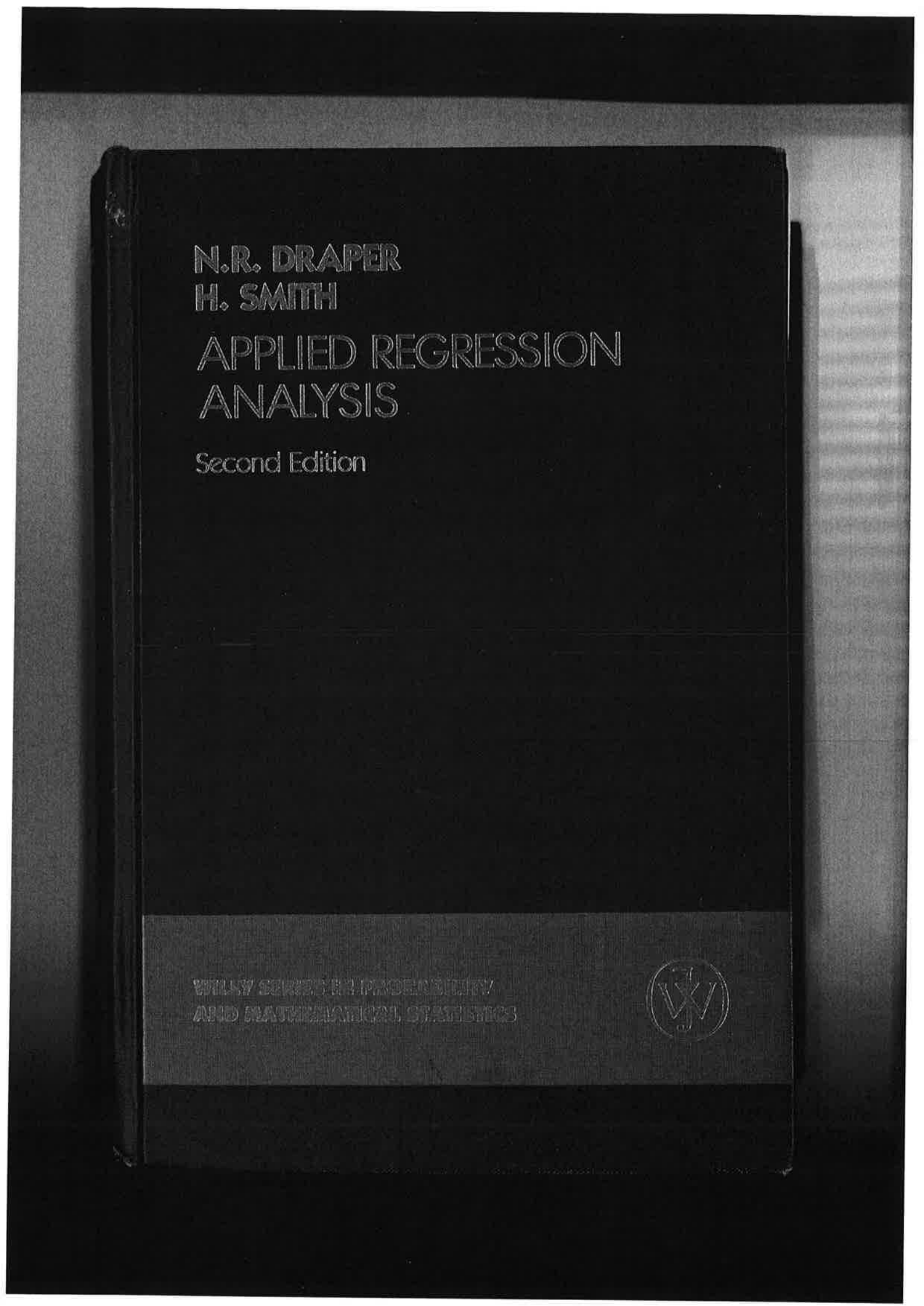
N.R. DRAPER
H. SMITH
APPLIED REGRESSION ANALYSIS
Second Edition

The second table is often rewritten with the *corrected* total sum of squares at the bottom, omitting the sum of squares due to the mean $n\bar{Y}^2$. (Incidentally, as we noted previously, we can write $n\bar{Y}^2$ in matrix form as $\mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n$ if we wish, although this is not usually done. This computation is, in fact, less subject to roundoff error if performed as $(\sum Y_i)^2/n$.) The abbreviated table takes the following form

