s, Sample Standard Deviation
2,000	2.479	.0070
4,000	2.569	.0110
6,000	2.652	.0056
8,000	2.769	.0423
10,000	2.866	.0114

The original data in this example can be thought of as organized into $r = 5$ separate samples of size $m = 3$, one for each of the pressures 2,000 psi, 4,000 psi, 6,000 psi, 8,000 psi, and 10,000 psi. It is instructive to consider what this thinking leads to for an alternative estimate of σ—namely, s_{P}. Table 9.2 gives $\bar{y}$ and s values for the five samples.

The sample standard deviations in Table 9.2 can be employed in the usual way to calculate s_{P}. That is, exactly as in Definition 1 of Chapter 7

$$s_{\text{P}}^2 = \frac{(3-1)(.0070)^2 + (3-1)(.0110)^2 + \cdots + (3-1)(.0114)^2}{(3-1) + (3-1) + \cdots + (3-1)}$$

$$= .000424 \text{ (g/cc)}^2$$

Example 1 (*continued*)

from which

$$s_P = \sqrt{s_P^2} = .0206 \text{ g/cc}$$

Comparing s_{LF} and s_P, there is no indication of poor fit carried by these values.

Section 4.1 includes some plotting of the residuals (9.9) for the pressure/density data (in particular, a normal plot that appears as Figure 4.7). Although the (raw) residuals (9.9) are most easily calculated, most commercially available regression programs provide standardized residuals as well as, or even in preference to, the raw residuals. (At this point, the reader should review the discussion concerning standardized residuals surrounding Definition 2 of Chapter 7.) In curve- and surface-fitting analyses, the variances of the residuals depend on the corresponding x's. Standardizing before plotting is a way to prevent mistaking a pattern on a residual plot that is explainable on the basis of these different variances for one that is indicative of problems with the basic model. Under model (9.4), for a given x with corresponding response y,

$$\text{Var}(y - \hat{y}) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}\right) \tag{9.11}$$

So using formula (9.11) and Definition 7.2, corresponding to the data pair (x_i, y_i) is the standardized residual for simple linear regression

Standardized residuals for simple linear regression

$$e_i^* = \frac{e_i}{s_{LF}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum (x - \bar{x})^2}}} \tag{9.12}$$

The more sophisticated method of examining residuals under model (9.4) is thus to make plots of the values (9.12) instead of plotting the raw residuals (9.9).

Example 1 (*continued*)

Consider how the standardized residuals for the pressure/density data set are related to the raw residuals. Recalling that

$$\sum (x - \bar{x})^2 = 120{,}000{,}000$$

and that the x_i values in the original data included only the pressures 2,000 psi, 4,000 psi, 6,000 psi, 8,000 psi, and 10,000 psi, it is easy to obtain the necessary values of the radical in the denominator of expression (9.12). These are collected in Table 9.3.

Table 9.3
Calculations for Standardized Residuals in the Pressure/Density Study

x	$\sqrt{1-\frac{1}{15}-\frac{(x-6{,}000)^2}{120{,}000{,}000}}$
2,000	.894
4,000	.949
6,000	.966
8,000	.949
10,000	.894

The entries in Table 9.3 show, for example, that one should expect residuals corresponding to $x = 6{,}000$ psi to be (on average) about $.966/.894 = 1.08$ times as large as residuals corresponding to $x = 10{,}000$ psi. Division of raw residuals by s_{LF} times the appropriate entry of the second column of Table 9.3 then puts them all on equal footing, so to speak. Table 9.4 shows both the raw residuals (taken from Table 4.5) and their standardized counterparts.

In the present case, since the values .894, .949, and .966 are roughly comparable, standardization via formula (9.12) doesn't materially affect conclusions about model adequacy. For example, Figures 9.3 and 9.4 are normal plots of (respectively) raw residuals and standardized residuals. For all intents and purposes, they are identical. So any conclusions (like those made in Section 4.1 based on Figure 4.7) about model adequacy supported by Figure 9.3 are equally supported by Figure 9.4, and vice versa.

In other situations, however (especially those where a data set contains a few very extreme x values), standardization can involve more widely varying denominators for formula (9.12) than those implied by Table 9.3 and thereby affect the results of a residual analysis.

Table 9.4
Residuals and Standardized Residuals for the Pressure/Density Study

x	e	Standardized Residual
2,000	.0137, .0067, −.0003	.77, .38, −.02
4,000	−.0117, .0003, .0103	−.62, .02, .55
6,000	−.0210, −.0100, −.0140	−1.09, −.52, −.73
8,000	−.0403, .0097, .0437	−2.13, .51, 2.31
10,000	−.0007, .0173, −.0037	−.04, .97, −.21

Example 1
(*continued*)

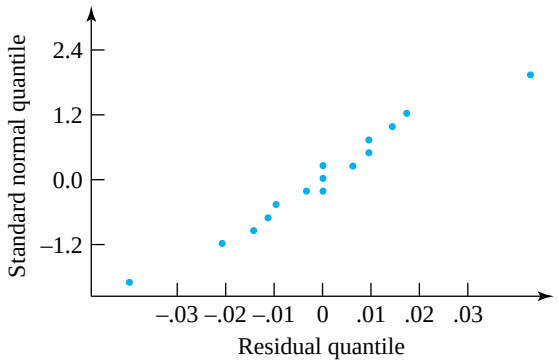

Figure 9.3 Normal plot of residuals for a linear fit to the pressure/density data

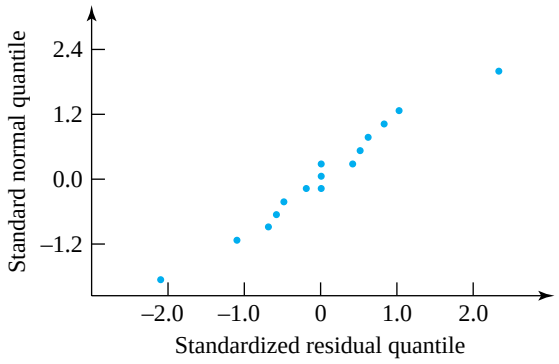

Figure 9.4 Normal plot of standardized residuals for a linear fit to the pressure/density data

9.1.2 Inference for the Slope Parameter

Especially in applications of the simple linear regression model (9.4) where x represents a variable that can be physically manipulated by the engineer, the slope parameter β_1 is of fundamental interest. It is the **rate of change of average response with respect to** x, and it governs the impact of a change in x on the system output. Inference for β_1 is fairly simple, because of the distributional properties that b_1 (the slope of the least squares line) inherits from the model. That is, under model (9.4), b_1 has a normal distribution with

$$Eb_1 = \beta_1$$

and

$$\text{Var}\, b_1 = \frac{\sigma^2}{\sum (x - \bar{x})^2} \qquad \textbf{(9.13)}$$

which in turn imply that

$$Z = \frac{b_1 - \beta_1}{\dfrac{\sigma}{\sqrt{\sum (x - \bar{x})^2}}}$$

is standard normal. In a manner similar to many of the arguments in Chapters 6 and 7, this motivates the fact that the quantity

$$T = \frac{b_1 - \beta_1}{\dfrac{s_{\text{LF}}}{\sqrt{\sum (x - \bar{x})^2}}} \tag{9.14}$$

has a t_{n-2} distribution. The standard arguments of Chapter 6 applied to expression (9.14) then show that

$$\text{H}_0\colon \beta_1 = \# \tag{9.15}$$

can be tested using the test statistic

Test statistic for $H_0\colon \beta_1 = \#$

$$T = \frac{b_1 - \#}{\dfrac{s_{\text{LF}}}{\sqrt{\sum (x - \bar{x})^2}}} \tag{9.16}$$

and a t_{n-2} reference distribution. More importantly, under the simple linear regression model (9.4), a two-sided confidence interval for β_1 can be made using endpoints

Confidence limits for the slope, β_1

$$b_1 \pm t \frac{s_{\text{LF}}}{\sqrt{\sum (x - \bar{x})^2}} \tag{9.17}$$

where the associated confidence is the probability assigned to the interval between $-t$ and t by the t_{n-2} distribution. A one-sided interval is made in the usual way, based on one endpoint from formula (9.17).

Example 1 (*continued*)

In the context of the powder pressing study, Section 4.1 showed that the slope of the least squares line through the pressure/density data is

$$b_1 = .0000486 \text{ (g/cc)/psi}$$

Example 1 (*continued*)

Then, for example, a 95% two-sided confidence interval for β_1 can be made using the .975 quantile of the t_{13} distribution in formula (9.17). That is, one can use endpoints

$$.0000486 \pm 2.160\frac{.0199}{\sqrt{120{,}000{,}000}}$$

that is,

$$.0000486 \pm .0000039$$

that is,

$$.0000448 \text{ (g/cc)/psi} \quad \text{and} \quad .0000526 \text{ (g/cc)/psi}$$

A confidence interval like this one for β_1 can be translated into a confidence interval for a difference in mean responses for two different values of x. According to model (9.4), two different values of x differing by Δx have mean responses differing by $\beta_1 \Delta x$. One then simply multiplies endpoints of a confidence interval for β_1 by Δx to obtain a confidence interval for the difference in mean responses. For example, since $8{,}000 - 6{,}000 = 2{,}000$, the difference between mean densities at 8,000 psi and 6,000 psi levels has a 95% confidence interval with endpoints

$$2{,}000(.0000448) \text{ g/cc} \quad \text{and} \quad 2{,}000(.0000526) \text{ g/cc}$$

that is,

$$.0896 \text{ g/cc} \quad \text{and} \quad .1052 \text{ g/cc}$$

Considerations in the selection of x values

Formula (9.17) allows a kind of precision to be attached to the slope of the least squares line. It is useful to consider how that precision is related to study characteristics that are potentially under an investigator's control. Notice that both formulas (9.13) and (9.17) indicate that the larger $\sum(x - \bar{x})^2$ is (i.e., the more spread out the x_i values are), the more precision b_1 offers as an estimator of the underlying slope β_1. Thus, as far as the estimation of β_1 is concerned, in studies where x represents the value of a system variable under the control of an experimenter, he or she should choose settings of x with the largest possible sample variance. (In fact, if one has n observations to spend and can choose values of x anywhere in some interval $[a, b]$, taking $\frac{n}{2}$ of them at $x = a$ and $\frac{n}{2}$ at $x = b$ produces the best possible precision for estimating the slope β_1.)

However, this advice (to spread the x_i's out) must be taken with a grain of salt. The approximately linear relationship (9.4) may hold over only a limited range of possible x values. Choosing experimental values of x beyond the limits where it is reasonable to expect formula (9.4) to hold, hoping thereby to obtain a good estimate

of slope, is of course nonsensical. And it is also important to recognize that precise estimation of β_1 under the assumptions of model (9.4) is not the only consideration when planning data collection. It is usually also important to be in a position to tell when the linear form of (9.4) is inappropriate. That dictates that data be collected at a number of different settings of x, not simply at the smallest and largest values possible.

9.1.3 Inference for the Mean System Response for a Particular Value of x

Chapters 7 and 8 repeatedly considered the problem of estimating the mean of y under a particular one (or combination) of the levels of the factor (or factors) of interest. In the present context, the analog is the problem of estimating the mean response for a fixed value of the system variable x,

$$\mu_{y|x} = \beta_0 + \beta_1 x \tag{9.18}$$

The natural data-based approximation of the mean in formula (9.18) is the corresponding y value taken from the least squares line. The notation

Estimator of $\mu_{y|x} = \beta_0 + \beta_1 x$

$$\hat{y} = b_0 + b_1 x \tag{9.19}$$

will be used for this value on the least squares lines. (This is in spite of the fact that the value in formula (9.19) may not be a fitted value in the sense that the phrase has most often been used to this point. x need not be equal to any of $x_1, x_2, \ldots, x_n$ for both expressions (9.18) and (9.19) to make sense.) The simple linear regression model (9.4) leads to simple distributional properties for $\hat{y}$ that then produce inference methods for $\mu_{y|x}$.

Under model (9.4), $\hat{y}$ has a normal distribution with

$$E\hat{y} = \mu_{y|x} = \beta_0 + \beta_1 x$$

and

$$\operatorname{Var}\hat{y} = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}\right) \tag{9.20}$$

(In expression (9.20), notation is being abused somewhat. The i subscripts and indices of summation in $\sum(x-\bar{x})^2$ have been suppressed. This summation runs over the n values x_i included in the original data set. On the other hand, in the $(x-\bar{x})^2$ term appearing as a numerator in expression (9.20), the x involved is not

necessarily equal to any of $x_1, x_2, \ldots, x_n$. Rather, it is simply the value of the system variable at which the mean response is to be estimated.) Then

$$Z = \frac{\hat{y} - \mu_{y|x}}{\sigma\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}}$$

has a standard normal distribution. This in turn motivates the fact that

$$T = \frac{\hat{y} - \mu_{y|x}}{s_{\text{LF}}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}} \tag{9.21}$$

has a t_{n-2} distribution. The standard arguments of Chapter 6 applied to expression (9.21) then show that

$$\text{H}_0\colon \mu_{y|x} = \# \tag{9.22}$$

can be tested using the test statistic

Test statistic for $H_0\colon \mu_{y|x} = \#$

$$T = \frac{\hat{y} - \#}{s_{\text{LF}}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}} \tag{9.23}$$

and a t_{n-2} reference distribution. Further, under the simple linear regression model (9.4), a two-sided individual confidence interval for $\mu_{y|x}$ can be made using endpoints

Confidence limits for the mean repsonse, $\mu_{y|x} = \beta_0 + \beta_1 x$

$$\hat{y} \pm t s_{\text{LF}}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} \tag{9.24}$$

where the associated confidence is the probability assigned to the interval between $-t$ and t by the t_{n-2} distribution. A one-sided interval is made in the usual way based on one endpoint from formula (9.24).

Example 1 (*continued*)

Returning again to the pressure/density study, consider making individual 95% confidence intervals for the mean densities of cylinders produced first at 4,000 psi and then at 5,000 psi.

Treating first the 4,000 psi condition, the corresponding estimate of mean density is

$$\hat{y} = 2.375 + .000048\bar{6}(4{,}000) = 2.5697 \text{ g/cc}$$

Further, from formula (9.24) and the fact that the .975 quantile of the t_{13} distribution is 2.160, a precision of plus-or-minus

$$2.160(.0199)\sqrt{\frac{1}{15} + \frac{(4{,}000 - 6{,}000)^2}{120{,}000{,}000}} = .0136 \text{ g/cc}$$

can be attached to the 2.5697 g/cc figure. That is, endpoints of a two-sided 95% confidence interval for the mean density under the 4,000 psi condition are

$$2.5561 \text{ g/cc} \quad \text{and} \quad 2.5833 \text{ g/cc}$$

Under the $x = 5{,}000$ psi condition, the corresponding estimate of mean density is

$$\hat{y} = 2.375 + .000048\bar{6}(5{,}000) = 2.6183 \text{ g/cc}$$

Using formula (9.24), a precision of plus-or-minus

$$2.160(.0199)\sqrt{\frac{1}{15} + \frac{(5{,}000 - 6{,}000)^2}{120{,}000{,}000}} = .0118 \text{ g/cc}$$

can be attached to the 2.6183 g/cc figure. That is, endpoints of a two-sided 95% confidence interval for the mean density under the 5,000 psi condition are

$$2.6065 \text{ g/cc} \quad \text{and} \quad 2.6301 \text{ g/cc}$$

The reader should compare the plus-or-minus parts of the two confidence intervals found here. The interval for $x = 5{,}000$ psi is shorter and therefore more informative than the interval for $x = 4{,}000$ psi. The origin of this discrepancy should be clear, at least upon scrutiny of formula (9.24). For the students' data, $\bar{x} = 6{,}000$ psi. $x = 5{,}000$ psi is closer to $\bar{x}$ than is $x = 4{,}000$ psi, so the $(x - \bar{x})^2$ term (and thus the interval length) is smaller for $x = 5{,}000$ psi than for $x = 4{,}000$ psi.

The phenomenon noted in the preceding example—that the length of a confidence interval for $\mu_{y|x}$ increases as one moves away from $\bar{x}$—is an important one. And it has an intuitively plausible implication for the planning of experiments where an approximately linear relationship between y and x is expected, and x is under

the investigator's control. If there is an interval of values of x over which one wants good precision in estimating mean responses, it is only sensible to center one's data collection efforts in that interval.

Inference for the intercept, β_0

Proper use of displays (9.22), (9.23), and (9.24) gives inference methods for the parameter β_0 in model (9.4). β_0 is the y intercept of the linear relationship (9.18). So by setting $x = 0$ in displays (9.22), (9.23), and (9.24), tests and confidence intervals for β_0 are obtained. However, unless $x = 0$ is a feasible value for the input variable and the region where the linear relationship (9.18) is a sensible description of physical reality includes $x = 0$, inference for β_0 alone is rarely of practical interest.

The confidence intervals represented by formula (9.24) carry individual associated confidence levels. Section 7.3 showed that it is possible (using the P-R method) to give simultaneous confidence intervals for r possibly different means, μ_i. This comes about essentially by appropriately increasing the t multiplier used in the plus-or-minus part of the formula for individual confidence limits. Here it is possible, by replacing t in formula (9.24) with a larger value, to give simultaneous confidence intervals for *all* means $\mu_{y|x}$. That is, under model (9.4), simultaneous two-sided confidence intervals for all mean responses $\mu_{y|x}$ can be made using respective endpoints

Simultaneous two-sided confidence limits for all means, $\mu_{y|x}$

$$(b_0 + b_1 x) \pm \sqrt{2f}\, s_{\text{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}} \qquad \textbf{(9.25)}$$

where for positive f, the associated simultaneous confidence is the $F_{2,n-2}$ probability assigned to the interval $(0, f)$.

Of course, the practical meaning of the phrase "for all means $\mu_{y|x}$" is more like "for all mean responses in an interval where the simple linear regression model (9.4) is a workable description of the relationship between x and y." As is always the case in curve- and surface-fitting situations, *extrapolation* outside of the range of x values where one has data (and even to some extent *interpolation* inside that range) is risky business. When it is done, it should be supported by subject-matter expertise to the effect that it is justifiable.

It may be somewhat difficult to grasp the meaning of a simultaneous confidence figure applicable to *all* possible intervals of the form (9.25). To this point, the confidence levels considered have been for finite sets of intervals. Probably the best way to understand the theoretically infinite set of intervals given by formula (9.25) is as defining a region in the (x, y)-plane thought likely to contain the line $\mu_{y|x} = \beta_0 + \beta_1 x$. Figure 9.5 is a sketch of a typical confidence region represented by formula (9.25). There is a region indicated about the least squares line whose vertical extent increases with distance from $\bar{x}$ and which has the stated confidence in covering the line describing the relationship between x and $\mu_{y|x}$.

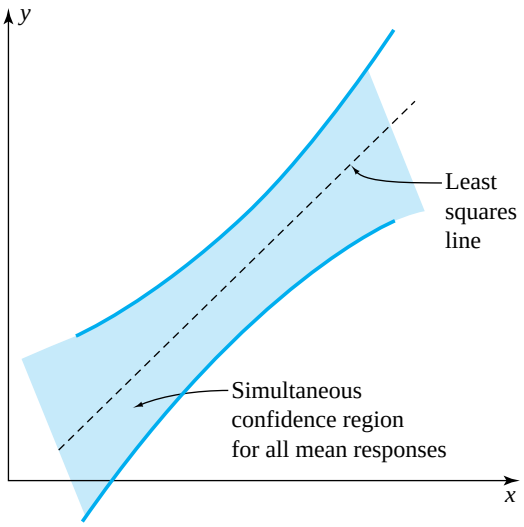

Figure 9.5 Region in the (x, y)-plane defined by simultaneous confidence intervals for all values of $\mu_{y|x}$

Example 1 (*continued*)

It is instructive to compare what the P-R method of Section 7.3 and formula (9.25) give for simultaneous 95% confidence intervals for mean cylinder densities produced under the five conditions actually used by the students in their study.