Source		df	SS	MS
	$SS(b_1 \mid b_0)$	1	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	
Residual	Lack of fit	$n-2-n_e$	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - SS(p.e.)$	MS_L
	Pure error	n_e	SS(p.e)	s_e^2
Total, corrected		$n-1$	$\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$	

The tests for lack of fit and for β_1 are performed as described in Chapter 1. An additional measure of the regression is provided by the ratio

$$R^2 = \frac{(\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2)}{(\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2)}$$

4. If no lack of fit is shown, $(\mathbf{X}'\mathbf{X})^{-1}s^2$ will provide estimates of $V(b_0)$, $V(b_1)$, and $\text{cov}(b_0, b_1)$ and enable individual coefficients to be tested or other calculations made as in Chapter 1.

5. The following quantities can be found:

The vector of fitted values:	$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$
A prediction of Y at X_0:	$\hat{Y}_0 = \mathbf{X}_0'\mathbf{b} = \mathbf{b}'\mathbf{X}_0$
with variance:	$V(\hat{Y}_0) = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2$

2.6. The General Regression Situation

We have seen how the problem of fitting a straight line by least squares can be handled through the use of matrices. This approach is important for the following reason. If we wish to fit *any* model linear in parameters β_0, β_1, β_2, ..., by least squares, the calculations necessary are of exactly the same form (in matrix terms) as those for the straight line involving only two parameters β_0 and β_1. The mechanics of calculation, however, increase sharply with the number of parameters. Thus while the formulae are easy

to remember, the use of a digital computer is an essential for nearly all problems except when

1. the number of parameters is small—say less than five.
2. the data arise from a designed experiment which provides an $\mathbf{X}'\mathbf{X}$ matrix of simple, or "patterned," form.

A general statement of linear regression methods will now be given. For the theoretical derivation of these results, the reader should consult, for example, *Regression Analysis* by R. L. Plackett, Clarendon Press, Oxford, 1960.

Suppose we have a model under consideration which can be written in the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.6.1}$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector of observations,
$\mathbf{X}$ is an $(n \times p)$ matrix of known form,
$\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters,
$\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of errors,

and where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$, so the elements of $\boldsymbol{\varepsilon}$ are uncorrelated.

Since $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, an alternative way of writing the model is

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}. \tag{2.6.1}$$

The error sum of squares is then

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{2.6.2}$$

(This follows due to the fact that $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}$ is a 1×1 matrix, or a scalar, whose transpose $(\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\boldsymbol{\beta}$ must have the same value.)

The least squares estimate of $\boldsymbol{\beta}$ is the value $\mathbf{b}$ which, when substituted in Eq. (2.6.2), minimizes $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. It can be determined by differentiating Eq. (2.6.2) with respect to $\boldsymbol{\beta}$ and setting the resultant matrix equation equal to zero, at the same time replacing $\boldsymbol{\beta}$ by $\mathbf{b}$. (Differentiating $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ with respect to a vector quantity $\boldsymbol{\beta}$ is equivalent to differentiating $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ separately with respect to each element of $\boldsymbol{\beta}$ in order, writing down the resulting derivatives one below the other, and rearranging the whole into matrix form.) This provides the *normal equations*

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{2.6.3}$$

Two main cases arise; either Eq. (2.6.3) consists of p independent equations in p unknowns, or some equations depend on others so that there are fewer than p independent equations in the p unknowns (the p unknowns are

the elements of **b**). If some of the normal equations depend on others, $\mathbf{X}'\mathbf{X}$ is singular, so that $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. Then either the model should be expressed in terms of fewer parameters or else additional restrictions on the parameters must be given or assumed. Some examples of this situation are given in Chapter 9. If the p normal equations are independent, $\mathbf{X}'\mathbf{X}$ is nonsingular, and its inverse exists. In this case the solution of the normal equations can be written

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2.6.4}$$

This solution **b** has the following properties:

1. It is an estimate of $\boldsymbol{\beta}$ which minimizes the error sum of squares $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, *irrespective* of any distribution properties of the errors.

Note. An assumption that the errors $\boldsymbol{\varepsilon}$ are normally distributed is *not* required in order to obtain the estimates **b** but it *is* required later in order to make tests which depend on the assumption of normality, such as t- or F-tests, or for obtaining confidence intervals based on the t- and F-distributions.

2. The elements of **b** are linear functions of the observations $Y_1, Y_2, \ldots, Y_n$, and provide unbiased estimates of the elements of $\boldsymbol{\beta}$ which have the minimum variances (of *any* linear functions of the Y's which provide unbiased estimates), irrespective of distribution properties of the errors.

Note. Suppose we have an expression $T = l_1 Y_1 + l_2 Y_2 + \cdots + l_n Y_n$, which is a linear function of observations $Y_1, Y_2, \ldots, Y_n$, and which we use as an estimate of a parameter θ. Then T is a random variable whose probability distribution will depend on the distribution from which the Y's arise. If we repeatedly take samples of Y's and evaluate the corresponding T's we shall generate the distribution of T empirically. Whether we do this or not, the distribution of T will have some definite mean value which we can write $E(T)$ and a variance which we can write $V(T)$. If it happens that the mean of the distribution of T is equal to the parameter θ we are estimating by T—that is, if $E(T) = \theta$—then we say that T is an unbiased estimator of θ. The word *estimator* is normally used when referring to the theoretical expression for T in terms of a sample of Y's. A specific numerical value of T would be called an unbiased *estimate* of θ. This distinction, though correct, is not always maintained in statistical writings. If we have all possible linear functions $T_1, T_2, \ldots$, say, of n observations $Y_1, Y_2, \ldots, Y_n$, and if the T's satisfy

$$\theta = E(T_1) = E(T_2) \cdots$$

that is, they are all unbiased estimators of θ, then the one with the smallest value of $V(T_j), j = 1, 2, \ldots$, is the *minimum variance unbiased estimator* of θ. (The result (2) is "Gauss's Theorem.")

3. If the errors are independent and $\varepsilon_i \sim N(0, \sigma^2)$, then $\mathbf{b}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. (In vector terms we can write $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, meaning that $\boldsymbol{\varepsilon}$ follows an n-dimensional multivariate normal distribution with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ (where $\mathbf{0}$ denotes a vector consisting entirely of zeros and of the same length as $\boldsymbol{\varepsilon}$) and $\mathbf{V}(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$; that is, $\boldsymbol{\varepsilon}$ has a variance-covariance matrix whose diagonal elements, $V(\varepsilon_i)$, $i = 1, 2, \ldots, n$ are all σ^2 and whose off-diagonal elements, covariance $(\varepsilon_i, \varepsilon_j)$, $i \neq j = 1, \ldots, n$ are all zero. The likelihood function for the sample $Y_1, Y_2, \ldots, Y_n$ is defined in this case as the product

$$\prod_{i=1}^{n} \frac{1}{\sigma(2\pi)^{1/2}} e^{-\varepsilon_i^2/(2\sigma^2)} = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/2\sigma^2}. \tag{2.6.5}$$

Thus for a fixed value of σ, maximizing the likelihood function is equivalent to minimizing the quantity $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. Note that this fact can be used to provide a justification for the least squares procedure (i.e., for minimizing the sum of *squares* of errors), because in many physical situations the assumption that errors are normally distributed is quite sensible. We shall, in any case, find out if this assumption appears to be violated by examining the residuals from the regression analysis. If any definite *a priori* knowledge *is* available about the error distribution, perhaps from theoretical considerations or from sound prior knowledge of the process under study, the maximum likelihood argument could be used to obtain estimates based on a criterion other than least squares. For example, suppose the errors ε_i, $i = 1, 2, \ldots, n$ were independent and followed the double exponential distribution:

$$f(\varepsilon_i) = (2\sigma)^{-1} e^{-|\varepsilon_i|/\sigma} \qquad (-\infty \leq \varepsilon_i \leq \infty) \tag{2.6.6}$$

rather than the normal distribution:

$$f(\varepsilon_i) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-\varepsilon_i^2/2\sigma^2} \tag{2.6.7}$$

as is usually assumed. The double exponential frequency function has a pointed peak of height $1/2\sigma$ at $\varepsilon_i = 0$, and tails off to zero as ε_i goes to both plus and minus infinity. Then application of the maximum likelihood principle for estimating $\boldsymbol{\beta}$, assuming σ fixed, would involve minimization of

$$\sum_{i=1}^{n} |\varepsilon_i|$$

the sum of absolute errors and not the minimization of

$$\sum_{i=1}^{n} \varepsilon_i^2$$

the sum of *squares* of errors. For further reading on minimizing the sum of absolute errors, see the useful reference "Least absolute values estimation: an introduction," by J. E. Gentle, *Communications Statistics—Simulated Computations*, **B6**(4), 1977, 313–328. For computational aspects, see (Fortran) Algorithm AS 110, "L_p norm fit of a straight line," by V. A. Sposito, W. J. Kennedy, and J. E. Gentle, *Applied Statistics*, **26**, 1977, 114–118, and (Fortran) Algorithm AS 108, "Multiple linear regression with minimum sum of absolute errors," by S. C. Narula and J. F. Wellington, *Applied Statistics*, **26**, 1977, 106–111.