First, formula (7.28) of Section 7.3 shows that with $n - r = 15 - 5 = 10$ degrees of freedom for s_p and $r = 5$ conditions under study, 95% simultaneous two-sided confidence limits for all five mean densities are of the form

$$\bar{y}_i \pm 3.103 \frac{s_p}{\sqrt{n_i}}$$

which in the present context is

$$\bar{y}_i \pm 3.103 \frac{.0206}{\sqrt{3}}$$

that is,

$$\bar{y}_i \pm .0369 \text{ g/cc}$$

Then, since $\nu_1 = 2$ and $\nu_2 = 13$ degrees of freedom are involved in the use of formula (9.25), simultaneous limits of the form

$$\hat{y} \pm \sqrt{2(3.81)}\, s_{LF} \sqrt{\frac{1}{15} + \frac{(x - 6{,}000)^2}{120{,}000{,}000}}$$

Example 1 (*continued*)

Table 9.5
Simultaneous (and Individual) 95% Confidence Intervals for Mean Cylinder Densities

x, Pressure	$\mu_{y\|x}$ (P-R Method) Mean Density	$\mu_{y\|x}$ (from formula (9.25)) Mean Density	$\mu_{y\|x}$ (from formula (9.24)) Mean Density
2,000 psi	$2.4790 \pm .0369$ g/cc	$2.4723 \pm .0246$ g/cc	$2.4723 \pm .0136$ g/cc
4,000 psi	$2.5693 \pm .0369$ g/cc	$2.5697 \pm .0174$ g/cc	$2.5697 \pm .0118$ g/cc
6,000 psi	$2.6520 \pm .0369$ g/cc	$2.6670 \pm .0142$ g/cc	$2.6670 \pm .0111$ g/cc
8,000 psi	$2.7687 \pm .0369$ g/cc	$2.7643 \pm .0174$ g/cc	$2.7643 \pm .0118$ g/cc
10,000 psi	$2.8660 \pm .0369$ g/cc	$2.8617 \pm .0246$ g/cc	$2.8617 \pm .0136$ g/cc

are indicated. Table 9.5 shows the five intervals that result from the use of each of the two simultaneous confidence methods, together with individual intervals (9.24).

Two points are evident from Table 9.5. First, the intervals that result from formula (9.25) are somewhat wider than the corresponding individual intervals given by formula (9.24). But it is also clear that the use of the simple linear regression model assumptions in preference to the more general one-way assumptions of Chapter 7 can lead to shorter simultaneous confidence intervals and correspondingly sharper real-world engineering inferences.

9.1.4 Prediction and Tolerance Intervals (*Optional*)

Inference for $\mu_{y|x}$ is one kind of answer to the qualitative question, "If I hold the input variable x at some particular level, what can I expect in terms of a system response?" It is an answer in terms of *mean* or long-run average response. Sometimes an answer in terms of *individual responses* is of more practical use. And in such cases it is helpful to know that the simple linear regression model assumptions (9.4) lead to their own specialized formulas for prediction and tolerance intervals.

The basic fact that makes possible prediction intervals under assumptions (9.4) is that if y_{n+1} is one additional observation, coming from the distribution of responses corresponding to a particular x, and $\hat{y}$ is the corresponding fitted value at that x (based on the original n data pairs), then

$$T = \frac{y_{n+1} - \hat{y}}{s_{\text{LF}}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}}$$

has a t_{n-2} distribution. This fact leads in the usual way to the conclusion that under model (9.4) the two-sided interval with endpoints

Simple linear regression prediction limits for an additional y at a given x

$$\hat{y} \pm t s_{\text{LF}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}} \tag{9.26}$$

can be used as a prediction interval for an additional observation y at a particular value of the input variable x. The associated prediction confidence is the probability that the t_{n-2} distribution assigns to the interval between $-t$ and t. One-sided intervals are made in the usual way, by employing only one of the endpoints (9.26) and adjusting the confidence level appropriately.

It is possible not only to derive prediction interval formulas from the simple linear regression model assumptions but also to develop relatively simple formulas for approximate one-sided tolerance bounds. That is, the intervals

A one-sided tolerance interval for the y distribution at x

$$(\hat{y} - \tau s_{\text{LF}}, \infty) \tag{9.27}$$

and

Another one-sided tolerance interval for the y distribution at x

$$(-\infty, \hat{y} + \tau s_{\text{LF}}) \tag{9.28}$$

can be used as one-sided tolerance intervals for a fraction p of the underlying distribution of responses corresponding to a particular value of the system variable x, provided τ is appropriately chosen (depending upon the data, p, x, and the desired confidence level).

In order to write down a reasonably clean formula for τ, the notation

The ratio of $\sqrt{\text{Var}\,\hat{y}}$ to σ for simple linear regression

$$A = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}} \tag{9.29}$$

will be adopted for the multiplier that is used (e.g., in formula (9.24)) to go from an estimate of σ to an estimate of the standard deviation of $\hat{y}$. Then, for approximate

γ level confidence in locating a fraction p of the responses y at the x of interest, τ appropriate for use in interval (9.27) or (9.28) is

Multiplier to use in interval (9.27) or (9.28)

$$\tau = \frac{Q_z(p) + AQ_z(\gamma)\sqrt{1 + \frac{1}{2(n-2)}\left(\frac{Q_z^2(p)}{A^2} - Q_z^2(\gamma)\right)}}{1 - \frac{Q_z^2(\gamma)}{2(n-2)}} \tag{9.30}$$

Example 1 (*continued*)

To illustrate the use of prediction and tolerance interval formulas in the simple linear regression context, consider a 90% lower prediction bound for a single additional density in powder pressing, if a pressure of 4,000 psi is employed. Then, additionally consider finding a 95% lower tolerance bound for 90% of many additional cylinder densities if that pressure is used.

Treating first the prediction problem, formula (9.26) shows that an appropriate prediction bound is

$$2.5697 - 1.350(.0199)\sqrt{1 + \frac{1}{15} + \frac{(4{,}000 - 6{,}000)^2}{120{,}000{,}000}} = 2.5796 - .0282$$

that is,

$$2.5514 \text{ g/cc}$$

If, rather than predicting a single additional density for $x = 4{,}000$ psi, it is of interest to locate 90% of additional densities corresponding to a 4,000 psi pressure, a tolerance bound is in order. First use formula (9.29) and find that

$$A = \sqrt{\frac{1}{15} + \frac{(4{,}000 - 6{,}000)^2}{120{,}000{,}000}} = .3162$$

Next, for 95% confidence, applying formula (9.30),

$$\tau = \frac{1.282 + (.3162)(1.645)\sqrt{1 + \frac{1}{2(15-2)}\left(\frac{(1.282)^2}{(.3162)^2} - (1.645)^2\right)}}{1 - \frac{(1.645)^2}{2(15-2)}} = 2.149$$

So finally, an approximately 95% lower tolerance bound for 90% of densities produced using a pressure of 4,000 psi is (via formula (9.27))

$$2.5697 - 2.149(.0199) = 2.5697 - .0428$$

that is,

$$2.5269 \text{ g/cc}$$

Cautions about prediction and tolerance intervals in regression

The fact that curve-fitting facilitates interpolation and extrapolation makes it imperative that care be taken in the interpretation of prediction and tolerance intervals. All of the warnings regarding the interpretation of prediction and tolerance intervals raised in Section 6.6 apply equally to the present situation. But the new element here (that formally, the intervals can be made for values of x where one has absolutely no data) requires additional caution. If one is to use formulas (9.26), (9.27), and (9.28) at a value of x not represented among $x_1, x_2, \ldots, x_n$, it must be plausible that model (9.4) not only describes system behavior at those x values where one has data, but at the additional value of x as well. And even when this is "plausible" the application of formulas (9.26), (9.27), and (9.28) to new values of x should be treated with a good dose of care. Should one's (unverified) judgment prove wrong, the nominal confidence level has unknown practical relevance.

9.1.5 Simple Linear Regression and ANOVA

Section 7.4 illustrates how, for unstructured studies, partition of the total sum of squares into interpretable pieces provides both (1) intuition and quantification regarding the origin of observed variation and also (2) the basis for an F test of "no differences between mean responses." It turns out that something similar is possible in simple linear regression contexts.

In the unstructured context of Section 7.4, it was useful to name the difference between *SSTot* and *SSE*. The corresponding convention for curve- and surface-fitting situations is stated next in definition form.

Definition 2

In curve- and surface-fitting analyses of multisample studies, the difference

$$SSR = SSTot - SSE$$

will be called the **regression sum of squares**.

It is not obvious, but the difference referred to in Definition 2 in general has the form of a sum of squares of appropriate quantities. In the present context of fitting a line by least squares,

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Without using the particular terminology of Definition 2, this text has already made fairly extensive use of $SSR = SSTot - SSE$. A review of Definition 3 in Chapter 4 (page 130), and Definitions 4 and 6 in Chapter 7 (page 484) will show that in curve- and surface-fitting contexts,

The coefficient of determination for simple linear regression in sum of squares notation

$$R^2 = \frac{SSR}{SSTot} \tag{9.31}$$

That is, SSR is the numerator of the coefficient of determination defined first in Definition 3 (Chapter 4). It is commonly thought of as the part of the raw variability in y that is accounted for in the curve- or surface-fitting process.

SSR and SSE not only provide an appealing partition of $SSTot$ but also form the raw material for an F test of

$$\mathrm{H}_0 : \beta_1 = 0 \tag{9.32}$$

versus

$$\mathrm{H}_a : \beta_1 \neq 0 \tag{9.33}$$

Under model (9.4), hypothesis (9.32) can be tested using the statistic

An F statistic for testing $H_0: \beta_1 = 0$

$$F = \frac{SSR/1}{s^2_{\mathrm{LF}}} = \frac{SSR/1}{SSE/(n-2)} \tag{9.34}$$

and an $F_{1,n-2}$ reference distribution, where large observed values of the test statistic constitute evidence against H_0.

Earlier in this section, the general null hypothesis $\mathrm{H}_0 : \beta_1 = \#$ was tested using the t statistic (9.16). It is thus reasonable to consider the relationship of the F test indicated in displays (9.32), (9.33), and (9.34) to the earlier t test. The null hypothesis $\mathrm{H}_0 : \beta_1 = 0$ is a special form of hypothesis (9.15), $\mathrm{H}_0 : \beta_1 = \#$. It is the most frequently tested version of hypothesis (9.15) because it can (within limits) be interpreted as the null hypothesis that mean response doesn't depend on x. This is because when hypothesis (9.32) is true within the simple linear regression model (9.4), $\mu_{y|x} = \beta_0 + 0 \cdot x = \beta_0$, which doesn't depend on x. (Actually, a better interpretation of a test of hypothesis (9.32) is as a test of whether a linear term in

x adds significantly to one's ability to model the response y after accounting for an overall mean response.)

If one then considers testing hypotheses (9.32) and (9.33), it might appear that the $\# = 0$ version of formula (9.16) and formula (9.34) represent two different testing methods. But they are equivalent. The statistic (9.34) turns out to be the square of the $\# = 0$ version of statistic (9.16), and (two-sided) observed significance levels based on statistic (9.16) and the t_{n-2} distribution turn out to be the same as observed significance levels based on statistic (9.34) and the $F_{1,n-2}$ distribution. So, from one point of view, the F test specified here is redundant, given the earlier discussion. But it is introduced here because of its relationship to the ANOVA ideas of Section 7.4, and because it has an important natural generalization to more complex curve- and surface-fitting contexts. (This generalization is discussed in Section 9.2 and cannot be made equivalent to a t test.)

The partition of *SSTot* into its parts, *SSR* and *SSE*, and the calculation of the statistic (9.34) can be organized in ANOVA table format. Table 9.6 shows the general format that this book will use in the simple linear regression context.

Table 9.6
General Form of the ANOVA Table for Simple Linear Regression

ANOVA Table (for testing $H_0: \beta_1 = 0$)				
Source	*SS*	*df*	*MS*	*F*
Regression	*SSR*	1	$SSR/1$	MSR/MSE
Error	*SSE*	$n-2$	$SSE/(n-2)$	
Total	*SSTot*	$n-1$		

Example 1 (*continued*)

Recall again from the discussion of the pressure/density example in Section 4.1 that

$$SSTot = \sum (y - \bar{y})^2 = .289366$$

Also, from page 654 recall that

$$SSE = \sum (y - \hat{y})^2 = .005153$$

Thus,

$$SSR = SSTot - SSE = .289366 - .005153 = .284213$$

and the specific version of Table 9.6 for the present example is given as Table 9.7.

Example 1 (*continued*)

Then the observed level of significance for testing $H_0: \beta_1 = 0$ is

$$P[\text{an } F_{1,13} \text{ random variable} > 717] < .001$$

and one has very strong evidence against the possibility that $\beta_1 = 0$. A linear term in Pressure is an important contributor to one's ability to describe the behavior of Cylinder Density. This is, of course, completely consistent with the earlier interval-oriented analysis that produced 95% confidence limits for β_1 of

$$.0000448 \text{ (g/cc)/psi} \quad \text{and} \quad .0000526 \text{ (g/cc)/psi}$$

that do not bracket 0.

The value of $R^2 = .9822$ (found first in Section 4.1) can also be easily derived, using the entries of Table 9.7 and the relationship (9.31).

Table 9.7
ANOVA Table for the Pressure/Density Data

ANOVA Table (for testing $H_0: \beta_1 = 0$)

Source	*SS*	*df*	*MS*	*F*
Regression	.284213	1	.284213	717
Error	.005153	13	.000396	
Total	.289366	14		

9.1.6 Simple Linear Regression and Statistical Software

Many of the calculations needed for the methods of this section are made easier by statistical software packages. None of the methods of this section are so computationally intensive that they absolutely require the use of such software, but it is worthwhile to consider its use in the simple linear regression context. Learning where on a typical printout to find the various summary statistics corresponding to calculations made in this section helps in locating important summary statistics for the more complicated curve- and surface-fitting analyses of the next section. Printout 1 is from a MINITAB analysis of the pressure/density data.

Printout 1 Simple Linear Regression for the Pressure/Density Data (*Example 1*)

```
Regression Analysis


The regression equation is
density = 2.38 +0.000049 pressure
```

```
Predictor        Coef       StDev          T        P
Constant      2.37500     0.01206     197.01    0.000
pressure   0.00004867  0.00000182      26.78    0.000

S = 0.01991     R-Sq = 98.2%     R-Sq(adj) = 98.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1     0.28421     0.28421    717.06    0.000
Residual Error    13     0.00515     0.00040
Total             14     0.28937

Obs   pressure    density         Fit   StDev Fit    Residual    St Resid
  1       2000    2.48600     2.47233     0.00890     0.01367        0.77
  2       2000    2.47900     2.47233     0.00890     0.00667        0.37
  3       2000    2.47200     2.47233     0.00890    -0.00033       -0.02
  4       4000    2.55800     2.56967     0.00630    -0.01167       -0.62
  5       4000    2.57000     2.56967     0.00630     0.00033        0.02
  6       4000    2.58000     2.56967     0.00630     0.01033        0.55
  7       6000    2.64600     2.66700     0.00514    -0.02100       -1.09
  8       6000    2.65700     2.66700     0.00514    -0.01000       -0.52
  9       6000    2.65300     2.66700     0.00514    -0.01400       -0.73
 10       8000    2.72400     2.76433     0.00630    -0.04033       -2.14R
 11       8000    2.77400     2.76433     0.00630     0.00967        0.51
 12       8000    2.80800     2.76433     0.00630     0.04367        2.31R
 13      10000    2.86100     2.86167     0.00890    -0.00067       -0.04
 14      10000    2.87900     2.86167     0.00890     0.01733        0.97
 15      10000    2.85800     2.86167     0.00890    -0.00367       -0.21

R denotes an observation with a large standardized residual

Predicted Values

    Fit  StDev Fit         95.0% CI             95.0% PI
2.61833    0.00545   ( 2.60655, 2.63011)  ( 2.57374, 2.66293)
```

Printout 1 is typical of summaries of regression analyses printed by commercially available statistical packages. The most basic piece of information on the printout is, of course, the fitted equation. Immediately below it is a table giving (to more significant digits) the estimated coefficients (b_0 and b_1), their estimated standard deviations, and the t ratios (appropriate for testing whether coefficients β are 0) made up as the quotients. The printout includes the values of s_{LF} and R^2 and an ANOVA table much like Table 9.7. For the several observed values of test statistics printed out (including the observed value of F from formula (9.34)), MINITAB gives observed levels of significance. The ANOVA table is followed by a table of values of y, fitted y,

$$\text{“StDev Fit”} = s_{\text{LF}}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}}$$

and residual, and standardized residual corresponding to the n data points. MINITAB's regression program has an option that allows one to request fitted values, confidence intervals for $\mu_{y|x}$, and prediction intervals for x values of interest, and Printout 1 finishes with this information for the value $x = 5{,}000$.

The reader is encouraged to compare the information on Printout 1 with the various results obtained in Example 1 and verify that everything on the printout (except the "adjusted R^2" value) is indeed familiar.

Section 1 Exercises

1. Return to the situation of Exercise 3 of Section 4.1 and the polymer molecular weight study of R. Harris.
   (a) Find s_{LF} for these data. What does this intend to measure in the context of the engineering problem?
   (b) Plot both residuals versus x and the standardized residuals versus x. How much difference is there in the appearance of these two plots?
   (c) Give a 90% two-sided confidence interval for the increase in mean average molecular weight that accompanies a 1°C increase in temperature here.
   (d) Give individual 90% two-sided confidence intervals for the mean average molecular weight at 212°C and also at 250°C.
   (e) Give simultaneous 90% two-sided confidence intervals for the two means indicated in part (d).
   (f) Give 90% lower prediction bounds for the next average molecular weight, first at 212°C and then at 250°C.
   (g) Give approximately 95% lower tolerance bounds for 90% of average molecular weights, first at 212°C and then at 250°C.
   (h) Make an ANOVA table for testing $H_0: \beta_1 = 0$ in the simple linear regression model. What is the p-value here for a two-sided test of this hypothesis?

2. Return to the situation of Chapter Exercise 1 of Chapter 4 and the concrete strength study of Nicholson and Bartle.
   (a) Find estimates of the parameters β_0, β_1, and σ in the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$. How does your estimate of σ based on the simple linear regression model compare with the pooled sample standard deviation, s_{P}?
   (b) Compute residuals and standardized residuals. Plot both against x and $\hat{y}$ and normal-plot them. How much do the appearances of the plots of the standardized residuals differ from those of the raw residuals?
   (c) Make a 90% two-sided confidence interval for the increase in mean compressive strength that accompanies a .1 increase in the water/cement ratio. (This is $.1\beta_1$).
   (d) Test the hypothesis that the mean compressive strength doesn't depend on the water/cement ratio. What is the p-value?
   (e) Make a 95% two-sided confidence interval for the mean strength of specimens with the water/cement ratio .5 (based on the simple linear regression model).
   (f) Make a 95% two-sided prediction interval for the strength of an additional specimen with the water/cement ratio .5 (based on the simple linear regression model).
   (g) Make an approximately 95% lower tolerance bound for the strengths of 90% of additional specimens with the water/cement ratio .5 (based on the simple linear regression model).

9.2 Inference Methods for General Least Squares Curve- and Surface-Fitting (Multiple Linear Regression)

The previous section presented formal inference methods available under the (normal) simple linear regression model. Confidence interval estimation, hypothesis testing, prediction and tolerance intervals, and ANOVA were all seen to have simple linear regression versions. This section makes a parallel study of more general curve- and surface-fitting contexts. First, the multiple linear regression model and its corresponding variance estimate and standardized residuals are introduced. Then, in turn, there are discussions of how multiple linear regression computer programs can (1) facilitate inference for rate of change parameters in the model, (2) make possible inference for the mean system response at a given combination of values for the input/system variables and the making of prediction and tolerance intervals, and (3) allow the use of ANOVA methods in multiple regression contexts.

9.2.1 The Multiple Linear Regression Model, Corresponding Variance Estimate, and Standardized Residuals

This section considers situations like those treated on a descriptive level in Section 4.2, where for k system variables $x_1, x_2, \ldots, x_k$ and a response y, an approximate relationship like

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{9.35}$$

holds. As in Section 4.2, the form (9.35) not only covers those circumstances where $x_1, x_2, \ldots, x_k$ all represent physically different variables but also describes contexts where some of the variables are functions of others. For example, the relationship

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

can be thought of as a $k = 2$ version of formula (9.35), where x_2 is a deterministic function of x_1, $x_2 = x_1^2$.

As in Section 4.2, a double subscript notation will be used for the values of the input variables. Thus, the problem considered is that of inference based on the data vectors $(x_{11}, x_{21}, \ldots, x_{k1}, y_1), (x_{12}, x_{22}, \ldots, x_{k2}, y_2), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$. As always, a probability model is needed to support formal inferences for such data, and the one considered here is an appropriate specialization of the general one-way normal model of Section 7.1. That is, the standard assumptions of the multiple linear regression model are that there are underlying normal distributions for the response

y with a common variance σ^2 but means $\mu_{y|x_1,x_2,\ldots,x_k}$ that change linearly with each of $x_1, x_2, \ldots, x_k$. In symbols, it is typical to write that for $i = 1, 2, \ldots, n$,

The (normal) multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \tag{9.36}$$

where the ϵ_i are (unobservable) iid normal $(0, \sigma^2)$ random variables, the $x_{1i}, x_{2i}, \ldots, x_{ki}$ are known constants, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and σ^2 are unknown model parameters (fixed constants). This is the specialization of the general one-way model

$$y_{ij} = \mu_i + \epsilon_{ij}$$

to the situation where the means $\mu_{y|x_1,x_2,\ldots,x_k}$ satisfy the relationship

$$\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{9.37}$$

If one thinks of formula (9.37) as defining a surface in $(k+1)$-dimensional space, then the model equation (9.36) simply says that responses y differ from corresponding values on that surface by mean 0, variance σ^2 random noise. Figure 9.6 illustrates this point for the simple $k = 2$ case (where x_1 and x_2 are not functionally related).

Inferences about quantities involving those $(x_1, x_2, \ldots, x_k)$ combinations represented in the data, like the mean response at a single $(x_1, x_2, \ldots, x_k)$ or the difference between two such mean responses, will typically be sharper when methods based on model (9.36) can be used in place of the general methods of Chapter 7. And as was true for simple linear regression, to the extent that it is sensible to assume that model (9.36) describes system behavior for values of $x_1, x_2, \ldots, x_k$ not included

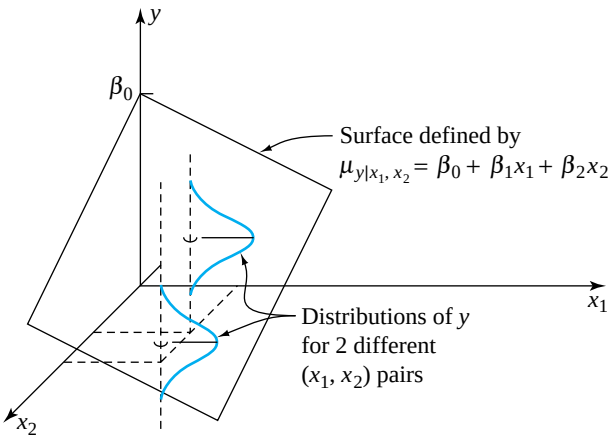

Figure 9.6 Graphical representation of the multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

in the data, it provides the basis for inferences involving limited interpolation and extrapolation on the system variables $x_1, x_2, \ldots, x_k$.

Section 4.2 contains a discussion of using statistical software in the least squares fitting of the approximate relationship (9.35) to a set of $(x_1, x_2, \ldots, x_k, y)$ data. That discussion can be thought of as covering the fitting and use of residuals in model checking for the multiple linear regression model (9.36). Section 4.2 did not produce explicit formulas for $b_0, b_1, b_2, \ldots, b_k$, the (least squares) estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Instead it relied on the software to produce those estimates. Of course, once one has estimates of the β's, corresponding fitted values immediately become

Estimators of the coefficients β in the multiple linear regression model

Fitted values for the multiple linear regression model

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} \tag{9.38}$$

with residuals

Residuals for the multiple linear regression model

$$e_i = y_i - \hat{y}_i \tag{9.39}$$

The residuals (9.39) can be used to make up an estimate of σ^2. One divides a sum of squared residuals by an appropriate number of degrees of freedom. That is, one can make the following definition of a **(multiple linear regression or) surface-fitting sample variance**.

Definition 3

For a set of n data vectors $(x_{11}, x_{21}, \ldots, x_{k1}, y_1), (x_{12}, x_{22}, \ldots, x_{k2}, y_2), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$ where least squares fitting produces fitted values given by formula (9.38) and residuals (9.39),

$$s_{\text{SF}}^2 = \frac{1}{n-k-1}\sum(y-\hat{y})^2 = \frac{1}{n-k-1}\sum e^2 \tag{9.40}$$

will be called a **surface-fitting sample variance**. Associated with it are $\nu = n - k - 1$ degrees of freedom and an estimated standard deviation of response, $s_{\text{SF}} = \sqrt{s_{\text{SF}}^2}$.

Compare Definitions 1 and 3 and notice that the $k = 1$ version of s_{SF}^2 is just s_{LF}^2 from simple linear regression. s_{SF} estimates the level of basic background variation, σ, whenever the model (9.36) is an adequate description of the system under study. When it is not, s_{SF} will tend to overestimate σ. So comparing s_{SF} to s_{P} is another way of investigating the appropriateness of that description. (s_{SF} much larger than s_{P} suggests that model (9.36) is a poor one.)

Example 2
(*Example 5, Chapter 4, revisited—page 150*)

Inference in the Nitrogen Plant Study

The main example in this section will be the nitrogen plant data set given in Table 4.8. Recall that in the discussion of the example, with

$$x_1 = \text{a measure of air flow}$$
$$x_2 = \text{the cooling water inlet temperature}$$
$$y = \text{a measure of stack loss}$$

the fitted equation

$$\hat{y} = -15.409 - .069x_1 + .528x_2 + .007x_1^2$$

appeared to be a sensible data summary. Accordingly, consider the making of inferences based on the $k = 3$ version of model (9.36),

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \epsilon_i \qquad \textbf{(9.41)}$$

Printout 2 is from a MINITAB analysis of the data of Table 4.8. Among many other things, it gives the values of the residuals from the fitted version of formula (9.41) for all $n = 17$ data points. It is then possible to apply Definition 3 and produce a surface-fitting estimate of the parameter σ^2 in the model (9.41). That is,

$$s_{\text{SF}}^2 = \frac{1}{17 - 3 - 1}\left((.053)^2 + (-.125)^2 + \cdots + (.265)^2 + (2.343)^2\right)$$
$$= 1.26$$

so a corresponding estimate of σ is

$$s_{\text{SF}} = \sqrt{1.26}$$
$$= 1.125$$

(The units of y—and therefore s_{SF}—are .1% of incoming ammonia escaping unabsorbed.)

In routine practice it is a waste to do even these calculations, since multiple regression programs typically output s_{SF} as part of their analysis. The reader should take time to locate the value $s_{\text{SF}} = 1.125$ on Printout 2. If one accepts the relevance of model (9.41), for fixed values of airflow and inlet temperature (and therefore airflow squared), the standard deviation associated with many days' stack losses produced under those conditions would then be expected to be approximately .1125%.

Printout 2 Multiple Linear Regression for the Stack Loss Data (*Example 2*)

```
Regression Analysis

The regression equation is
y = - 15.4 - 0.069 x1 + 0.528 x2 + 0.00682 x1**2

Predictor        Coef       StDev          T        P
Constant      -15.41       12.60      -1.22    0.243
x1           -0.0691      0.3984      -0.17    0.865
x2            0.5278      0.1501       3.52    0.004
x1**2       0.006818    0.003178       2.15    0.051

S = 1.125       R-Sq = 98.0%     R-Sq(adj) = 97.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      799.80      266.60    210.81    0.000
Residual Error    13       16.44        1.26
Total             16      816.24

Source       DF      Seq SS
x1            1      775.48
x2            1       18.49
x1**2         1        5.82

Obs         x1          y         Fit   StDev Fit    Residual    St Resid
  1       80.0     37.000      36.947       1.121       0.053       0.57 X
  2       62.0     18.000      18.125       0.407      -0.125      -0.12
  3       62.0     18.000      18.653       0.462      -0.653      -0.64
  4       62.0     19.000      19.181       0.553      -0.181      -0.18
  5       62.0     20.000      19.181       0.553       0.819       0.84
  6       58.0     15.000      15.657       0.513      -0.657      -0.66
  7       58.0     14.000      13.018       0.475       0.982       0.96
  8       58.0     14.000      13.018       0.475       0.982       0.96
  9       58.0     13.000      12.490       0.595       0.510       0.53
 10       58.0     11.000      13.018       0.475      -2.018      -1.98
 11       58.0     12.000      13.546       0.378      -1.546      -1.46
 12       50.0      8.000       7.680       0.493       0.320       0.32
 13       50.0      7.000       7.680       0.493      -0.680      -0.67
 14       50.0      8.000       8.208       0.499      -0.208      -0.21
 15       50.0      8.000       8.208       0.499      -0.208      -0.21
 16       50.0      9.000       8.735       0.548       0.265       0.27
 17       56.0     15.000      12.657       0.298       2.343       2.16R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Predicted Values

       Fit  StDev Fit         95.0% CI             95.0% PI
    15.544      0.383   (  14.717,  16.372)  (  12.978,  18.111)
```

Example 2 (*continued*)

Among the 17 data points in Table 4.8, there are only 12 different airflow/inlet temperature combinations (and therefore 12 different (x_1, x_2, x_1^2) vectors). The original data can be thought of as organized into $r = 12$ separate samples, one for each different (x_1, x_2, x_1^2) vector and there is thus an estimate of σ that doesn't depend for its validity on the appropriateness of the assumption that $\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$. That is, s_{P} can be computed and compared it to s_{SF} as a check on the appropriateness of model (9.41). Table 9.8 organizes the calculation of that pooled estimate of σ.

Table 9.8
Twelve Sample Means and Four Sample Variances for the Stack Loss Data

x_1, Air Flow	x_2, Inlet Temperature	y, Stack Loss	$\bar{y}$	s^2
50	18	8, 7	7.5	.5
50	19	8, 8	8.0	0.0
50	20	9	9.0	—
56	20	15	15.0	—
58	17	13	13.0	—
58	18	14, 14, 11	13.0	3.0
58	19	12	12.0	—
58	23	15	15.0	—
62	22	18	18.0	—
62	23	18	18.0	—
62	24	19, 20	19.5	.5
80	27	37	37.0	—

Then

$$s_{\text{P}}^2 = \frac{1}{17-12}\left((2-1)(.5) + (2-1)(0.0) + (3-1)(3.0) + (2-1)(.5)\right)$$
$$= 1.40$$

so

$$s_{\text{P}} = \sqrt{s_{\text{P}}^2} = \sqrt{1.40} = 1.183$$

The fact that $s_{SF} = 1.125$ and $s_P = 1.183$ are in substantial agreement is consistent with the work in Example 5 of Chapter 4, which found the equation

$$\hat{y} = -15.409 - .069x_1 + .528x_2 + .007x_1^2$$

to be a good summarization of the nitrogen plant data.

s_{SF} is basic to all of formal statistical inference based on the multiple linear regression model. But before using it to make statistical intervals and do significance testing, note also that it is useful for producing standardized residuals for the multiple linear regression model. That is, it is possible to find positive constants $a_1, a_2, \ldots, a_n$ (which are each complicated functions of all of $x_{11}, x_{21}, \ldots, x_{k1}, x_{12}, x_{22}, \ldots, x_{k2}, \ldots, x_{1n}, x_{2n}, \ldots, x_{kn}$) such that the ith residual $e_i = y_i - \hat{y}_i$ has

$$\text{Var}(y_i - \hat{y}_i) = a_i\sigma^2$$

Then, recalling Definition 2 in Chapter 7 (page 458), corresponding to the data point $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ is the standardized residual for multiple linear regression

Standardized residuals for multiple linear regression

$$e_i^* = \frac{e_i}{s_{SF}\sqrt{a_i}} \tag{9.42}$$

It is not possible to include here a simple formula for the a_i that are needed to compute standardized residuals. (They are of interest only as building blocks in formula (9.42) anyway.) But it is easy to read the standardized residuals (9.42) off a typical multiple regression printout and to plot them in the usual ways as means of checking the apparent appropriateness of a candidate version of model (9.36) fit to a set of n data points $(x_1, x_2, \ldots, x_k, y)$.

Example 2 (*continued*)

As an illustration of the use of standardized residuals, consider again Printout 2 on page 679. The annotations on that printout locate the columns of residuals and standardized residuals for model (9.41). Figure 9.7 depicts normal probability plots, first of the raw residuals and then of the standardized residuals.

There are only the most minor differences between the appearances of the two plots in Figure 9.7, suggesting that decisions concerning the appropriateness of model (9.41) based on raw residuals will not be much altered by the more sophisticated consideration of standardized residuals instead.

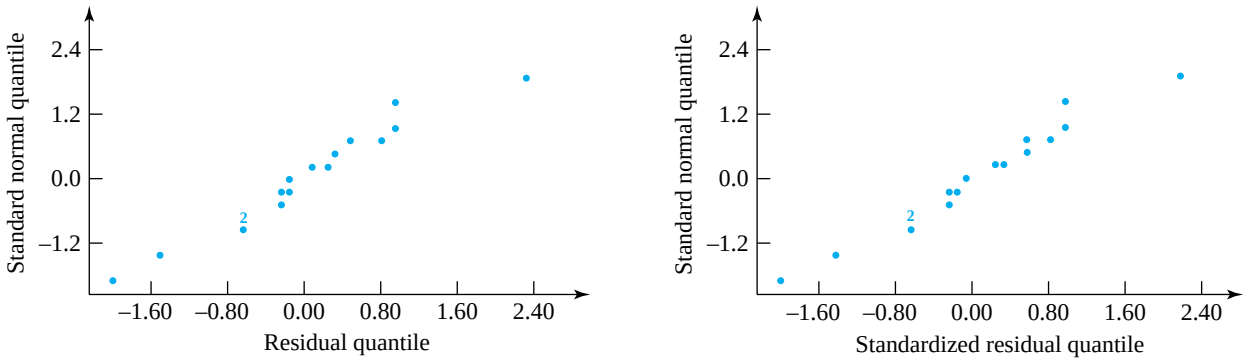

Figure 9.7 Normal plots of residuals and standardized residuals for the stack loss data (*Example 2*)

9.2.2 Inference for the Parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$

Section 9.1 considered inference for the slope parameter β_1 in simple linear regression, treating it as a rate of change (of average y as a function of x). In the multiple regression context, if $x_1, x_2, \ldots, x_k$ are all physically different system variables, the coefficients $\beta_1, \beta_2, \ldots, \beta_k$ can again be thought of as rates of change of average response with respect to $x_1, x_2, \ldots, x_k$, respectively. (They are partial derivatives of $\mu_{y|x_1,x_2,\ldots,x_k}$ with respect to the x's.) On the other hand, when some x's are functionally related to others (for instance, if $k = 2$ and $\mu_{y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$), individual interpretation of the β's can be less straightforward. In any case, the β's do determine the nature of the surface represented by

$$\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

and it is possible to do formal inference for $\beta_0, \beta_1, \ldots, \beta_k$ individually. In many instances, important physical interpretations can be found for such inferences. (For example, beginning with $\mu_{y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$, an inference that β_2 is positive says that the mean response is concave up as a function of x and has a minimum value.)

The key to formal inference for the β's is that under model (9.36), there are positive constants $d_0, d_1, d_2, \ldots, d_k$ (which are each complicated functions of all of $x_{11}, \ldots, x_{k1}, x_{12}, \ldots, x_{k2} \ldots, x_{1n}, \ldots, x_{kn}$) such that the least squares coefficients $b_0, b_1, \ldots, b_k$ are normally distributed with

$$E b_l = \beta_l$$

and

$$\operatorname{Var} b_l = d_l \sigma^2$$

This in turn makes it plausible that for $l = 0, 1, 2, \ldots, k$, the quantity

Estimated standard deviation of b_l

$$s_{\text{SF}}\sqrt{d_l} \tag{9.43}$$

is an estimate of the standard deviation of b_l and that

$$T = \frac{b_l - \beta_l}{s_{\text{SF}}\sqrt{d_l}} \tag{9.44}$$

has a t_{n-k-1} distribution.

There is no simple way to write down formulas for the constants d_l, but the estimated standard deviations of the coefficients, $s_{\text{SF}}\sqrt{d_l}$, are a typical part of the output from multiple linear regression programs.

The usual arguments of Chapter 6 applied to expression (9.44) then show that

$$\text{H}_0\colon \beta_l = \# \tag{9.45}$$

can be tested using the test statistic

Test statistic for $H_0\colon \beta_l = \#$

$$\boxed{T = \frac{b_l - \#}{s_{\text{SF}}\sqrt{d_l}}} \tag{9.46}$$

and a t_{n-k-1} reference distribution. More importantly, under the multiple linear regression model (9.36), a two-sided individual confidence interval for β_l can be made using endpoints

Confidence limits for β_l

$$\boxed{b_l \pm t s_{\text{SF}}\sqrt{d_l}} \tag{9.47}$$

where the associated confidence is the probability assigned to the interval between $-t$ and t by the t_{n-k-1} distribution. Appropriate use of only one of the endpoints (9.47) gives a one-sided interval for β_l.

Example 2 (*continued*)

Looking again at Printout 2 (see page 679), note that MINITAB's multiple regression output includes a table of estimated coefficients (b_l) and (estimated) standard deviations ($s_{\text{SF}}\sqrt{d_l}$). These are collected in Table 9.9.

Example 2 (*continued*)

Table 9.9
Fitted Coefficients and Estimates of Their Standard Deviations for the Stack Loss Data

Estimated Coefficient	(Estimated) Standard Deviation of the Estimate
$b_0 = -15.41$	$s_{SF}\sqrt{d_0} = 12.60$
$b_1 = -.0691$	$s_{SF}\sqrt{d_1} = .3984$
$b_2 = .5278$	$s_{SF}\sqrt{d_2} = .1501$
$b_3 = .006818$	$s_{SF}\sqrt{d_3} = .003178$

Then since the upper .05 point of the t_{13} distribution is 1.771, from formula (9.47) a two-sided 90% confidence interval for β_2 in model (9.41) has endpoints

$$.5278 \pm 1.771(.1501)$$

that is,

.2620 (.1% nitrogen loss/degree) and .7936 (.1% nitrogen loss/degree)

This interval establishes that there is an increase in mean stack loss y with increased inlet temperature x_2 (the interval contains only positive values). It further gives a way of assessing the likely impact on y of various changes in x_2. For example, if x_1 (and therefore $x_3 = x_1^2$) is held constant but x_2 is increased by 2°, one can anticipate an increase in mean stack loss of between

.5240 (.1% nitrogen loss) and 1.5873 (.1% nitrogen loss)

As a second example of the use of formula (9.47), note that a 90% two-sided confidence interval for β_3 has endpoints

$$.006818 \pm 1.771(.003178)$$

that is,

.0012 and .0124

β_3 controls the amount and direction of curvature (in the variable x_1) possessed by the surface specified by $\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$. Since the interval contains only positive values, it shows that at the 90% confidence level, there is some important concave-up curvature in the airflow variable needed to describe the stack loss variable. This is consistent with the picture of fitted mean response given previously in Figure 4.15 (see page 155).

However, check that if 95% confidence is used in the calculation of the two-sided interval for β_3, the resulting confidence interval contains values on both sides of 0. If this higher level of confidence is needed, the data in hand are not adequate to establish definitively the nature of any curvature in mean stack loss as a function of airflow. Any real curvature appears weak enough in comparison to the basic background variation that more data are needed to decide whether the surface is concave up, linear, or concave down in the variable x_1.

Very often multiple regression programs output not only the estimated standard deviations of fitted coefficients (9.43) but also the ratios

$$t = \frac{b_l}{s_{\text{SF}}\sqrt{d_l}}$$

and associated two-sided p-values for testing

$$\text{H}_0 : \beta_l = 0$$

Review Printout 2 and note that, for example, the two-sided p-value for testing $\text{H}_0 : \beta_3 = 0$ in model (9.41) is slightly larger than .05. This is completely consistent with the preceding discussion regarding the interpretation of interval estimates of β_3.

9.2.3 Inference for the Mean System Response for a Particular Set of Values for $x_1, x_2, \ldots, x_k$

Inference methods for the parameters $\beta_0, \beta_1, \ldots, \beta_k$ provide insight into the nature of the relationships between $x_1, x_2, \ldots, x_k$ and the mean response y. But other methods are needed to answer the important engineering question, *"What can be expected in terms of system response if I use a particular combination of levels of the system variables $x_1, x_2, \ldots, x_k$?"* An answer to this question will first be phrased in terms of inference methods for the mean system response $\mu_{y|x_1,x_2,\ldots,x_k}$.

In a manner similar to what was done in Section 9.1, the notation

Estimator of $\mu_{y|x_1,x_2,\ldots,x_k}$

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \qquad \textbf{(9.48)}$$

will here be used for the value produced by the least squares equation when a particular set of numbers $x_1, x_2, \ldots, x_k$ is plugged into it. ($\hat{y}$ may not be a fitted value in the strict sense of the phrase, as the vector $(x_1, x_2, \ldots, x_k)$ may not match any data vector $(x_{1i}, x_{2i}, \ldots, x_{ki})$ used to produce the least squares coefficients $b_0, b_1, \ldots, b_k$.) As it turns out, the multiple linear regression model (9.36) leads to simple distributional properties for $\hat{y}$, which then produce inference methods for $\mu_{y|x_1,x_2,\ldots,x_k}$.

Under model (9.36), it is possible to find a positive constant A depending in a complicated way upon $x_1, x_2, \ldots, x_k$ and all of $x_{11}, \ldots, x_{k1}, x_{12}, \ldots, x_{k2}, \ldots, x_{1n}, \ldots, x_{kn}$ (the locations at which inference is desired and at which the original data points were collected) so that $\hat{y}$ has a normal distribution with

$$E\hat{y} = \mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

and

$A = \sqrt{\text{Var}\,\hat{y}}/\sigma$

$$\text{Var}\,\hat{y} = \sigma^2 A^2 \tag{9.49}$$

In view of formula (9.49), it is thus plausible that

Estimated standard deviation of $\hat{y}$

$$s_{\text{SF}} \cdot A \tag{9.50}$$

can be used as an estimated standard deviation for $\hat{y}$ and that inference methods for the mean system response can be based on the fact that

$$T = \frac{\hat{y} - \mu_{y|x_1,x_2,\ldots,x_k}}{s_{\text{SF}} \cdot A}$$

has a t_{n-k-1} distribution. That is,

$$\text{H}_0 : \mu_{y|x_1,x_2,\ldots,x_k} = \# \tag{9.51}$$

can be tested using the test statistic

Test statistic for $H_0 : \mu_{y|x_1,x_2,\ldots,x_k} = \#$

$$\boxed{T = \frac{\hat{y} - \#}{s_{\text{SF}} \cdot A}} \tag{9.52}$$

and a t_{n-k-1} reference distribution. Further, under the multiple linear regression model (9.36), a two-sided confidence interval for $\mu_{y|x_1,x_2,\ldots,x_k}$ can be made using endpoints

Confidence limits for the mean response $\mu_{y|x_1,x_2,\ldots,x_k}$

$$\boxed{\hat{y} \pm t s_{\text{SF}} \cdot A} \tag{9.53}$$

where the associated confidence is the probability assigned to the interval between $-t$ and t by the t_{n-k-1} distribution. One-sided intervals based on formula (9.53) are made in the usual way.