Without Distributional Assumptions

Suppose we have used the method of least squares to estimate $\boldsymbol{\beta}$ by $\mathbf{b}$. We can proceed with the following steps whether the errors are normally distributed or not.

1. The fitted values are obtained from $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$.

2. The vector of residuals is given by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ (see Chapter 3 for their examination). It is true that $\sum_{i=1}^{n} e_i \hat{Y}_i = 0$, whatever the model. This can be seen by multiplying the jth normal equation by the jth b and adding the results. If there is a β_0 term in the model, it is also true that $\sum_{i=1}^{n} e_i = 0$. (The e_i and $\hat{Y}_i$, $i = 1, 2, \ldots, n$ are the ith elements of the vectors $\mathbf{e}$ and $\hat{\mathbf{Y}}$, respectively.)

3. $\mathbf{V(b)} = (\mathbf{X'X})^{-1}\sigma^2$ provides the variances (diagonal terms) and covariances (off-diagonal terms) of the estimates. (An estimate of σ^2 is obtained as described below.)

4. Suppose $\mathbf{X}_0'$ is a specified $1 \times p$ vector whose elements are of the same form as a row of $\mathbf{X}$ so that $\hat{Y}_0 = \mathbf{X}_0'\mathbf{b} = \mathbf{b}'\mathbf{X}_0$ is *the fitted value at a specified point* $\mathbf{X}_0$. For example, if the model were $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$, then $\mathbf{X}_0' = (1, X_0, X_0^2)$ for a given value X_0. Then $\hat{Y}_0$ is the value *predicted at* $\mathbf{X}_0$ *by the regression equation* and has variance

$$V(\hat{Y}_0) = \mathbf{X}_0'\mathbf{V(b)}\mathbf{X}_0 = \mathbf{X}_0'(\mathbf{X'X})^{-1}\mathbf{X}_0\sigma^2. \tag{2.6.8}$$

5. A basic analysis of variance table can be constructed as follows.

Source	df	SS	MS
Regression	p	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	MS_R
Residual	$n - p$	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	MS_E
Total	n	$\mathbf{Y}'\mathbf{Y}$	

A further subdivision of the parts of this table can be carried out as follows.

5a. If a β_0 term is in the model we can subdivide the regression sum of squares into

$$\mathrm{SS}(b_0) = \frac{(\sum Y_i)^2}{n} = n\bar{Y}^2 \tag{2.6.9}$$

$$\mathrm{SS}(\text{Regression}\,|\,b_0) = \mathrm{SS}(R\,|\,b_0) = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{(\sum Y_i)^2}{n}. \tag{2.6.10}$$

These sums of squares are based on 1 and $p-1$ degrees of freedom, respectively. (More extensive subdivision of the regression sum of squares will be discussed in Section 2.7.)

5b. If repeat observations are available we can split the residual SS into SS(pure error) with n_e degrees of freedom, which estimates $n_e\sigma^2$ and SS(lack of fit) with $(n-p-n_e)$ degrees of freedom.

"Repeats" now must be repeats in *all* coordinates $X_1, X_2, \ldots, X_k$ of the independent variables (though approximate use of "very close" points is sometimes seen in practice). This provides an analysis of variance table as follows.

Source	df	SS	MS	
b_0	1	$\mathrm{SS}(b_0)$		$\}\mathrm{MS}_R$
Regression $\|\, b_0$	$p-1$	$\mathrm{SS}(R\,\|\,b_0)$	$\mathrm{MS}(R\,\|\,b_0)$	$\}$
Lack of fit	$n-p-n_e$	SS(l.o.f.)	MS(l.o.f.)	$\}\mathrm{MS}_E$
Pure error	n_e	SS(p.e.)	MS(p.e)	$\}$
Total	n	$\mathbf{Y}'\mathbf{Y}$		

Note. The order in which terms are given in the table is not important. Most of the tables in this text are rearranged in the order often seen in computer printouts.