Finding the factor A

The practical obstacle to be overcome in the use of these methods is the computation of A. Although it is not possible to give a simple formula for A, most multiple regression programs provide A for $(x_1, x_2, \ldots, x_k)$ vectors of interest. MINITAB, for example, will fairly automatically produce values of $s_{\text{SF}} \cdot A$ corresponding to

each data point $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, labeled as (the estimated) **standard deviation (of the) fit**. And an option makes it possible to obtain similar information for *any* user-specified choice of $(x_1, x_2, \ldots, x_k)$. (Division of this by s_{SF} then produces A.)

Example 2 (*continued*)

Consider the problem of estimating the mean stack loss if the nitrogen plant of Example 5 in Chapter 4 is operated consistently with $x_1 = 58$ and $x_2 = 19$. (Notice that this means that $x_3 = x_1^2 = 3{,}364$ is involved.) Now the conditions $x_1 = 58$, $x_2 = 19$, and $x_3 = 3{,}364$ match perfectly those of data point number 11 on Printout 2 (see page 679). Thus, $\hat{y}$ and $s_{SF} \cdot A$ for these conditions may be read directly from the printout as 13.546 and .378, respectively. Then, for example, from formula (9.53), a 90% two-sided confidence interval for the mean stack loss corresponding to an airflow of 58 and water inlet temperature of 19 has endpoints

$$13.546 \pm 1.771(.378)$$

that is,

$$12.88 \text{ (.1\% nitrogen loss)} \quad \text{and} \quad 14.22 \text{ (.1\% nitrogen loss)}$$

As a second illustration of the use of formula (9.53), suppose that setting plant operating conditions at an airflow of $x_1 = 60$ and a water inlet temperature of $x_2 = 20$ is contemplated and it is desireable to have an interval estimate for the mean stack loss implied by those conditions. Notice that the $x_1 = 60$, $x_2 = 20$, and $x_3 = x_1^2 = 3{,}600$ vector does not exactly match that of any of the $n = 17$ data points available. Therefore, some interpolation/extrapolation is required to make the desired interval. And it will not be possible to simply read appropriate values of $\hat{y}$ and $s_{SF} \cdot A$ off Printout 2 as related to one of the data points used to fit the equation.

Location of the point with coordinates $x_1 = 60$ and $x_2 = 20$ on a scatterplot of (x_1, x_2) values for the original $n = 17$ data points (like Figure 4.19) reveals that the candidate operating conditions are not wildly different from those used to develop the fitted equation. So there is hope that the use of formula (9.53) will provide an inference of some practical relevance. Accordingly, the coordinates $x_1 = 60$, $x_2 = 20$, and $x_3 = x_1^2 = 3{,}600$ were input into MINITAB and a "prediction" request made, resulting in the final section of Printout 2. Reading from that final section of the printout, $\hat{y} = 15.544$ and $s_{SF} \cdot A = .383$, so a 90% two-sided confidence interval for the mean stack loss has endpoints

$$15.544 \pm 1.771(.383)$$

that is,

$$14.86 \text{ (.1\% nitrogen loss)} \quad \text{and} \quad 16.22 \text{ (.1\% nitrogen loss)}$$

(Of course, endpoints of a 95% interval can be read directly from the printout.)

Example 2 (*continued*)

It is impossible to overemphasize the fact that the preceding two intervals are dependent for their practical relevance on that of model (9.41) for not only those (x_1, x_2) pairs in the original data but (in the second case) also for the $x_1 = 60$ and $x_2 = 20$ set of conditions. Formulas like (9.53) always allow for imprecision due to statistical fluctuations/background noise in the data. They *do not*, however, allow for discrepancies related to the application of a model in a regime over which it is not appropriate. Formula (9.53) is an important and useful formula. But it should be used thoughtfully, with no expectation that it will magically do more than help quantify the precision provided by the data in the context of a particular set of model assumptions.

Lacking a simple explicit formula for A, it is difficult to be very concrete about how this quantity varies. In qualitative terms, it does change with the $(x_1, x_2, \ldots, x_k)$ vector under consideration. It is smallest when this vector is near the center of the cloud of points $(x_{1i}, x_{2i}, \ldots, x_{ki})$ in k-dimensional space corresponding to the n data points used to fit model (9.36). The fact that it can vary substantially is obvious from Printout 2. There for the nitrogen plant case, the estimated standard deviation of $\hat{y}$ given in display (9.50) varies from .298 to 1.121, indicating that A for data point 1 is about 3.8 times the size of A for data point 17 ($\frac{1.121}{.298} \approx 3.8$). That is, the precision with which a mean response is determined can vary widely over the region where it is sensible to use a fitted equation.

Formula (9.53) provides individual confidence intervals for mean responses. Simultaneous intervals are also easily obtained by a modification of formula (9.53) similar to the one provided for simple linear regression. That is, under the multiple linear regression model, simultaneous two-sided confidence intervals for all mean responses $\mu_{y|x_1,x_2,\ldots,x_k}$ can be made using respective endpoints

Simultaneous two-sided confidence limits for all mean repsonses $\mu_{y|x_1,x_2,\ldots,x_k}$

$$\hat{y} \pm \sqrt{(k+1)\,f}\, s_{\mathrm{SF}} \cdot A \qquad \textbf{(9.54)}$$

where for positive f, the associated confidence is the $F_{k+1,n-k-1}$ probability assigned to the interval $(0, f)$. Formula (9.54) is related to formula (9.53) through the replacement of the multiplier t by the (larger for a given nominal confidence) multiplier $\sqrt{(k+1)\,f}$. When it is applied only to $(x_1, x_2, \ldots, x_k)$ vectors found in the original n data points, formula (9.54) is an alternative to the P-R method of simultaneous intervals for means, appropriate to surface-fitting problems. When the multiple linear regression model is indeed appropriate, formula (9.54) will usually give shorter simultaneous intervals than the P-R method.

Example 2 (*continued*)

For making simultaneous 90% confidence intervals for the mean stack losses at the 12 different sets of plant conditions represented in the original data set, one can use formula (9.54) with $k = 3$, $f = 2.43$ (the .9 quantile of the $F_{4,13}$ distribution) and the $\hat{y}$ and corresponding $s_{\mathrm{SF}} \cdot A$ values appearing on Printout 2 (see page 679). For example, considering the $x_1 = 80$ and $x_2 = 27$ conditions of

observation 1 on the printout, $s_{SF} \cdot A = 1.121$ and one of the simultaneous 90% confidence intervals associated with these conditions has endpoints

$$36.947 \pm \sqrt{(3+1)(2.43)(1.121)}$$

or

33.452 (.1% nitrogen loss) and 40.442 (.1% nitrogen loss)

9.2.4 Prediction and Tolerance Intervals (*Optional*)

The second kind of answer that statistical theory can provide to the question, "What is to be expected in terms of system response if one uses a particular $(x_1, x_2, \ldots, x_k)$?", has to do with individual responses rather than mean responses. That is, the same factor A referred to in making confidence intervals for mean responses can be used to develop prediction and tolerance intervals for surface-fitting situations.

In the first place, under model (9.36), the two-sided interval with endpoints

Multiple regression prediction limits for an additional y at $(x_1, x_2, \ldots, x_k)$

$$\hat{y} \pm t s_{SF} \sqrt{1 + A^2} \tag{9.55}$$

can be used as a prediction interval for an additional observation at a particular combination of levels of the variables $x_1, x_2, \ldots, x_k$. The associated prediction confidence is the probability that the t_{n-k-1} distribution assigns to the interval between $-t$ and t. One-sided intervals are made in the usual way, by employing only one of the endpoints (9.55) and adjusting the confidence level appropriately.

In order to use formula (9.55), $s_{SF} \cdot A$ and s_{SF} can be taken from a multiple regression printout and A obtained via division. Equivalently, it is possible to use a small amount of algebra to rewrite formula (9.55) as

An alternative formula for prediction limits

$$\hat{y} \pm t \sqrt{s_{SF}^2 + (s_{SF} \cdot A)^2} \tag{9.56}$$

and substitute s_{SF} and $s_{SF} \cdot A$ directly into formula (9.56).

In order to find one-sided tolerance bounds in the surface-fitting context, begin with the value of A corresponding to a particular $(x_1, x_2, \ldots, x_k)$. If a confidence level of γ is desired in locating a fraction p of the underlying distribution of responses, compute

Multiplier to use in making tolerance intervals in multiple regression

$$\tau = \frac{Q_z(p) + A Q_z(\gamma) \sqrt{1 + \dfrac{1}{2(n-k-1)} \left(\dfrac{Q_z^2(p)}{A^2} - Q_z^2(\gamma) \right)}}{1 - \dfrac{Q_z^2(\gamma)}{2(n-k-1)}} \tag{9.57}$$

Then, the interval

A one-sided tolerance interval for the y distribution at $(x_1, x_2, \ldots, x_k)$

$$(\hat{y} - \tau s_{SF}, \infty) \tag{9.58}$$

or

Another one-sided tolerance interval for the y distribution at $(x_1, x_2, \ldots, x_k)$

$$(-\infty, \hat{y} + \tau s_{SF}) \tag{9.59}$$

can be used as an approximately γ level one-sided tolerance interval for a fraction p of the underlying distribution of responses corresponding to $(x_1, x_2, \ldots, x_k)$.

Example 2 (*continued*)

Returning to the nitrogen plant example, consider first the calculation of a 90% lower prediction bound for a single additional stack loss y, if airflow of $x_1 = 58$ and water inlet temperature of $x_2 = 19$ are used. Then consider also a 95% lower tolerance bound for 90% of many additional stack loss values if the plant is run under those conditions.

Treating the prediction interval problem, recall that for $x_1 = 58$ and $x_2 = 19$, $\hat{y} = 13.546$ and $s_{SF} \cdot A = .378$. Since $s_{SF} = 1.125$ and the .9 quantile of the t_{13} distribution is 1.350, formula (9.56) shows that the desired 90% lower prediction bound for an additional stack loss under such plant operating conditions is

$$13.546 - 1.350\sqrt{(1.125)^2 + (.378)^2}$$

that is, approximately

$$11.94 \text{ (.1\% nitrogen loss)}$$

To not predict a single additional stack loss, but rather to locate 90% of many additional stack losses with 95% confidence, expression (9.57) is the place to begin. Note that for $x_1 = 58$ and $x_2 = 19$,

$$A = .378/1.125 = .336$$

so, using expression (9.57),

$$\tau = \frac{1.282 + (.378)(1.645)\sqrt{1 + \frac{1}{2(17-3-1)}\left(\frac{(1.282)^2}{(.378)^2} - (1.645)^2\right)}}{1 - \frac{(1.645)^2}{2(17-3-1)}} = 2.234$$

So finally, a 95% lower tolerance bound for 90% of stack losses produced under operating conditions of $x_1 = 58$ and $x_2 = 19$ is, via display (9.58),

$$13.546 - 2.234(1.125) = 13.546 - 2.513$$

that is,

$$11.033 \text{ (.1\% nitrogen loss)}$$

The warnings raised in the previous section concerning prediction and tolerance intervals in simple regression all apply equally to the present case of multiple regression. So do points similar to those made in Example 2 (page 688) in reference to confidence intervals for the mean system response. Although they are extremely useful engineering tools, statistical intervals are never any better than the models on which they are based.

9.2.5 Multiple Regression and ANOVA

Formal inference in curve- and surface-fitting contexts can (and typically should) be carried out primarily using interval-oriented methods. Nevertheless, testing and ANOVA methods do have their place. So the discussion now turns to the matter of what ANOVA ideas provide in multiple regression.

As always, *SSTot* will stand for $\sum(y - \bar{y})^2$ and *SSE* for $\sum(y - \hat{y})^2$. Remember also that Definition 2 introduced the notation *SSR* for the difference *SSTot* − *SSE*. As remarked following Definition 2, the coefficient of determination can be written in terms of *SSR* and *SSTot* as

$$R^2 = \frac{SSTot - SSE}{SSTot} = \frac{SSR}{SSTot}$$

Further, under model (9.36), these sums of squares (*SSTot*, *SSE*, and *SSR*) form the basis of an F test of the hypothesis

$$\text{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{9.60}$$

versus

$$\text{H}_\text{a}\text{: not H}_0 \tag{9.61}$$

Hypothesis (9.60) can be tested using the statistic

F statistic for testing $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$$F = \frac{SSR/k}{SSE/(n-k-1)} \tag{9.62}$$

and an $F_{k,n-k-1}$ reference distribution, where large observed values of the test statistic constitute evidence against H_0. (The denominator of statistic (9.62) is another way of writing s^2_{SF}.)

Hypothesis (9.60) in the context of the multiple linear regression model implies that the mean response doesn't depend on any of the process variables $x_1, x_2, \ldots, x_k$. That is, if all of β_1 through β_k are 0, model statement (9.36) reduces to

$$y_i = \beta_0 + \epsilon_i$$

Interpreting a test of $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

So a test of hypothesis (9.60) is often interpreted as a test of whether the mean response is related to any of the input variables under consideration. The calculations leading to statistic (9.62) are most often organized in a table quite similar to the one discussed in Section 9.1 for testing $H_0: \beta_1 = 0$ in simple linear regression. The general form of that table is given as Table 9.10.

Table 9.10
General Form of the ANOVA Table for Testing $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ in Multiple Regression

ANOVA Table (for testing $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$)				
Source	*SS*	*df*	*MS*	*F*
Regression	*SSR*	k	SSR/k	MSR/MSE
Error	*SSE*	$n-k-1$	$SSE/(n-k-1)$	
Total	*SSTot*	$n-1$		

Example 2 (*continued*)

Once again turning to the analysis of the nitrogen plant data under the model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \epsilon_i$, consider testing $H_0: \beta_1 = \beta_2 = \beta_3 = 0$—that is, mean stack loss doesn't depend on airflow (or its square) or water inlet temperature. Printout 2 (see page 679) includes an ANOVA table for testing this hypothesis, which is essentially reproduced here as Table 9.11.

From Table 9.11, the observed value of the F statistic is 210.81, which is to be compared to $F_{3,13}$ quantiles in order to produce an observed level of significance. As indicated in Printout 2, the $F_{3,13}$ probability to the right of the value 210.81 is 0 (to three decimal places). This is definitive evidence that not all of β_1, β_2, and β_3 can be 0. Taken as a group, the variables x_1, x_2, and $x_3 = x_1^2$ definitely enhance one's ability to predict stack loss.

Table 9.11
ANOVA Table for Testing $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ for the Stack Loss Data

ANOVA Table (for testing $H_0: \beta_1 = \beta_2 = \beta_3 = 0$)

Source	*SS*	*df*	*MS*	*F*
Regression (on x_1, x_2, x_1^2)	799.80	3	266.60	210.81
Error	16.44	13	1.26	
Total	816.24	16		

Note also that the value of the coefficient of determination here can be calculated using sums of squares given in Table 9.11 as

$$R^2 = \frac{SSR}{SSTot} = \frac{799.80}{816.24} = .980$$

This is the value for R^2 advertised long ago in Example 5 in Chapter 4. Also, the error mean square, $MSE = 1.26$, is (as expected) exactly the value of s_{SF}^2 calculated earlier in this example.

It is a matter of simple algebra to verify that R^2 and the F statistic (9.62) are equivalent in the sense that

An expression for the F statistic (9.62) in terms of R^2

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \tag{9.63}$$

so the F test of hypothesis (9.60) can be thought of in terms of attaching a p-value to the statistic R^2. This is a valuable development, but it should be remembered that it is R^2 (rather than F) that has the direct interpretation as a measure of what fraction of raw variability the fitted equation accounts for. F and its associated p-value take account of the sample size n in a way that R^2 doesn't. They really measure statistical detectability rather than variation accounted for. This means that an equation that accounts for a fraction of observed variation that is relatively small by most standards can produce a very impressive (small) p-value. If this point is not clear, try using formula (9.63) to find the p-value for a situation where $n = 1{,}000$, $k = 4$, and $R^2 = .1$.

From Section 4.2 on, R^2 values have been used in this book for informal comparisons of various potential summary equations for a single data set. It turns out that it is sometimes possible to attach p-values to such comparisons through the use of the corresponding regression sums of squares and *another* F test.

Suppose that there are two different regression models for describing a data set—the first of the usual form (9.36) for k input variables $x_1, x_2, \ldots, x_k$,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

and the second being a specialization of the first where some p of the coefficients β (say, $\beta_{l_1}, \beta_{l_2}, \ldots, \beta_{l_p}$) are all 0 (i.e., a specialization not involving input variables $x_{l_1}, x_{l_2}, \ldots, x_{l_p}$). The first of these models will be called the **full regression model** and the second a **reduced regression model**. When one informally compares R^2 values for two such models, the comparison is essentially between *SSR* values, since the two R^2 values share the same denominator, *SSTot*. The two *SSR* values can be used to produce an observed level of significance for the comparison.

Under model the full model (9.36), the hypothesis

$$\mathrm{H}_0 : \beta_{l_1} = \beta_{l_2} = \cdots = \beta_{l_p} = 0 \tag{9.64}$$

(that the reduced model holds) can be tested against

$$\mathrm{H}_a : \text{ not } \mathrm{H}_0 \tag{9.65}$$

using the test statistic

F statistic for testing $H_0 : \beta_{l_1} = \cdots = \beta_{l_p} = 0$ in multiple regression

$$F = \frac{(SSR_f - SSR_r)/p}{SSE_f/(n-k-1)} \tag{9.66}$$

and an $F_{p,n-k-1}$ reference distribution, where large observed values of the test statistic constitute evidence against H_0 in favor of H_a. In expression (9.66), the "f" and "r" subscripts refer to the *full* and *reduced* regressions. The calculation of statistic (9.66) can be facilitated by expanding the basic ANOVA table for the full model (Table 9.10). Table 9.12 shows one form this can take.

Table 9.12
Expanded ANOVA Table for Testing $\mathrm{H}_0 : \beta_{l_1} = \beta_{l_2} = \cdots = \beta_{l_p} = 0$ in Multiple Regression

ANOVA Table (for testing $\mathrm{H}_0 : \beta_{l_1} = \beta_{l_2} = \cdots = \beta_{l_p} = 0$)				
Source	*SS*	*df*	*MS*	*F*
Regression (full)	SSR_f	k		
Regression (reduced)	SSR_r	$k-p$		
Regression (full | reduced)	$SSR_f - SSR_r$	p	$(SSR_f - SSR_r)/p$	$MSR_{f\|r}/MSE_f$
Error	SSE_f	$n-k-1$	$SSE_f/(n-k-1)$	
Total	*SSTot*	$n-1$		

Example 2 (*continued*)

In the nitrogen plant example, consider the comparison of the two possible descriptions of stack loss

$$y \approx \beta_0 + \beta_1 x_1 \tag{9.67}$$

(stack loss is approximately a linear function of airflow only) and

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 \tag{9.68}$$

(the description of stack loss that has been used throughout this section). Although a printout won't be included here to show it, it is a simple matter to verify that the fitting of expression (9.67) to the nitrogen plant data produces $SSR = 775.48$ and therefore $R^2 = .950$. Fitting expression (9.68), on the other hand, gives $SSR = 799.80$ and $R^2 = .980$. Since expression (9.67) is the specialization/reduction of expression (9.68) obtained by dropping the last $p = 2$ terms, the comparison of these two SSR (or R^2) values can be formalized with a p-value. A test of

$$\mathrm{H}_0 : \beta_2 = \beta_3 = 0$$

can be made in the (full) model (9.68). Table 9.13 organizes the calculation of the observed value of the statistic (9.66) for this problem. That is,

$$f = \frac{(799.80 - 775.48)/2}{16.44/13} = 9.7$$

When compared with tabled $F_{2,13}$ percentage points, the observed value of 9.7 is seen to produce a p-value between .01 and .001. There is strong evidence in the nitrogen plant data that an explanation of mean response in terms of expression (9.68) (pictured, for example, in Figure 4.15) is superior to one in terms of expression (9.67) (which could be pictured as a single linear mean response in x_1 for all x_2).

Table 9.13
ANOVA Table for Testing $\mathrm{H}_0 : \beta_2 = \beta_3 = 0$ in Model (9.68) for the Stack Loss Data

ANOVA Table (for testing $\mathrm{H}_0 : \beta_2 = \beta_3 = 0$)				
Source	*SS*	*df*	*MS*	*F*
Regression (x_1, x_2, x_1^2)	799.80	3		
Regression (x_1)	775.48	1		
Regression $(x_2, x_1^2 \mid x_1)$	24.32	2	12.16	9.7
Error (x_1, x_2, x_1^2)	16.44	13	1.26	
Total	816.24	16		

The F statistic (9.66) can be written in terms of R^2 values as

Alternative form of the F statistic for testing $H_0: \beta_{l_1} = \cdots = \beta_{l_p} = 0$

$$F = \frac{(R_{\text{f}}^2 - R_{\text{r}}^2)/p}{(1 - R_{\text{f}}^2)/(n - k - 1)} \tag{9.69}$$

so that the test of hypothesis (9.64) is indeed a way of attaching a p-value to the comparison of two R^2's. However, just as was remarked earlier concerning the test of hypothesis (9.60), it is the R^2's themselves that indicate how much additional variation a full model accounts for over a reduced model. The observed F value or associated p-value measures the extent to which that increase is distinguishable from background noise.

Interpreting full and reduced R^2's and the F test

p tests that single coefficients are 0 versus a test that p coefficients are all 0

To conclude this section, something needs to be said about the relationship between the tests of hypotheses (9.45) (with $\# = 0$), mentioned earlier, and the tests of hypothesis (9.64) based on the F statistic (9.66). When $p = 1$ (the full model contains only one more term than the reduced model), observed levels of significance based on statistic (9.66) are in fact equal to two-sided observed levels of significance based on $\# = 0$ versions of statistic (9.46). But for cases where $p \geq 2$, the tests of the hypotheses that individual β's are 0 (one at a time) are not an adequate substitute for the tests of hypothesis (9.64). For example, in the full model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{9.70}$$

testing

$$\text{H}_0: \beta_2 = 0 \tag{9.71}$$

and then testing

$$\text{H}_0: \beta_3 = 0 \tag{9.72}$$

need not be at all equivalent to making a single test of

$$\text{H}_0: \beta_2 = \beta_3 = 0 \tag{9.73}$$

This fact may at first seem paradoxical. But should the variables x_2 and x_3 be reasonably highly correlated in the data set, it is possible to get large p-values for tests of both hypothesis (9.71) and (9.72) and yet a tiny p-value for a test of hypothesis (9.73). The message carried by such an outcome is that (due to the fact that the variables x_2 and x_3 appear in the data set to be more or less equivalent) in the presence of x_1 and x_2, x_3 is not needed to model y. And in the presence of x_1 and x_3, x_2 is not needed to model y. But one or the other of the two variables x_2 and x_3 is needed to help model y even in the presence of x_1. So, the F test of hypothesis (9.64) is more than just a fancy version of several tests of hypotheses $\text{H}_0: \beta_l = 0$. It is an important addition to an engineer's curve- and surface-fitting tool kit.

Section 2 Exercises

1. Return to the situation of Chapter Exercise 2 of Chapter 4 and the carburetion study of Griffith and Tesdall. Consider an analysis of these data based on the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.
   (a) Find s_{SF} for these data. What does this intend to measure in the context of the engineering problem?
   (b) Plot both residuals versus x and the standardized residuals versus x. How much difference is there in the appearance of these two plots?
   (c) Give 90% individual two-sided confidence intervals for each of β_0, β_1, and β_2.
   (d) Give individual 90% two-sided confidence intervals for the mean elapsed time with a carburetor jetting size of 70 and then with a jetting size of 76.
   (e) Give simultaneous 90% two-sided confidence intervals for the two means indicated in part (d).
   (f) Give 90% lower prediction bounds for an additional elapsed time with a carburetor jetting size of 70 and also with a jetting size of 76.
   (g) Give approximate 95% lower tolerance bounds for 90% of additional elapsed times, first with a carburetor jetting size of 70 and then with a jetting size of 76.
   (h) Make an ANOVA table for testing $H_0: \beta_1 = \beta_2 = 0$ in the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. What is the meaning of this hypothesis in the context of the study and the quadratic model? What is the p-value?
   (i) Use a t statistic and test the null hypothesis $H_0: \beta_2 = 0$. What is the meaning of this hypothesis in the context of the study and the quadratic model?
2. Return to the situation of Exercise 2 of Section 4.2, and the chemithermomechanical pulp study of Miller, Shankar, and Peterson. Consider an analysis of the data there based on the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
   (a) Find s_{SF}. What does this intend to measure in the context of the engineering problem?
   (b) Plot both residuals and standardized residuals versus x_1, x_2, and $\hat{y}$. How much difference is there in the appearance of these pairs of plots?
   (c) Give 90% individual two-sided confidence intervals for all of β_0, β_1, and β_2.
   (d) Give individual 90% two-sided confidence intervals for the mean specific surface area, first when $x_1 = 9.0$ and $x_2 = 60$ and then when $x_1 = 10.0$ and $x_2 = 70$.
   (e) Give simultaneous 90% two-sided confidence intervals for the two means indicated in part (d).
   (f) Give 90% lower prediction bounds for the next specific surface area, first when $x_1 = 9.0$ and $x_2 = 60$ and then when $x_1 = 10.0$ and $x_2 = 70$.
   (g) Give approximate 95% lower tolerance bounds for 90% of specific surface areas, first when $x_1 = 9.0$ and $x_2 = 60$ and then when $x_1 = 10.0$ and $x_2 = 70$.
   (h) Make an ANOVA table for testing $H_0: \beta_1 = \beta_2 = 0$ in the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. What is the p-value?

9.3 Application of Multiple Regression in Response Surface Problems and Factorial Analyses

The discussions in Sections 4.1, 4.2, 9.1, and 9.2 have, we hope, given you a growing appreciation of the wide utility of regression methods in engineering. The purpose of this final section is to further expand your range of experience with multiple

regression by illustrating its usefulness in two additional contexts. First there is an illustration of how surface fitting is used in "response surface" (or response optimization) problems. Then there is a look at how regression has its applications even in factorial analyses.

9.3.1 Surface-Fitting and Response Surface Studies

Engineers are often called upon to address the following generic problem. A response or responses y are known to depend upon system variables $x_1, x_2, \ldots, x_k$. No simple physical theory is available for describing the dependence. Nevertheless, the variables $x_1, x_2, \ldots, x_k$ need adjustment to get good system behavior (as measured by the variables y). Multiple regression analysis and some specialized "response surface" considerations often prove effective in such problems.

Fitted linear and quadratic functions as empirical models

For one thing, **linear and quadratic functions** of $x_1, x_2, \ldots, x_k$ are often useful empirical descriptions of a relationship between $x_1, x_2, \ldots, x_k$ and y. The material in Sections 4.2 and 9.2 directly addresses fitting and inference for a linear approximate relationship like

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{9.74}$$

Response surfaces specified by equation (9.74) are "planar" (see again Figure 9.6 in this regard). When such surfaces fail to capture the nature of dependence of y on $x_1, x_2, \ldots, x_k$ because of their "lack of curvature," quadratic approximate relationships often prove effective. The general version of a quadratic equation for y in k variables x has k linear terms, k quadratic terms, and cross product terms for all pairs of x variables. For example, the general 3-variable quadratic response surface is specified by

$$\begin{aligned} y \approx \beta_0 &+ \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 \\ &+ \beta_8 x_1 x_3 + \beta_9 x_2 x_3 \end{aligned} \tag{9.75}$$

Gathering adequate data

One issue in using the k-variable version of quadratic function (9.75) is that of collecting adequate data to support the enterprise. 2^k factorial data are not sufficient. This is easy to see by considering the $k = 1$ case. Having data for only two different values of x_1, say $x_1 = 0$ and $x_1 = 1$, would not be adequate to support the fitting of

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \tag{9.76}$$

There are, as an arbitrary example, many different versions of equation (9.76) with $y = 5$ for $x_1 = 0$ and $y = 7$ for $x_1 = 1$, including

$$\begin{aligned} y &\approx 5 + 2x_1 + 0x_1^2 \\ y &\approx 5 - 8x_1 + 10x_1^2 \\ y &\approx 5 + 10x_1 - 8x_1^2 \end{aligned}$$

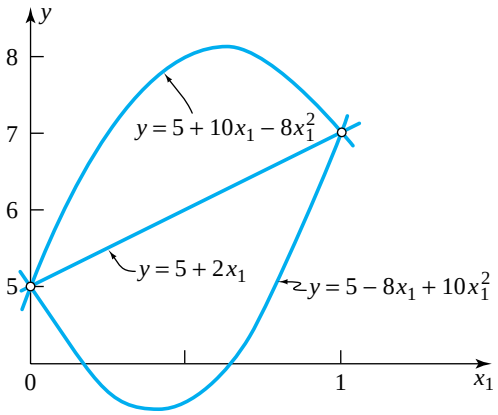

Figure 9.8 Plots of three different quadratic functions passing through the points $(x_1, y) = (0, 5)$ and $(x_1, y) = (1, 7)$

These three equations have plots with quite different shapes. The first is linear, the second is concave up with a minimum at $x_1 = .4$, and the third is concave down with a maximum at $x_1 = .625$. This is illustrated in Figure 9.8. The point is that data from at least three different x_1 values are needed in order to fit a one-variable quadratic equation.

What would happen if a regression program were used to fit equation (9.76) to a set of (x_1, y) data having only two different x_1 values in it? The program will typically refuse the user's request, perhaps fitting instead the simpler equation $y \approx \beta_0 + \beta_1 x_1$.

Exactly what *is* needed in the way of data in order to fit a k-variable quadratic equation is not easy to describe in elementary terms. 3^k factorial data are sufficient but for large k are really much more than are absolutely necessary. Statisticians have invested substantial effort in identifying patterns of $(x_1, x_2, \ldots, x_k)$ combinations that are both small (in terms of number of different combinations) and effective (in terms of facilitating precise estimation of the coefficients in a quadratic response function). See, for example, Section 7.2.2 of *Statistical Quality Assurance Methods for Engineers* by Vardeman and Jobe for a discussion of "central composite" plans often employed to gather data adequate to fit a quadratic. An early successful application of such a plan is described next.

Example 3

A Central Composite Study for Optimizing Bread Wrapper Seal Strength

The article "Sealing Strength of Wax-Polyethylene Blends" by Brown, Turner, and Smith (*Tappi*, 1958) contains an interesting central composite data set. The effects of the three process variables Seal Temperature, Cooling Bar Temperature, and % Polyethylene Additive on the seal strength y of a bread wrapper stock were studied. With the coding of the process variables indicated in Table 9.14, the data

Example 3
(*continued*)

Table 9.14
Coding of Three Process Variables in a Seal Strength Study

Factor		Variable	
A	Seal Temperature	$x_1 = \frac{t_1 - 255}{30}$	where t_1 is in °F
B	Cooling Bar Temperature	$x_2 = \frac{t_2 - 55}{9}$	where t_2 is in °F
C	Polyethylene Content	$x_3 = \frac{c - 1.1}{.6}$	where c is in %

Table 9.15
Seal Strengths Produced under 15 Different Sets of Process Conditions

x_1	x_2	x_3	Seal Strength, y (g/in.)
−1	−1	−1	6.6
1	−1	−1	6.9
−1	1	−1	7.9
1	1	−1	6.1
−1	−1	1	9.2
1	−1	1	6.8
−1	1	1	10.4
1	1	1	7.3
0	0	0	10.1
0	0	0	9.9
0	0	0	12.2
0	0	0	9.7
0	0	0	9.7
0	0	0	9.6
−1.682	0	0	9.8
1.682	0	0	5.0
0	−1.682	0	6.9
0	1.682	0	6.3
0	0	−1.682	4.0
0	0	1.682	8.6

in Table 9.15 were obtained. Notice that there are fewer than $3^3 = 27$ different (x_1, x_2, x_3) vectors in these data. (The central composite plan involves only 15 different combinations.)

If one fits a first-order (linear) model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{9.77}$$

to the data points listed in Table 9.15, a coefficient of determination of only $R^2 = .38$ is obtained, along with $s_{SF} = 1.79$. The pooled sample standard deviation (coming from the six points with $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$) is quite a bit smaller than s_{SF}—namely, $s_P = 1.00$. Between the small value of R^2 and the moderate difference between s_{SF} and s_P, there is already some indication that model (9.77) may be a poor description of the data. A residual analysis like those done in Section 4.2 would further confirm this.

On the other hand, fitting the expression (9.75) to the data in Table 9.15 produces the equation

$$\begin{aligned}\hat{y} = 10.165 &- 1.104x_1 + .0872x_2 + 1.020x_3 - .7596x_1^2 - 1.042x_2^2 \\ &- 1.148x_3^2 - .3500x_1x_2 - .5000x_1x_3 + .1500x_2x_3\end{aligned} \tag{9.78}$$

with a coefficient of determination of $R^2 = .86$ and $s_{SF} = 1.09$. At least on the basis of the two measures R^2 and s_{SF}, this quadratic description of seal strength seems much superior to a first-order description.

Plots and interpreting a fitted quadratic

For small values of k, the interpretation of a fitted quadratic response function can be facilitated through the use of various plots. One possibility is to **plot $\hat{y}$ versus a particular system variable** x, with values of any other system variables held fixed. This was the method used in Figure 4.15 for the nitrogen plant data, in Figure 4.16 (see page 158) for the lift/drag ratio data of Burris, and in Figure 9.8 of this section for the hypothetical one-variable quadratics. (It is also worth noting that in light of the inference material presented in Section 9.2, one can enhance such plots of $\hat{y}$ by adding error bars based on confidence limits for the means $\mu_{y|x_1,x_2,\ldots,x_k}$.)

A second kind of plot that can help in understanding a fitted quadratic function is the **contour plot**. A contour plot is essentially a topographic map. For a given pair of system variables (say x_1 and x_2) one can, for fixed values of all other input variables, sketch out the loci of points in the (x_1, x_2)-plane that produce several particular values of $\hat{y}$. Most statistical packages and engineering mathematics packages will make contour plots.

Example 3 (*continued*)

Figure 9.9 shows a series of five contour plots made using the fitted equation (9.78) for seal strength. These correspond to $x_3 = -2, -1, 0, 1$, and 2. The figure suggests that optimum predicted seal strength may be achievable for x_3 between 0 and 1, with x_1 between -2 and -1, and x_2 between 0 and 1.

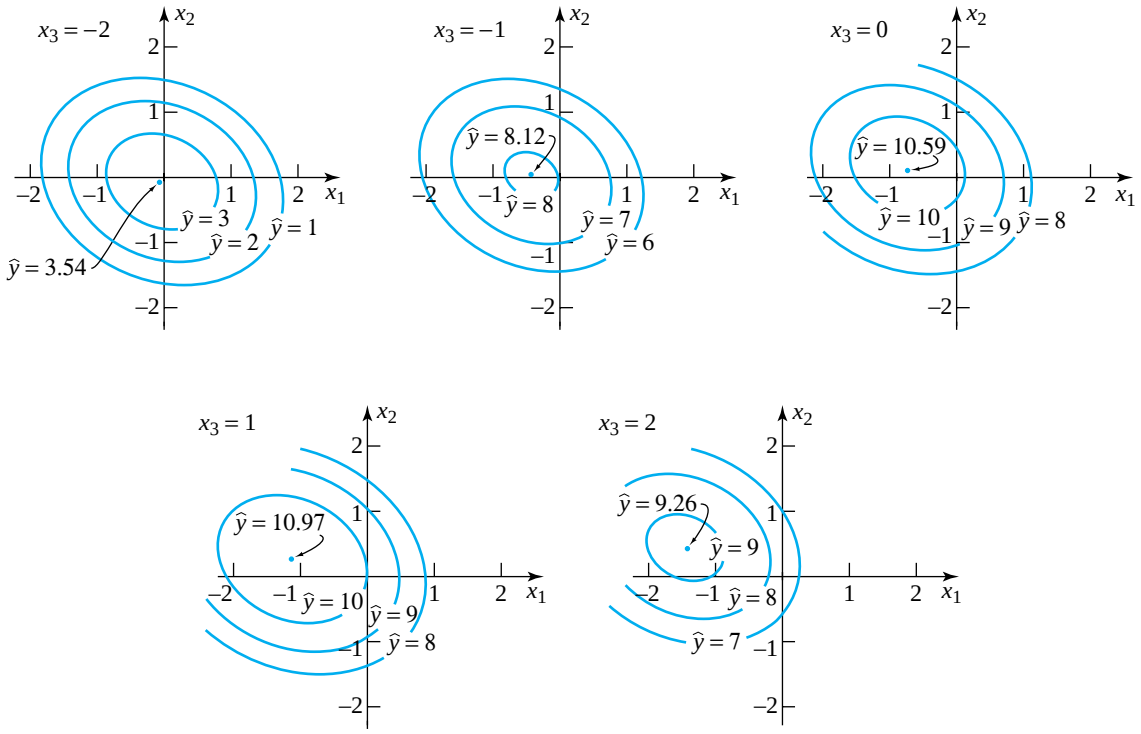

Figure 9.9 A series of contour plots for seal strength

Analytic interpretation of a fitted quadratic

Plotting is helpful in understanding a fitted quadratic primarily for small k. So it is important that there are also **analytical tools** that can be employed. To illustrate their character, consider the simple case of $k = 1$. The basic nature of the quadratic equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

is governed by b_2. For $b_2 > 0$ it describes a parabola opening up. For $b_2 < 0$ it describes a parabola opening down. And for $b_2 = 0$ it describes a line. Provided $b_2 \neq 0$ the value

$$x_1 = -\frac{b_1}{2b_2}$$

produces the minimum ($b_2 > 0$) or maximum ($b_2 < 0$) value of $\hat{y}$. Something like this story is also true for $k > 1$.

It is necessary to use some matrix notation to say what happens for $k > 1$. Temporarily modify the way the b's are subscripted as follows. The meaning of b_0 will remain unchanged. b_1 through b_k will be the coefficients for the k system

variables x_1 through x_k. b_{11} through b_{kk} will be the coefficients for the k squares x_1^2 through x_k^2. And for each $i \neq j$, b_{ij} will be the coefficient of the $x_i x_j$ cross product. One can define a $k \times 1$ vector $\boldsymbol{b}$ and a $k \times k$ matrix $\boldsymbol{B}$ as

Vector of linear coefficients and matrix of quadratic coefficients

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & \frac{1}{2}b_{12} & \cdots & \frac{1}{2}b_{1k} \\ \frac{1}{2}b_{12} & b_{22} & \cdots & \frac{1}{2}b_{2k} \\ \vdots & \vdots & & \vdots \\ \frac{1}{2}b_{1k} & \frac{1}{2}b_{2k} & \cdots & b_{kk} \end{bmatrix}$$

With

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

Provided the matrix $\boldsymbol{B}$ is nonsingular, the corresponding k-variable quadratic then has a **stationary point** (i.e., a point at which first partial derivatives with respect to $x_1, x_2, \ldots, x_k$ are all 0) where

Location of a stationary point for a k-variable fitted quadratic

$$\boldsymbol{x} = -\tfrac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b} \tag{9.79}$$

And depending upon the nature of $\boldsymbol{B}$, the stationary point will be either **a minimum, a maximum, or a saddle point** of the fitted response. (Moving away from a saddle point in some directions produces an increase in $\hat{y}$, while moving away in other directions produces a decrease.)

It is the **eigenvalues** of $\boldsymbol{B}$ that are critical in determining the shape of the fitted quadratic surface. The eigenvalues of $\boldsymbol{B}$ are the k solutions of the equation (in λ)

Equation solved by the eigenvalues λ of the matrix $\boldsymbol{B}$

$$\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = 0 \tag{9.80}$$

where $\boldsymbol{I}$ is the identity matrix. (Most statistical analysis packages and engineering mathematics packages will compute eigenvalues quite painlessly.)

When all solutions to equation (9.80) are positive, a fitted quadratic is bowl-shaped up and has a minimum at the point (9.79). When all solutions to equation

(9.80) are negative, a fitted quadratic is bowl-shaped down and has a maximum at the point (9.79). When some solutions to equation (9.80) are positive and some are negative, the fitted quadratic surface has neither a maximum nor minimum (unless one restricts attention to some bounded region of $\boldsymbol{x}$ vectors).

Printout 3 Analysis of the Fitted Quadratic for the Bread Wrapper Data (*Example 3*)

```
MTB > Read 3 3 M1.
DATA> -.7596 -.175 -.250
DATA> -.175 -1.042 .075
DATA> -.250 .075 -1.148
      3 rows read.
MTB > Read 3 1 M2.
DATA> -1.104
DATA> .0872
DATA> 1.020
      3 rows read.
MTB > Eigen M1 C1.
MTB > Print C1.

Data Display


C1
  -1.27090  -1.11680  -0.56190

MTB > Invert M1 M3.
MTB > Multiply M3 M2 M4.
MTB > Multiply M4 -.5 M5.
MTB > Print  M5.

Data Display


 Matrix M5

-1.01104
  0.26069
  0.68146
```

Example 3 (*continued*)

Printout 3 illustrates the use of MINITAB in the analytic investigation of the nature of the fitted surface (9.78) in the bread wrapper seal strength study. The printout shows the three eigenvalues of $\boldsymbol{B}$ to be negative. The fitted seal strength therefore has a maximum. This maximum is predicted to occur at the combination of values $x_1 = -1.01$, $x_2 = .26$, and $x_3 = .68$. (The MINITAB matrix functions used to make the printout are under the "Calc/Matrices" menu, and the display routine is under the "Manip/Display Data" menu.)

The discussion of response surface studies in this subsection isn't intended to be complete. Whole books, like, for example, Box and Draper's *Empirical Model-Building and Response Surfaces*, have been written on the subject. (Section 9.3 of Vardeman's *Statistics for Engineering Problem Solving* contains a more complete discussion than the present one, is still short of a book-length treatment.) We hope, however, this brief look at the topic suffices to indicate its importance to engineering practice.

9.3.2 Regression and Factorial Analyses

Many of the factorial inference methods discussed in this book are applicable only in balanced-data situations. For example, remember that the use of the reverse Yates algorithm to fit few-effects 2^p factorial models and the methods of interval-oriented inference for 2^p studies under few-effects models discussed in Section 8.2 are limited to balanced-data applications.

But by accident if not by design, an engineer will eventually face the analysis of unbalanced factorial data. Happily enough, this can be accomplished through use of the multiple regression formulas provided in Section 9.2. This subsection shows how factorial analyses can be thought of in multiple regression terms. It begins with a discussion of two-way factorial cases and then considers three-way (and higher) situations.

The basic multiple regression model equation used in Section 9.2,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \tag{9.81}$$

looks deceptively simple. With proper choice of the inputs x, versions of it can be used in a wide variety of contexts, including factorial analyses. For purposes of illustration, consider the case of a complete two-way factorial study with $I = 3$ levels of factor A and $J = 3$ levels of factor B. In the usual two-way factorial notation introduced in Definitions 1 and 2 of Chapter 8, the basic constraints on the main effects and two-factor interactions are $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, and $\sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} = 0$. These imply that the $I \cdot J = 3 \cdot 3 = 9$ different mean responses in such a study,

$$\mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \alpha\beta_{ij} \tag{9.82}$$

can be written as displayed in Table 9.16.

At first glance, the advantage of writing out these mean responses in terms of only effects corresponding to the first 2 ($= I - 1$) levels of A and first 2 ($= J - 1$) levels of B is not obvious. But doing so expresses the 9 ($= I \cdot J$) different means in terms of only as many different parameters as there are means, and helps one find a regression-type analog of expression (9.82).

Notice first that $\mu_{\cdot\cdot}$ appears in each mean response listed and therefore plays a role much like that of the intercept term β_0 in a regression model. Further, the two A main effects, α_1 and α_2, appear with positive signs when (respectively) $i = 1$

Table 9.16
Mean Responses in a 3^2 Factorial Study

i, Level of A	j, Level of B	Mean Response
1	1	$\mu_{..} + \alpha_1 + \beta_1 + \alpha\beta_{11}$
1	2	$\mu_{..} + \alpha_1 + \beta_2 + \alpha\beta_{12}$
1	3	$\mu_{..} + \alpha_1 - \beta_1 - \beta_2 - \alpha\beta_{11} - \alpha\beta_{12}$
2	1	$\mu_{..} + \alpha_2 + \beta_1 + \alpha\beta_{21}$
2	2	$\mu_{..} + \alpha_2 + \beta_2 + \alpha\beta_{22}$
2	3	$\mu_{..} + \alpha_2 - \beta_1 - \beta_2 - \alpha\beta_{21} - \alpha\beta_{22}$
3	1	$\mu_{..} - \alpha_1 - \alpha_2 + \beta_1 - \alpha\beta_{11} - \alpha\beta_{21}$
3	2	$\mu_{..} - \alpha_1 - \alpha_2 + \beta_2 - \alpha\beta_{12} - \alpha\beta_{22}$
3	3	$\mu_{..} - \alpha_1 - \alpha_2 - \beta_1 - \beta_2 + \alpha\beta_{11} + \alpha\beta_{12} + \alpha\beta_{21} + \alpha\beta_{22}$

or 2 but with negative signs when $i = 3\ (= I)$. In a similar manner, the first two B main effects, β_1 and β_2, appear with positive signs when (respectively) $j = 1$ or 2 but with negative signs when $j = 3\ (= J)$. If one thinks of the four A and B main effects used in Table 9.16 in terms of coefficients β in a regression model, it soon becomes clear how to invent "system variables" x to make the regression coefficients β appear with correct signs in the expressions for means μ_{ij}. That is, define four **dummy variables**

$$x_1^{\mathrm{A}} = \begin{cases} 1 & \text{if the response } y \text{ is from level 1 of A} \\ -1 & \text{if the response } y \text{ is from level 3 of A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2^{\mathrm{A}} = \begin{cases} 1 & \text{if the response } y \text{ is from level 2 of A} \\ -1 & \text{if the response } y \text{ is from level 3 of A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_1^{\mathrm{B}} = \begin{cases} 1 & \text{if the response } y \text{ is from level 1 of B} \\ -1 & \text{if the response } y \text{ is from level 3 of B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2^{\mathrm{B}} = \begin{cases} 1 & \text{if the response } y \text{ is from level 2 of B} \\ -1 & \text{if the response } y \text{ is from level 3 of B} \\ 0 & \text{otherwise} \end{cases}$$

Then, making the correspondences indicated in Table 9.17, $\mu_{..} + \alpha_i + \beta_j$ can be written in regression notation as

$$\beta_0 + \beta_1 x_1^{\mathrm{A}} + \beta_2 x_2^{\mathrm{A}} + \beta_3 x_1^{\mathrm{B}} + \beta_4 x_2^{\mathrm{B}}$$

Table 9.17
Correspondences between Regression Coefficients and the Grand Mean and Main Effects in a 3^2 Factorial Study

Regression Coefficient	Corresponding 3×3 Factorial Effect
β_0	$\mu_{..}$
β_1	α_1
β_2	α_2
β_3	β_1
β_4	β_2

What is more, since the x's used here take only the values -1, 0, and 1, so also do their products. And taken in pairs (one x^{A} variable with one x^{B} variable), their products produce the correct (-1, 0, or 1) multipliers for the 2-factor interactions $\alpha\beta_{11}$, $\alpha\beta_{12}$, $\alpha\beta_{21}$, and $\alpha\beta_{22}$ appearing in Table 9.16. That is, if one thinks of the interactions $\alpha\beta_{ij}$ in terms of regression coefficients β, with the additional correspondences listed in Table 9.18, the entire expression (9.82) can be written in regression notation as

$$\mu_{y|x_1^{\mathrm{A}},x_2^{\mathrm{A}},x_1^{\mathrm{B}},x_2^{\mathrm{B}}} = \beta_0 + \beta_1 x_1^{\mathrm{A}} + \beta_2 x_2^{\mathrm{A}} + \beta_3 x_1^{\mathrm{B}} + \beta_4 x_2^{\mathrm{B}} + \beta_5 x_1^{\mathrm{A}} x_1^{\mathrm{B}} + \beta_6 x_1^{\mathrm{A}} x_2^{\mathrm{B}} + \beta_7 x_2^{\mathrm{A}} x_1^{\mathrm{B}} + \beta_8 x_2^{\mathrm{A}} x_2^{\mathrm{B}} \quad \textbf{(9.83)}$$

By rewriting the factorial-type expression (9.82) as a regression-type expression (9.83) it is then obvious how to fit few-effects models and do inference under those models even for unbalanced data. Nowhere in Section 9.2 was there any requirement that the data set be balanced. So the methods there can be used (employing properly constructed x variables and properly interpreting a corresponding regression printout) to fit reduced versions of model (9.83) and make confidence, prediction, and tolerance intervals under those reduced models.

Table 9.18
Correspondence between Regression Coefficients and Interactions in a 3^2 Factorial Study

Regression Coefficient	Corresponding 3×3 Factorial Effect
β_5	$\alpha\beta_{11}$
β_6	$\alpha\beta_{12}$
β_7	$\alpha\beta_{21}$
β_8	$\alpha\beta_{22}$

The general $I \times J$ two-way factorial version of this story is similar. One defines $I - 1$ factor A dummy variables $x_1^A, x_2^A, \ldots, x_{I-1}^A$ according to

$I - 1$ dummy variables for factor A

$$x_i^A = \begin{cases} 1 & \text{if the response } y \text{ is from level } i \text{ of A} \\ -1 & \text{if the response } y \text{ is from level } I \text{ of A} \\ 0 & \text{otherwise} \end{cases} \tag{9.84}$$

and $J - 1$ factor B dummy variables $x_1^B, x_2^B, \ldots, x_{J-1}^B$ according to

$J - 1$ dummy variables for factor B

$$x_j^B = \begin{cases} 1 & \text{if the response } y \text{ is from level } j \text{ of B} \\ -1 & \text{if the response } y \text{ is from level } J \text{ of B} \\ 0 & \text{otherwise} \end{cases} \tag{9.85}$$

Multiple regression and two-way factorial analyses

and uses a regression program to do the computations. Estimated regression coefficients of x_i^A or x_j^B variables alone are estimated main effects, while those for $x_i^A x_j^B$ cross products are estimated 2-factor interactions.

Example 4
(*Examples 7, Chapter 4, and 1, Chapter 8, revisited—see pages 163, 547*)

A Factorial Analysis of Unbalanced Wood Joint Strength Data Using a Regression Program

Consider again the wood joint strength study of Kotlers, MacFarland, and Tomlinson. The discussion in Section 8.1 showed that if only the wood types pine and oak are considered, a no-interaction description of joint strength for butt, beveled, and lap joints might be appropriate. The corresponding part of the (originally 3×3 factorial) data of Kotlers, MacFarland, and Tomlinson is given here in Table 9.19.

Table 9.19
Strengths of 11 Wood Joints

			B *Wood Type*	
			1 (Pine)	2 (Oak)
A	*Joint Type*	1 (Butt)	829, 596	1169
		2 (Beveled)	1348, 1207	1518, 1927
		3 (Lap)	1000, 859	1295, 1561

Table 9.20
Joint Strength Data Prepared for a Factorial Analysis Using a Regression Program

i, Joint Type	j, Wood Type	x_1^A	x_2^A	x_1^B	y
1	1	1	0	1	829, 596
1	2	1	0	−1	1169
2	1	0	1	1	1348, 1207
2	2	0	1	−1	1518, 1927
3	1	−1	−1	1	1000, 859
3	2	−1	−1	−1	1295, 1561

Notice that because these data are unbalanced (due to the unfortunate loss of one butt/oak response), it is not possible to fit a no-interaction model to these data by simply adding together fitted effects (defined in Section 4.3) or to use anything said in Chapter 8 to make inferences based on such a model. But it *is* possible to use the dummy variable regression approach based on formulas (9.84) and (9.85) to do so.

Consider the regression-data-set version of Table 9.19 given in Table 9.20. Printouts 4 and 5 show the results of fitting the two regression models

$$y = \beta_0 + \beta_1 x_1^A + \beta_2 x_2^A + \beta_3 x_1^B + \beta_4 x_1^A x_1^B + \beta_5 x_2^A x_1^B + \epsilon \quad \textbf{(9.86)}$$

$$y = \beta_0 + \beta_1 x_1^A + \beta_2 x_2^A + \beta_3 x_1^B + \epsilon \quad \textbf{(9.87)}$$

to the data of Table 9.20. Printout 4 corresponding to model (9.86) is the full model or $\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij}$ description of the data. For that regression run, the reader should verify the correspondences between fitted regression coefficients b and fitted effects (defined in Section 4.3), listed in Table 9.21. (For example,

Table 9.21
Correspondence between Fitted Regression Coefficients and Fitted Factorial Effects for the Wood Joint Strength Data

Fitted Regression Coefficient	Value	Corresponding Fitted Effect
b_0	1206.5	$\bar{y}_{..}$
b_1	−265.75	a_1
b_2	293.50	a_2
b_3	−233.33	b_1
b_4	5.08	ab_{11}
b_5	10.83	ab_{21}

Example 4 (*continued*)

$\bar{y}_{\cdot\cdot} = 1206.5$ and $\bar{y}_{1\cdot} = 940.75$, so $a_1 = 940.75 - 1206.5 = -265.75$, which is the value of the fitted regression coefficient b_1.)

Model (9.86), like the two-way model (8.4) of Section 8.1, represents no restriction or simplification of the basic one-way model. So least squares estimates of parameters that are linear combinations of underlying means are simply the same linear combinations of sample means. Further, the fitted y values are (as expected) simply the sample means $\bar{y}_{ij}$.

Printout 5 corresponding to model (9.87) is the $\mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j$ description of the data. The fitted regression coefficients b for model (9.87) are not equal to the (full-model) fitted factorial effects defined in Section 4.3. (The b's are least squares estimates of the underlying effects for the no-interaction model. When factorial data are unbalanced, these are not necessarily equal to the quantities defined in Section 4.3. For example, b_1 from Printout 5 is -264.48, which is the least squares estimate of α_1 in a no-interaction model but differs from $a_1 = -264.75$.) In a similar vein, the fitted responses are neither sample means nor sums of $\bar{y}_{\cdot\cdot}$ plus the full-model fitted main effects defined in Section 4.3. (Of course, since the x variables take only values -1, 0, and 1, the fitted responses *are* sums and differences of the least squares estimates of the underlying parameters $\mu_{\cdot\cdot}, \alpha_1, \alpha_2, \beta_1$ in the no-interaction model.)

Inference under model (9.86) is simply inference under the usual one-way normal model, and all of Sections 7.1 through 7.4 and 8.1 can be used. It is then reassuring that on Printout 4, $s_{\text{SF}} = s_{\text{P}} = 182.2$ and that (for example) for butt joints and pine wood (levels 1 of both A and B), the estimated standard deviation for $\hat{y} = \bar{y}_{11}$ is

$$128.9 = s_{\text{SF}} \cdot A = \frac{s_{\text{P}}}{\sqrt{n_{11}}} = \frac{182.2}{\sqrt{2}}$$

To illustrate how inference under a no-interaction model would proceed for the unbalanced 3×2 factorial joint strength data, consider making a 95% two-sided confidence interval for the mean strength of butt/pine joints and then a 90% lower prediction bound for the strength of a single joint of the same kind. Note that for data point 1 (a butt/pine observation) on Printout 5, $\hat{y} = 708.7$ and $s_{\text{SF}} \cdot A = 94.8$, where $s_{\text{SF}} = 154.7$ has seven associated degrees of freedom. So from formula (9.53) of Section 9.2 (page 686), two-sided 95% confidence limits for mean butt/pine joint strength are

$$708.7 \pm 2.365(94.8)$$

that is,

484.5 psi and 932.9 psi

Similarly, using formula (9.56) on page 689, a 90% lower prediction limit for a single additional butt/pine joint strength is

$$708.7 - 1.415\sqrt{(154.7)^2 + (94.8)^2} = 452.0 \text{ psi}$$

From these two calculations, it should be clear that other methods from Section 9.2 could be used here as well. The reader should have no trouble finding and using residuals and standardized residuals for the no-interaction model based on formulas (9.39) and (9.42), giving simultaneous confidence intervals for all six mean responses under the no-interaction model using formula (9.54) or giving one-sided tolerance bounds for certain joint/wood combinations under the no-interaction model using formula (9.58) or (9.59).

Printout 4 Multiple Regression Version of the With-Interactions Factorial Analysis of Joint Strength (*Example 4*)

```
Data Display


 Row   xa1   xa2   xb1      y

   1     1     0     1    829
   2     1     0     1    596
   3     1     0    -1   1169
   4     0     1     1   1348
   5     0     1     1   1207
   6     0     1    -1   1518
   7     0     1    -1   1927
   8    -1    -1     1   1000
   9    -1    -1     1    859
  10    -1    -1    -1   1295
  11    -1    -1    -1   1561


Regression Analysis


The regression equation is
y = 1207 - 266 xa1 + 294 xa2 - 233 xb1 + 5.1 xa1*xb1 + 10.8 xa2*xb1

Predictor       Coef       StDev          T        P
Constant     1206.50       56.82      21.23    0.000
xa1          -265.75       85.91      -3.09    0.027
xa2           293.50       77.43       3.79    0.013
xb1          -233.33       56.82      -4.11    0.009
xa1*xb1         5.08       85.91       0.06    0.955
xa2*xb1        10.83       77.43       0.14    0.894

S = 182.2       R-Sq = 88.5%     R-Sq(adj) = 77.1%
```

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     1283527      256705      7.73    0.021
Residual Error     5      166044       33209
Total             10     1449571

Source       DF      Seq SS
xa1           1      120144
xa2           1      577927
xb1           1      583908
xa1*xb1       1         897
xa2*xb1       1         650

Obs        xa1           y         Fit   StDev Fit    Residual    St Resid
  1       1.00      829.0       712.5       128.9       116.5        0.90
  2       1.00      596.0       712.5       128.9      -116.5       -0.90
  3       1.00     1169.0      1169.0       182.2        -0.0           * X
  4       0.00     1348.0      1277.5       128.9        70.5        0.55
  5       0.00     1207.0      1277.5       128.9       -70.5       -0.55
  6       0.00     1518.0      1722.5       128.9      -204.5       -1.59
  7       0.00     1927.0      1722.5       128.9       204.5        1.59
  8      -1.00     1000.0       929.5       128.9        70.5        0.55
  9      -1.00      859.0       929.5       128.9       -70.5       -0.55
 10      -1.00     1295.0      1428.0       128.9      -133.0       -1.03
 11      -1.00     1561.0      1428.0       128.9       133.0        1.03

X denotes an observation whose X value gives it large influence.
```

Printout 5 Multiple Regression Version of the No-Interactions Factorial Analysis of Joint Strength (*Example 4*)

```
Regression Analysis


The regression equation is
y = 1207 - 264 xa1 + 293 xa2 - 234 xb1

Predictor       Coef       StDev          T        P
Constant     1207.14       47.38      25.48    0.000
xa1          -264.48       70.62      -3.74    0.007
xa2           292.86       65.11       4.50    0.003
xb1          -233.97       47.38      -4.94    0.002

S = 154.7       R-Sq = 88.4%     R-Sq(adj) = 83.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1281980      427327     17.85    0.001
Residual Error     7      167591       23942
Total             10     1449571

Source       DF      Seq SS
xa1           1      120144
xa2           1      577927
xb1           1      583908
```

Obs	xa1	y	Fit	StDev Fit	Residual	St Resid
1	1.00	829.0	708.7	94.8	120.3	0.98
2	1.00	596.0	708.7	94.8	-112.7	-0.92
3	1.00	1169.0	1176.6	109.4	-7.6	-0.07
4	0.00	1348.0	1266.0	90.7	82.0	0.65
5	0.00	1207.0	1266.0	90.7	-59.0	-0.47
6	0.00	1518.0	1734.0	90.7	-216.0	-1.72
7	0.00	1927.0	1734.0	90.7	193.0	1.54
8	-1.00	1000.0	944.8	90.7	55.2	0.44
9	-1.00	859.0	944.8	90.7	-85.8	-0.68
10	-1.00	1295.0	1412.7	90.7	-117.7	-0.94
11	-1.00	1561.0	1412.7	90.7	148.3	1.18

Dummy variables for regression analysis of p-way factorials

The pattern of analysis set out for two-way factorials carries over quite naturally to three-way and higher factorials. To use a multiple regression program to fit and make inferences based on simplified versions of the p-way factorial model, proceed as follows. $I-1$ dummy variables $x_1^{\mathrm{A}}, x_2^{\mathrm{A}}, \ldots, x_{I-1}^{\mathrm{A}}$ are defined (as before) to carry information about I levels of factor A, $J-1$ dummy variables $x_1^{\mathrm{B}}, x_2^{\mathrm{B}}, \ldots, x_{J-1}^{\mathrm{B}}$ are defined (as before) to carry information about J levels of factor B, $K-1$ dummy variables $x_1^{\mathrm{C}}, x_2^{\mathrm{C}}, \ldots, x_{K-1}^{\mathrm{C}}$ are defined to carry information about K levels of factor C, . . . , etc. Products of pairs of these, one each from the groups representing two different factors, carry information about 2-factor interactions of the factors. Products of triples of these, one each from the groups representing three different factors, carry information about 3-factor interactions of the factors. And so on.

When something short of the largest possible regression model is fitted to an unbalanced factorial data set, the estimated coefficients b that result are the least squares estimates of the underlying factorial effects *in the few-effects model*. (Usually, these differ somewhat from the (full-model) fitted effects defined in Section 4.3.) All of the regression machinery of Section 9.2 can be applied to create fitted values, residuals, and standardized residuals; to plot these to do model checking; to make confidence intervals for mean responses; and to create prediction and tolerance intervals.

When the regression with dummy variables approach is used as just described, the fitted coefficients b correspond to fitted effects for the levels 1 through $I-1$, $J-1$, $K-1$, etc. of the factors. For two-level factorials, this means that the fitted coefficients are estimated factorial effects for the "all low" treatment combination. However, because of extensive use of the Yates algorithm in this text, you will probably think first in terms of the 2^p factorial effects for the "all high" treatment combination.

Alternative choice of x variables for regression analysis of 2^p factorials

Two sensible courses of action then suggest themselves for the analysis of unbalanced 2^p factorial data. You can proceed exactly as just indicated, using dummy variables x_1^{A}, x_1^{B}, x_1^{C}, etc. and various products of the same, taking care to remember to interpret b's as "all low" fitted effects and subsequently to switch signs as appropriate to get "all high" fitted effects. The other possibility is to depart slightly from the program laid out for general p-way factorials in 2^p cases: Instead

of using the variables x_1^{A}, x_1^{B}, x_1^{C}, etc. and their products when doing regression, one may use the variables

$$x_2^{\mathrm{A}} = -x_1^{\mathrm{A}} = \begin{cases} 1 & \text{if the response } y \text{ is from the high level of A} \\ -1 & \text{if the response } y \text{ is from the low level of A} \end{cases}$$

$$x_2^{\mathrm{B}} = -x_1^{\mathrm{B}} = \begin{cases} 1 & \text{if the response } y \text{ is from the high level of B} \\ -1 & \text{if the response } y \text{ is from the low level of B} \end{cases}$$

$$x_2^{\mathrm{C}} = -x_1^{\mathrm{C}} = \begin{cases} 1 & \text{if the response } y \text{ is from the high level of C} \\ -1 & \text{if the response } y \text{ is from the low level of C} \end{cases}$$

etc. and their products when doing regression. When the variables x_2^{A}, x_2^{B}, x_2^{C}, etc. are used, the fitted b's are the estimated "all high" 2^p factorial effects.

Example 5
(*Example 4, Chapter 8, revisited—page 569*)

A Factorial Analysis of Unbalanced 2^3 Power Requirement Data Using Regression

Return to the situation of the 2^3 metalworking power requirement study of Miller. The original data set (given in Table 8.8) is balanced, with the common sample size being $m = 4$. For the sake of illustrating how regression with dummy variables can be used in the analysis of unbalanced higher-way factorial data, consider artificially unbalancing Miller's data by supposing that the first data point appearing in Table 8.8 has gotten lost. The portion of Miller's data that will be used here is then given in Table 9.22.

Table 9.22
Dynamometer Readings for 2^3 Treatment Combinations

Tool Type	Bevel Angle	Type of Cut	y, Dynamometer Reading (mm)
1	15°	continuous	26.5, 30.5, 27.0
2	15°	continuous	28.0, 28.5, 28.0, 25.0
1	30°	continuous	28.5, 28.5, 30.0, 32.5
2	30°	continuous	29.5, 32.0, 29.0, 28.0
1	15°	interrupted	28.0, 25.0, 26.5, 26.5
2	15°	interrupted	24.5, 25.0, 28.0, 26.0
1	30°	interrupted	27.0, 29.0, 27.5, 27.5
2	30°	interrupted	27.5, 28.0, 27.0, 26.0

For this slightly altered data set, the Yates algorithm produces the fitted effects

$$\begin{array}{lll} a_2 = -.2656 & ab_{22} = .0469 & abc_{222} = -.0469 \\ b_2 = .8281 & ac_{22} = -.0469 & \\ c_2 = -.9531 & bc_{22} = -.2031 & \end{array}$$

and $s_P = 1.51$ with $\nu = 23$ associated degrees of freedom. Formula (8.12) (page 575) of Section 8.2 then shows that (say) two-sided 90% confidence intervals for effects have plus-and-minus parts

$$\pm 1.714(1.51)\frac{1}{2^3}\sqrt{\frac{7}{4}+\frac{1}{3}} = \pm .47$$

Just as in Example 4 in Chapter 8, where all $n = 32$ data points were used, one might thus judge only the B and C main effects to be clearly larger than background noise.

Printout 6 supports exactly these conclusions. This regression run was made using all seven of the terms

$$x_2^A,\ x_2^B,\ x_2^C,\ x_2^A x_2^B,\ x_2^A x_2^C,\ x_2^B x_2^C, \quad \text{and} \quad x_2^A x_2^B x_2^C$$

(i.e., using the *full model* in regression terminology and the *unrestricted* 2^3 *factorial model* in the terminology of Section 8.2). On Printout 6, one can identify the fitted regression coefficients b with the fitted factorial effects in the pairs indicated in Table 9.23.

Table 9.23
Correspondence Between Fitted Regression Coefficients and Fitted Factorial Effects for the Regression Run of Printout 6

Fitted Regression Coefficient	Fitted Factorial Effect
b_0	$\bar{y}_{\ldots}$
b_1	a_2
b_2	b_2
b_3	c_2
b_4	ab_{22}
b_5	ac_{22}
b_6	bc_{22}
b_7	abc_{222}

Example 5 (*continued*)

Analysis of the data of Table 9.22 based on a full factorial model

$$y_{ijkl} = \mu_{\cdots} + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkl}$$

that is,

$$y_i = \beta_0 + \beta_1 x_{2i}^{\mathrm{A}} + \beta_2 x_{2i}^{\mathrm{B}} + \beta_3 x_{2i}^{\mathrm{C}} + \beta_4 x_{2i}^{\mathrm{A}} x_{2i}^{\mathrm{B}} + \beta_5 x_{2i}^{\mathrm{A}} x_{2i}^{\mathrm{C}} + \beta_6 x_{2i}^{\mathrm{B}} x_{2i}^{\mathrm{C}} + \beta_7 x_{2i}^{\mathrm{A}} x_{2i}^{\mathrm{B}} x_{2i}^{\mathrm{C}} + \epsilon_i$$

is a logical first step. Based on that step, it seems desirable to fit and draw inferences based on a "B and C main effects only" description of y. Since the data in Table 9.22 are unbalanced, the naive use of the reverse Yates algorithm with the (full-model) fitted effects will not produce appropriate fitted values. $\bar{y}_{\cdots}$, b_2, and c_2 are simply *not* the least squares estimates of $\mu_{\cdots}$, β_2, and γ_2 for the "B and C main effects only" model in this unbalanced data situation.

However, what can be done is to fit the reduced regression model

$$y_i = \beta_0 + \beta_2 x_{2i}^{\mathrm{B}} + \beta_3 x_{2i}^{\mathrm{C}} + \epsilon_i$$

to the data. Printout 7 represents the use of this technique. Locate on that printout the (reduced-model) estimates of the factorial effects $\mu_{\cdots}$, β_2, and γ_2 and note that they differ somewhat from $\bar{y}_{\cdots}$, b_2, and c_2 as defined in Section 4.3 and displayed on Printout 6. Note also that the four different possible fitted mean responses, along with their estimated standard deviations, are as given in Table 9.24.

The values in Table 9.24 can be used in the formulas of Section 9.2 to produce confidence intervals for the four mean responses, prediction intervals, tolerance intervals, and so on based on the "B and C main effects only" model. All of this can be done despite the fact that the data of Table 9.22 are unbalanced.

Table 9.24
Fitted Values and Their Estimated Standard Deviations for a "B and C Main Effects Only" Analysis of the Unbalanced Power Requirement Data

Bevel Angle	x_2^{B}	Type of Cut	x_2^{C}	$\hat{y}$	$s_{\mathrm{SF}} \cdot A$
15°	−1	continuous	−1	27.88	.46
30°	1	continuous	−1	29.54	.44
15°	−1	interrupted	1	25.98	.44
30°	1	interrupted	1	27.64	.44

Printout 6 Multiple Regression Version of the With-Interactions Factorial Analysis of Power Requirement (*Example 5*)

```
Regression Analysis

The regression equation is
y = 27.8 - 0.266 xa2 + 0.828 xb2 - 0.953 xc2 + 0.047 xa*xb - 0.047 xa*xc
          - 0.203 xb*xc - 0.047 xa*xb*xc

Predictor       Coef       StDev          T        P
Constant     27.7656      0.2731     101.68    0.000
xa2          -0.2656      0.2731      -0.97    0.341
xb2           0.8281      0.2731       3.03    0.006
xc2          -0.9531      0.2731      -3.49    0.002
xa*xb         0.0469      0.2731       0.17    0.865
xa*xc        -0.0469      0.2731      -0.17    0.865
xb*xc        -0.2031      0.2731      -0.74    0.465
xa*xb*xc     -0.0469      0.2731      -0.17    0.865

S = 1.514       R-Sq = 51.0%     R-Sq(adj) = 36.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         7      54.748       7.821      3.41    0.012
Residual Error    23      52.687       2.291
Total             30     107.435

Source       DF      Seq SS
xa2           1       2.202
xb2           1      22.645
xc2           1      28.398
xa*xb         1       0.091
xa*xc         1       0.051
xb*xc         1       1.293
xa*xb*xc      1       0.068

Obs        xa2          y         Fit   StDev Fit    Residual    St Resid
  1      -1.00     26.500      28.000       0.874      -1.500       -1.21
  2      -1.00     30.500      28.000       0.874       2.500        2.02R
  3      -1.00     27.000      28.000       0.874      -1.000       -0.81
  4       1.00     28.000      27.375       0.757       0.625        0.48
  5       1.00     28.500      27.375       0.757       1.125        0.86
  6       1.00     28.000      27.375       0.757       0.625        0.48
  7       1.00     25.000      27.375       0.757      -2.375       -1.81
  8      -1.00     28.500      29.875       0.757      -1.375       -1.05
  9      -1.00     28.500      29.875       0.757      -1.375       -1.05
 10      -1.00     30.000      29.875       0.757       0.125        0.10
 11      -1.00     32.500      29.875       0.757       2.625        2.00R
 12       1.00     29.500      29.625       0.757      -0.125       -0.10
 13       1.00     32.000      29.625       0.757       2.375        1.81
 14       1.00     29.000      29.625       0.757      -0.625       -0.48
 15       1.00     28.000      29.625       0.757      -1.625       -1.24
 16      -1.00     28.000      26.500       0.757       1.500        1.14
 17      -1.00     25.000      26.500       0.757      -1.500       -1.14
 18      -1.00     26.500      26.500       0.757      -0.000       -0.00
 19      -1.00     26.500      26.500       0.757      -0.000       -0.00
 20       1.00     24.500      25.875       0.757      -1.375       -1.05
 21       1.00     25.000      25.875       0.757      -0.875       -0.67
 22       1.00     28.000      25.875       0.757       2.125        1.62
 23       1.00     26.000      25.875       0.757       0.125        0.10
```

24	-1.00	27.000	27.750	0.757	-0.750	-0.57
25	-1.00	29.000	27.750	0.757	1.250	0.95
26	-1.00	27.500	27.750	0.757	-0.250	-0.19
27	-1.00	27.500	27.750	0.757	-0.250	-0.19
28	1.00	27.500	27.125	0.757	0.375	0.29
29	1.00	28.000	27.125	0.757	0.875	0.67
30	1.00	27.000	27.125	0.757	-0.125	-0.10
31	1.00	26.000	27.125	0.757	-1.125	-0.86

R denotes an observation with a large standardized residual

Printout 7 Multiple Regression Version of a "B and C Main Effects Only" Analysis of Power Requirement (*Example 5*)

Regression Analysis

The regression equation is
y = 27.8 + 0.832 xb2 - 0.949 xc2

Predictor	Coef	StDev	T	P
Constant	27.7619	0.2553	108.73	0.000
xb2	0.8319	0.2553	3.26	0.003
xc2	-0.9494	0.2553	-3.72	0.001

S = 1.420 R-Sq = 47.4% R-Sq(adj) = 43.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	50.972	25.486	12.64	0.000
Residual Error	28	56.463	2.017		
Total	30	107.435			

Source	DF	Seq SS
xb2	1	23.093
xc2	1	27.879

Obs	xb2	y	Fit	StDev Fit	Residual	St Resid
1	-1.00	26.500	27.879	0.457	-1.379	-1.03
2	-1.00	30.500	27.879	0.457	2.621	1.95
3	-1.00	27.000	27.879	0.457	-0.879	-0.65
4	-1.00	28.000	27.879	0.457	0.121	0.09
5	-1.00	28.500	27.879	0.457	0.621	0.46
6	-1.00	28.000	27.879	0.457	0.121	0.09
7	-1.00	25.000	27.879	0.457	-2.879	-2.14R
8	1.00	28.500	29.543	0.437	-1.043	-0.77
9	1.00	28.500	29.543	0.437	-1.043	-0.77
10	1.00	30.000	29.543	0.437	0.457	0.34
11	1.00	32.500	29.543	0.437	2.957	2.19R
12	1.00	29.500	29.543	0.437	-0.043	-0.03
13	1.00	32.000	29.543	0.437	2.457	1.82
14	1.00	29.000	29.543	0.437	-0.543	-0.40
15	1.00	28.000	29.543	0.437	-1.543	-1.14
16	-1.00	28.000	25.981	0.437	2.019	1.49
17	-1.00	25.000	25.981	0.437	-0.981	-0.73
18	-1.00	26.500	25.981	0.437	0.519	0.38
19	-1.00	26.500	25.981	0.437	0.519	0.38
20	-1.00	24.500	25.981	0.437	-1.481	-1.10
21	-1.00	25.000	25.981	0.437	-0.981	-0.73
22	-1.00	28.000	25.981	0.437	2.019	1.49
23	-1.00	26.000	25.981	0.437	0.019	0.01

24	1.00	27.000	27.644	0.437	-0.644	-0.48
25	1.00	29.000	27.644	0.437	1.356	1.00
26	1.00	27.500	27.644	0.437	-0.144	-0.11
27	1.00	27.500	27.644	0.437	-0.144	-0.11
28	1.00	27.500	27.644	0.437	-0.144	-0.11
29	1.00	28.000	27.644	0.437	0.356	0.26
30	1.00	27.000	27.644	0.437	-0.644	-0.48
31	1.00	26.000	27.644	0.437	-1.644	-1.22

R denotes an observation with a large standardized residual

Example 5 has been treated as if the lack of balance in the data came about by misfortune. And the lack of balance in Example 4 *did* come about in such a way. But lack of balance in p-way factorial data can also be the result of careful planning. Consider, for example, a 2^4 factorial situation where the budget can support collection of 20 observations but not as many as 32. In such a case, complete replication of the 16 combinations of two levels of four factors in order to achieve balance is not possible. But it makes far more sense to replicate four of the 16 combinations (and thus be able to calculate s_P and honestly assess the size of background variation) than to achieve balance by using no replication. By now it should be obvious how to subsequently go about the analysis of the resulting partially replicated (and thus unbalanced) factorial data.

Section 3 Exercises

1. Flood and Shankwitz reported the results of a metallurgical engineering design project involving the tempering response of a certain grade of stainless steel. Slugs of this steel were preprocessed to reasonably uniform hardnesses, which were measured and recorded. The slugs were then tempered at various temperatures for various lengths of time. The hardnesses were then remeasured and the change in hardness computed. The data in the accompanying tables were obtained in this replicated 4×4 factorial study.

Time, x_1 (min)	Temperature, x_2 (°F)	Increase in Hardness, y
5	800	0, 0, −1
5	900	−3, −2, 1
5	1000	−1, −1, 0
5	1100	−4, 1, 3
50	800	3, 4, −1
50	900	−3, −1, 1
50	1000	−4, −1, −3
50	1100	−4, −4, −2
150	800	4, 2, −2
150	900	−1, −1, −2
150	1000	−4, −5, −7
150	1100	−7, −5, −8
500	800	1, −3, 0
500	900	−2, −8, −2
500	1000	−8, −7, −7
500	1100	−11, −9, −5

(a) Fit the quadratic model

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 \left(\ln(x_1)\right)^2 + \beta_4 x_2^2 + \beta_5 x_2 \ln(x_1) + \epsilon$$

to these data. What fraction of the observed variability in hardness increase is accounted for in the fitting of the quadratic response surface? What is your estimate of the standard deviation of hardness changes that would be experienced

at any fixed combination of time and temperature? How does this estimate compare with s_P? Does there appear to be enough difference between the two values to cast serious doubt on the appropriateness of the regression model?

(b) There was some concern on the project group's part that the 5-minute time was completely unlike the other times and should not be considered in the same analysis as the longer times. Temporarily delete the 12 slugs treated only 5 minutes from consideration, refit the quadratic model, and compare fitted values for the 36 slugs tempered longer than 5 minutes for this regression to those from part (a). How different are these two sets of values?

Henceforth consider the quadratic model fitted to all 48 data points.

(c) Make a contour plot showing how y varies with $\ln(x_1)$ and x_2. In particular, use it to identify the region of $\ln(x_1)$ and x_2 values where the tempering seems to provide an increase in hardness. Sketch the corresponding region in the (x_1, x_2)-plane.

(d) For the $x_1 = 50$ and $x_2 = 800$ set of conditions,
(i) give a 95% two-sided confidence interval for the mean increase in hardness provided by tempering.
(ii) give a 95% two-sided prediction interval for the increase in hardness produced by tempering an additional slug.
(iii) give an approximate 95% lower tolerance bound for the hardness increases of 90% of such slugs undergoing tempering.

2. Return to the situation of Chapter Exercise 10 of Chapter 8 and the chemical product impurity study. The analysis suggested in that exercise leads to the conclusion that only the A and B main effects are detectably nonzero. The data are unbalanced, so it is not possible to use the reverse Yates algorithm to fit the "A and B main effects only" model to the data.

(a) Use the dummy variable regression techniques to fit the "A and B main effects only" model. (You should be able to pattern what you do after Example 5.) How do A and B main effects estimated on the basis of this few-effects/simplified description of the pattern of response compare with what you obtained for fitted effects using the Yates algorithm?

(b) Compute and plot standardized residuals for the few-effects model. (Plot against levels of A, B, and C, against $\hat{y}$, and normal-plot them.) Do any of these plots indicate any problems with the few-effects model?

(c) How does s_{FE} (which you can read directly off your printout as s_{SF}) compare with s_P in this situation? Do the two values carry any strong suggestion of lack of fit?

Chapter 9 Exercises

1. Return to the situation of Chapter Exercise 3 of Chapter 4 and the grain growth study of Huda and Ralph. Consider an analysis of the researchers' data based on the model

$$y = \beta_0 + \beta_1 x_1 + +\beta_2 \ln(x_2) + \beta_3 x_1 \ln(x_2) + \epsilon$$

(a) Fit this model to the data given in Chapter 4. Based on this fit, what is your estimate of the standard deviation of grain size, y, associated with different specimens treated using a fixed temperature and time?

(b) Make a plot of the observed y's versus the corresponding $\ln(x_2)$'s. On this plot, sketch the linear fitted response functions ($\hat{y}$ versus $\ln(x_2)$) for $x_1 = 1443$, 1493, and 1543. Notice that the fit to the researchers' data is excellent. However, notice also that the model has four β's and was fit based on only nine data points. What possibility therefore needs to be kept in mind when making predictions based on this model?

(c) Make a 95% two-sided confidence interval for the mean y when a temperature of $x_1 = 1493^\circ$K and a time of $x_2 = 120$ minutes are used.

(d) Make a 95% two-sided prediction interval for an additional grain size, y, when a temperature of $x_1 = 1493^\circ$K and a time of $x_2 = 120$ minutes are used.

(e) Find a 95% two-sided confidence interval for the mean y when a temperature of $x_1 = 1500^\circ$K and a time of $x_2 = 500$ minutes are used. (This is not a set of conditions in the original data set. So you will need to inform your regression program of where you wish to predict.)

(f) What does the hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ mean in the context of this study and the model being used in this exercise? Find the p-value associated with an F test of this hypothesis.

(g) What does the hypothesis $H_0: \beta_3 = 0$ mean in the context of this study and the model being used in this exercise? Find the p-value associated with a two-sided t test of this hypothesis.

2. The article "Orthogonal Design for Process Optimization and its Application in Plasma Etching" by Yin and Jillie (*Solid State Technology*, 1987) discusses a 4-factor experiment intended to guide optimization of a nitride etch process on a single wafer plasma etcher. Data were collected at only nine out of $3^4 = 81$ possible combinations of three levels of each of the four factors (making up a so-called *orthogonal array*). The factors involved in the experimentation were the Power applied to the cathode x_1, the Pressure in the reaction chamber x_2, the spacing or Gap between the anode and the cathode x_3, and the Flow of the reactant gas C_2F_6, x_4. Three different responses were measured, an etch rate for SiN y_1, a uniformity for SiN y_2, and a selectivity of the process (for silicon nitride) between silicon nitride and polysilicon y_3. Eight of the nine different combinations were run once, while one combination was run three times. The researchers reported the data given in the accompanying table.

x_1 (W)	x_2 (mTorr)	x_3 (cm)	x_4 (sccm)	y_1 (Å/min)	y_2 (%)	y_3 (SiN/poly)
275	450	0.8	125	1075	2.7	1.63
275	500	1.0	160	633	4.9	1.37
275	550	1.2	200	406	4.6	1.10
300	450	1.0	200	860	3.4	1.58
300	500	1.2	125	561	4.6	1.26
275	450	0.8	125	1052	1.7	1.72
300	550	0.8	160	868	4.6	1.65
325	450	1.2	160	669	5.0	1.42
325	500	0.8	200	1138	2.9	1.69
325	550	1.0	125	749	5.6	1.54
275	450	0.8	125	1037	2.6	1.72

The data are listed in the order in which they were actually collected. Notice that the conditions under which the first, sixth, and eleventh data points were collected are the same—that is, there is some replication in this fractional factorial data set.

(a) The fact that the first, sixth, and last data points were collected under the same set of process conditions provides some check on the consistency of experimental results across time in this study. What else might (should) have been done in this study to try to make sure that time trends in an extraneous variable don't get confused with the effects of the experimental variables (in particular, the effect of x_1, as the experiment was run)? (Consider again the ideas of Section 2.3.)

(b) Fit a linear model in all of x_1, x_2, x_3, and x_4 to each of the three response variables. Notice that although such a model appears to provide a good fit to the y_3 data, the situations for y_1 and y_2 are not quite so appealing. (Compare s_{SF} to s_P for y_1 and note that R^2 for the second variable is relatively low, at least compared to what one can achieve for y_3.)

(c) In search of better-fitting equations for the y_2 (or y_1) data, one might consider fitting a full quadratic equation in x_1, x_2, x_3, and x_4 to the data. What happens when you attempt to do this using a regression package? (The problem is that the data given here are not adequate to

distinguish between various possible quadratic response surfaces in four variables.)

(d) In light of the difficulty experienced in (c), a natural thing to do might be to try to fit quadratic surfaces involving only some of all possible second-order terms. Fit the two models for y_2 including (i) x_1, x_2, x_3, x_4, x_1^2, x_2^2, x_3^2, and x_4^2 terms, and (ii) x_1, x_2, x_4, x_1^2, x_2^2, x_4^2, x_1x_2, and x_2x_4 terms. How do these two fitted equations compare in terms of $\hat{y}_2$ values for (x_1, x_2, x_3, x_4) combinations in the data set? How do $\hat{y}_2$ values compare for the two fitted equations when $x_1 = 325, x_2 = 550, x_3 = 1.2$, and $x_4 = 200$? (Notice that although this last combination is not in the data set, there are values of the individual variables in the data set matching these.) What is the practical engineering difficulty faced in a situation like this, where there is not enough data available to fit a full quadratic model but it doesn't seem that a model linear in the variables is an adequate description of the response?

Henceforth, confine attention to y_3 and consider an analysis based on a model linear in all of x_1, x_2, x_3, and x_4.

(e) Give a 90% two-sided individual confidence interval for the increase in mean selectivity ratio that accompanies a 1 watt increase in power.

(f) What appear to be the optimal (large y_3) settings of the variables x_1, x_2, x_3, and x_4 (within their respective ranges of experimentation)? Refer to the coefficients of your fitted equation from (b).

(g) Give a 90% two-sided confidence interval for the mean selectivity ratio at the combination of settings that you identified in (f). What cautions would you include in a report in which this interval is to appear? (Under what conditions is your calculated interval going to have real-world meaning?)

3. The article "How to Optimize and Control the Wire Bonding Process: Part II" by Scheaffer and Levine (*Solid State Technology*, 1991) discusses the use of a $k = 4$ factor central composite design in the improvement of the operation of the K&S 1484XQ bonder. The effects of the variables Force, Ultrasonic Power, Temperature, and Time on the final ball bond shear strength were studied. The accompanying table gives data like those collected by the authors. (The original data were not given in the paper, but enough information was given to produce these simulated values that have structure like the original data.)

Force, x_1 (gm)	Power, x_2 (mw)	Temp., x_3 °C	Time, x_4 (ms)	Strength, y (gm)
30	60	175	15	26.2
40	60	175	15	26.3
30	90	175	15	39.8
40	90	175	15	39.7
30	60	225	15	38.6
40	60	225	15	35.5
30	90	225	15	48.8
40	90	225	15	37.8
30	60	175	25	26.6
40	60	175	25	23.4
30	90	175	25	38.6
40	90	175	25	52.1
30	60	225	25	39.5
40	60	225	25	32.3
30	90	225	25	43.0
40	90	225	25	56.0
25	75	200	20	35.2
45	75	200	20	46.9
35	45	200	20	22.7
35	105	200	20	58.7
35	75	150	20	34.5
35	75	250	20	44.0
35	75	200	10	35.7
35	75	200	30	41.8
35	75	200	20	36.5
35	75	200	20	37.6
35	75	200	20	40.3
35	75	200	20	46.0
35	75	200	20	27.8
35	75	200	20	40.3

(a) Fit both the full quadratic response surface and the simpler linear response surface to these data. On the basis of simple examination of the R^2 values, does it appear that the quadratic surface is enough better as a data summary to make it worthwhile to suffer the increased complexity that it brings with it? How do the s_{SF} values for the two fitted models compare to s_P computed from the final six data points listed here?

(b) Conduct a formal test (in the full quadratic model) of the hypothesis that the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ is an adequate description of the response. Does your p-value support your qualitative judgment from part (a)?

(c) In the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$, give a 90% confidence interval for β_2. Interpret this interval in the context of the original engineering problem. (What is β_2 supposed to measure?) Would you expect the p-value from a test of $H_0: \beta_2 = 0$ to be large or to be small?

(d) Use the linear model and find an approximate 95% lower tolerance bound for 98% of bond shear strengths at the center point $x_1 = 35$, $x_2 = 75$, $x_3 = 200$, and $x_4 = 20$.

4. (Testing for "Lack of Fit" to a Regression Model) In curve- and surface-fitting problems where there is some replication, this text has used the informal comparison of s_{SF} (or s_{LF}) to s_P as a means of detecting poor fit of a regression model. It is actually possible to use these values to conduct a formal significance test for lack of fit. That is, under the one-way normal model of Chapter 7, it is possible to test

$$H_0: \mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

using the test statistic

$$F = \frac{\dfrac{(n-k-1)s_{SF}^2 - (n-r)s_P^2}{r-k-1}}{s_P^2}$$

and an $F_{r-k-1,n-r}$ reference distribution, where large values of F count as evidence against H_0. (If s_{SF} is much larger than s_P, the difference in the numerator of F will be large, producing a large sample value and a small observed level of significance.)

(a) It is not possible to use the lack of fit test in any of Exercise 3 of Section 4.1, Exercise 2 of Section 4.2, or Chapter Exercises 2 or 3 of Chapter 4. Why?

(b) For the situation of Exercise 2 of Section 9.1, conduct a formal test of lack of fit of the linear relationship $\mu_{y|x} = \beta_0 + \beta_1 x$ to the concrete strength data.

(c) For the situation of Exercise 1 of Section 9.3, conduct a formal test of lack of fit of the full quadratic relationship

$$\mu_{y|x_1,x_2} = \beta_0 + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 \left(\ln(x_1)\right)^2 + \beta_4 x_2^2 + \beta_5 x_2 \ln(x_1)$$

to the hardness increase data.

(d) For the situation of Chapter Exercise 3, conduct a formal test of lack of fit of the linear relationship

$$\mu_{y|x_1,x_2,x_3,x_4} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

to the ball bond shear strength data.

5. Return to the situation of Chapter Exercises 18 and 19 of Chapter 4 and the ore refining study of S. Osoka. In that study, the object was to discover settings of the process variables x_1 and x_2 that would simultaneously maximize y_1 and minimize y_2.

(a) Fit full quadratic response functions for y_1 and y_2 to the data given in Chapter 4. Compute and plot standardized residuals for these two fitted equations. Comment on the appearance of these plots and what they indicate about the appropriateness of the fitted response surfaces.

(b) One useful rule of thumb in response surface studies (suggested by Box, Hunter, and Hunter in their book *Statistics for Experimenters*) is to

check that for a fitted surface involving a total of l coefficients b (including b_0),

$$\max \hat{y} - \min \hat{y} > 4\sqrt{\frac{l \cdot s_{SF}^2}{n}}$$

before trying to make decisions based on its nature (bowl-shape up or down, saddle, etc.) or do even limited interpolation or extrapolation. This criterion is a comparison of the movement of the fitted surface across those n data points in hand, to four times an estimate of the root of the average variance associated with the n fitted values $\hat{y}$. If the criterion is not satisfied, the interpretation is that the fitted surface is so flat (relative to the precision with which it is determined) as to make it impossible to tell with any certainty the true nature of how mean response varies as a function of the system variables.

Judge the usefulness of the surfaces fitted in part (a) against this criterion. Do the response surfaces appear to be determined adequately to support further analysis (involving optimization, for example)?

(c) Use the analytic method discussed in Section 9.3 to investigate the nature of the response surfaces fitted in part (a). According to the signs of the eigenvalues, what kinds of surfaces were fitted to y_1 and y_2, respectively?

(d) Make contour plots of the fitted y_1 and y_2 response surfaces from (a) on a single set of (x_1, x_2)-axes. Use these to help locate (at least approximately) a point (x_1, x_2) with maximum predicted y_1, subject to a constraint that predicted y_2 be no larger than 55.

(e) For the point identified in part (d), give 90% two-sided prediction intervals for the next values of y_1 and y_2 that would be produced by this refining process. Also give an approximate 95% lower tolerance bound for 90% of additional pyrite recoveries and an approximate 95% upper tolerance bound for 90% of additional kaolin recoveries at this combination of x_1 and x_2 settings.

6. Return to the concrete strength testing situation of Chapter Exercise 16 of Chapter 4.

(a) Find estimates of the parameters β_0, β_1, and σ in the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$.

(b) Compute standardized residuals and plot them in the same ways that you were asked to plot the ordinary residuals in part (g) of the problem in Chapter 4. How much do the appearances of the new plots differ from the earlier ones?

(c) Make a 95% two-sided confidence interval for the increase in mean compressive strength that accompanies a 5 psi increase in splitting tensile strength. (Note: This is $5\beta_1$.)

(d) Make a 90% two-sided confidence interval for the mean strength of specimens with splitting tensile strength 300 psi (based on the simple linear regression model).

(e) Make a 90% two-sided prediction interval for the strength of an additional specimen with splitting tensile strength 300 psi (based on the simple linear regression model).

(f) Find an approximate 95% lower tolerance bound for the strengths of 90% of additional specimens with splitting tensile strength 300 psi (based on the simple linear regression model).

7. Wiltse, Blandin, and Schiesel experimented with a grain thresher built for an agricultural engineering design project. They ran efficiency tests on the cleaning chamber of the machine. This part of the machine sucks air through threshed material, drawing light (nonseed) material out an exhaust port, while the heavier seeds fall into a collection tray. Airflow is governed by the spacing of an air relief door. The following are the weights, y (in grams), of the portions of 14 gram samples of pure oat seeds run through the cleaning chamber that ended up in the collection tray. Four different door spacings x were used, and 20 trials were made at each door spacing.

.500 in. Spacing
12.00, 12.30, 12.45, 12.45, 12.50, 12.50, 12.50, 12.60, 12.65, 12.70, 12.70, 12.80, 12.90, 12.90, 13.00, 13.00, 13.00, 13.10, 13.20, 13.20

.875 in. Spacing
12.40, 12.80, 12.80, 12.90, 12.90, 12.90, 12.90, 13.00, 13.00, 13.00, 13.00, 13.20, 13.20, 13.20, 13.30, 13.40, 13.40, 13.45, 13.45, 13.70

1.000 in. Spacing
12.00, 12.80, 12.80, 12.90, 12.90, 13.00, 13.00, 13.00, 13.15, 13.20, 13.20, 13.30, 13.40, 13.40, 13.45, 13.50, 13.60, 13.60, 13.60, 13.70

1.250 in. Spacing
12.10, 12.20, 12.25, 12.25, 12.30, 12.30, 12.30, 12.40, 12.50, 12.50, 12.50, 12.60, 12.60, 12.85, 12.90, 12.90, 13.00, 13.10, 13.15, 13.25

Use the quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ and do the following.

(a) Find an estimate of σ in the model above. What is this supposed to measure? How does your estimate compare to s_P here? What does this comparison suggest to you?

(b) Use an F statistic and test the null hypothesis $H_0: \beta_1 = \beta_2 = 0$. (You may take values off a printout to do this but show the whole five-step significance-testing format.) What is the meaning of this hypothesis in the present context?

(c) Use a t statistic and test the null hypothesis $H_0: \beta_2 = 0$. (Again, you may take values off a printout to do this but show the whole five-step significance-testing format.) What is the meaning of this hypothesis in the present context?

(d) Give a 90% lower confidence bound for the mean weight of the part of such samples that would wind up in the collection tray using a 1.000 in. door spacing.

(e) Give a 90% lower prediction bound for the next weight of the part of such a sample that would wind up in the collection tray using a 1.000 in. door spacing.

(f) Give an approximate 95% lower tolerance for 90% of the weights of all such samples that would wind up in the collection tray using a 1.000 in. door spacing.

8. Return to the armor testing context of Chapter Exercise 21 of Chapter 4. In what follows, base your answers on the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

(a) Based on this model, what is your estimate of the standard deviation of ballistic limit, y, associated with different specimens of a given thickness and Brinell hardness?

(b) Find and plot the standardized residuals. (Plot them versus x_1, versus x_2, and versus $\hat{y}$ and normal-plot them.) Comment on the appearance of your plots.

(c) Make 90% two-sided confidence intervals for β_1 and for β_2. Based on the second of these, what increase in mean ballistic limit would you expect to accompany a 20-unit increase in the Brinell hardness number?

(d) Make a 95% two-sided confidence interval for the mean ballistic limit when a thickness of $x_1 = 258$ (.001 in.) and a Brinell hardness of $x_2 = 391$ are involved.

(e) Make a 95% two-sided prediction interval for an additional ballistic limit when a thickness of $x_1 = 258$ (.001 in.) and a Brinell hardness of $x_2 = 391$ are involved.

(f) Find an approximate 95% lower tolerance bound for 98% of additional ballistic limits when a thickness of $x_1 = 258$ (.001 in.) and a Brinell hardness of $x_2 = 391$ are involved.

(g) Find a 95% two-sided confidence interval for the mean ballistic limit when a thickness of $x_1 = 260$ (.001 in.) and a Brinell hardness of $x_2 = 380$ are involved.

(h) What does the hypothesis $H_0: \beta_1 = \beta_2 = 0$ mean in the context of this study and the model being used in this exercise? Find the p-value associated with an F test of this hypothesis.

(i) What does the hypothesis $H_0: \beta_1 = 0$ mean in the context of this study and the model being used in this exercise? Find the p-value associated with a two-sided t test of this hypothesis.

9. Return to the PETN density/detonation velocity data of Chapter Exercise 23 of Chapter 4.

(a) Find estimates of the parameters β_0, β_1, and σ in the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$. How does your estimate of σ compare to s_P? What does this comparison suggest about the reasonableness of the regression model for the data in hand?

(b) Compute standardized residuals and plot them in the same ways that you plotted the residuals in part (g) of Chapter Exercise 23 of Chapter 4. How much do the appearances of the new plots differ from the earlier ones?

(c) Make a 90% two-sided confidence interval for the increase in mean detonation velocity that accompanies a 1 g/cc increase in PETN density.

(d) Make a 90% two-sided confidence interval for the mean detonation velocity of charges with PETN density 0.65 g/cc.

(e) Make a 90% two-sided prediction interval for the next detonation velocity of a charge with PETN density 0.65 g/cc.

(f) Make an approximate 99% lower tolerance bound for the detonation velocities of 95% of charges having a PETN density of 0.65 g/cc.

10. Return to the thread stripping problem of Chapter Exercise 24 of Chapter 4.

(a) Find estimates of the parameters β_0, β_1, β_2, and σ in the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. How does your estimate of σ compare to s_P? What does this comparison suggest about the reasonableness of the quadratic model for the data in hand? What is your estimate of σ supposed to be measuring?

(b) Use an F statistic and test the null hypothesis $H_0: \beta_1 = \beta_2 = 0$ for the quadratic model. (You may take values off a printout to help you do this but show the whole five-step significance testing format.) What is the meaning of this hypothesis in the present context?

(c) Use a t statistic and test the hypothesis $H_0: \beta_2 = 0$ in the quadratic model. (Again, show the whole five-step significance testing format.) What is the meaning of this hypothesis in the present context?

(d) Give a 95% two-sided confidence interval for the mean torque at failure for a thread engagement of 40 (in the units of the problem) using the quadratic model.

(e) Give a 95% two-sided prediction interval for an additional torque at failure for a thread engagement of 40 using the quadratic model.

(f) Give an approximate 99% lower tolerance bound for 95% of torques at failure for studs having thread engagements of 40 using the quadratic model.

11. Return to the situation of Chapter Exercise 28 of Chapter 4 and the metal cutting experiment of Mielnick. Consider an analysis of the torque data based on the model $y_1' = \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \epsilon$.

(a) Make a 90% two-sided confidence interval for the coefficient β_1.

(b) Make a 90% two-sided confidence interval for the mean log torque when a .318 in drill and a feed rate of .005 in./rev are used.

(c) Make a 95% two-sided prediction interval for an additional log torque when a .318 in drill and a feed rate of .005 in./rev are used. Exponentiate the endpoints of this interval to get a prediction interval for a raw torque under these conditions.

(d) Find a 95% two-sided confidence interval for the mean log torque for $x_1 = .300$ in and $x_2 = .010$ in./rev.

12. Return to Chapter Exercise 25 of Chapter 4 and the tire grip force study.

(a) Find estimates of the parameters β_0, β_1, and σ in the simple linear regression model $\ln(y) = \beta_0 + \beta_1 x + \epsilon$.

(b) Compute standardized residuals and plot them in the same ways you plotted the residuals in part (h) of Chapter Exercise 25 of

Chapter 4. How much do the appearances of the new plots differ from the earlier ones?

(c) Make a 90% two-sided confidence interval for the increase in mean log grip force that accompanies an increase in drag of 10% (e.g., from 30% drag to 40% drag). Note that this is $10\beta_1$.

(d) Make a 95% two-sided confidence interval for the mean log grip force of a tire of this type under 30% drag (based on the simple linear regression model).

(e) Make a 95% two-sided prediction interval for the raw grip force of another tire of this design under 30% drag. (*Hint:* Begin by making an interval for log grip force of such a tire.)

(f) Find an approximate 95% lower tolerance bound for the grip forces of 90% of tires of this design under 30% drag (based on the simple linear regression model for $\ln(y)$).

13. Consider again the asphalt permeability data of Woelfl, Wei, Faulstich, and Litwack given in Chapter Exercise 26 of Chapter 4. Use the quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ and do the following:

(a) Find an estimate of σ in the quadratic model. What is this supposed to measure? How does your estimate compare to s_{P} here? What does this comparison suggest to you?

(b) Use an F statistic and test the null hypothesis $\text{H}_0\colon \beta_1 = \beta_2 = 0$ for the quadratic model. (You may take values off a printout to help you do this, but show the whole five-step significance testing format.) What is the meaning of this hypothesis in the present context?

(c) Use a t statistic and test the null hypothesis $\text{H}_0\colon \beta_2 = 0$ in the quadratic model. Again, show the whole five-step significance testing format. What is the meaning of this hypothesis in the present context?

(d) Give a 90% two-sided confidence interval for the mean permeability of specimens of this type with a 6.5% asphalt content.

(e) Give a 90% two-sided prediction interval for the next permeability measured on a specimen of this type having a 6.5% asphalt content.

(f) Find an approximate 95% lower tolerance bound for the permeability of 90% of the specimens of this type having a 6.5% asphalt content.

14. Consider again the axial breaking strength data of Koh, Morden, and Ogbourne given in Chapter Exercise 27 of Chapter 4. At one point in that exercise, it is argued that perhaps the variable $x_3 = x_1^2/x_2$ is the principal determiner of axial breaking strength, y.

(a) Plot the 36 pairs (x_3, y) corresponding to the data given in Chapter 4. Note that a constant σ assumption is probably not a good one over the whole range of x_3's in the students' data.

In light of the point raised in part (a), for purposes of simple linear regression analysis, henceforth restrict attention to those 27 data pairs with $x_3 > .004$.

(b) Find estimates of the parameters β_0, β_1, and σ in the simple linear regression model $y = \beta_0 + \beta_1 x_3 + \epsilon$. How does your estimate of σ based on the simple linear regression model compare to s_{P}? What does this comparison suggest about the reasonableness of the regression model for the data in hand?

(c) Make a 98% two-sided confidence interval for the mean axial breaking strength of .250 in. dowels 8 in. in length based on the regression analysis. How does this interval compare with the use of formula (6.20) and the four measurements on dowels of this type contained in the data set?

(d) Make a 98% two-sided prediction interval for the axial breaking strength of a single additional .250 in. dowel 8 in. in length. Do the same if the dowel is only 6 in. in length.

(e) Make an approximate 95% lower tolerance bound for the breaking strengths of 98% of .250 in. dowels 8 in. in length.