THE R^2 STATISTIC. The ratio

$$R^2 = \frac{\mathrm{SS}(R\,|\,b_0)}{\mathbf{Y}'\mathbf{Y} - \mathrm{SS}(b_0)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \tag{2.6.11}$$

is an extension of the quantity defined for the straight line regression and is the square of the *multiple correlation coefficient*. Another name for R^2 is the *coefficient of multiple determination*. (The quantity R^2 must not be confused with the R in the expressions $\mathrm{SS}(R\,|\,b_0)$ and MS_R, where R is a label denoting the regression contribution.)

R^2 is the square of the correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ and $0 \leq R^2 \leq 1$. If pure error exists, it is impossible for R^2 to actually attain 1; see the remarks on pp. 33, 40–41, 64, and 547. A perfect fit to the data for which $\hat{Y}_i = Y_i$, an unlikely event in practice, would give $R^2 = 1$.

If $Y_i = \overline{Y}$, that is, $b_1 = b_2 = \cdots = b_{p-1} = 0$ (or a model $Y = \beta_0 + \varepsilon$ alone has been fitted), then $R^2 = 0$. Thus R^2 is a measure of the usefulness of the terms, other than β_0, in the model. It is important to realize that, if there is no pure error, R^2 can be made unity simply by employing n properly selected coefficients in the model, including β_0, since a model can then be chosen which fits the data exactly. (For example, if we have an observation of Y at four different values of X, a cubic polynomial

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

passes exactly through all four points.) Since R^2 is often used as a convenient measure of the success of the regression equation in explaining the variation in the data, we must be sure that an improvement in R^2 due to adding a new term to the model has some real significance and is not due to the fact that the number of parameters in the model is getting close to saturation point—that is, the number of distinct X-sites. This is an *especial* danger when there are *repeat* observations. For example, if we have one hundred observations which occur in five groups each of twenty repeats, we have effectively five pieces of information, represented by five mean values, and ninety-five error degrees of freedom for pure error, nineteen at each repeat point. Thus a five-parameter model will provide a perfect fit to the five means and may give a very large value of R^2, especially if the experimental error is small compared with the spread of the five means. In this case the fact that one hundred observations can be well predicted by a model with only five parameters is not surprising since there are really only five distinct data points and not one hundred as it first seemed. When there are no exact repeats, but the points in the X-space (at which observations Y are available) *are* close together, this type of situation can occur and yet be well concealed within the data. Plots of the data, and the residuals (see Chapter 3), will usually reveal such "clusters" of points.

ADJUSTED R^2 STATISTIC. Suppose p is the total number of parameters in a fitted model (including β_0) and RSS_p is the corresponding residual sum of squares. We have defined the R^2 statistic, a measure of the amount of variation about the mean explained by the fitted equation, as

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\overline{Y}^2} = 1 - \frac{\text{RSS}_p}{\text{CTSS}} \tag{2.6.11a}$$

where CTSS denotes the corrected total sum of squares $\mathbf{Y}'\mathbf{Y} - n\overline{Y}^2$, and where n is the total number of observations.

A related statistic, which has been used in the past, and is preferred by some workers, is the so-called *adjusted* R^2 defined, in our context, as

$$R_a^2 = 1 - \frac{(\text{RSS}_p)/(n-p)}{(\text{CTSS})/(n-1)} = 1 - (1-R^2)\left(\frac{n-1}{n-p}\right). \quad (2.6.11b)$$

An "adjustment" has been made for the corresponding degrees of freedom of the two quantities RSS_p and CTSS, the idea being that the statistic R_a^2 can be used to compare equations fitted not only to a specific set of data but also to two or more entirely different sets of data. (The value of this statistic for the latter purpose is, in our opinion, not high; R_a^2 might be useful as an initial gross indicator, but that is all.)

As pointed out by R. W. Kennard in "A note on the C_p statistic," *Technometrics*, **13**, 1971, 899–900, adjusted R^2 is closely related to the C_p statistic, a statistic used in one type of regression selection procedure. We discuss the use of C_p in Chapter 6. Apart from this, we do not use adjusted R^2 in this book.

(The equivalence of the numerators in Eqs. (2.6.11) and (2.6.11a) may be established as follows:

$$\sum (\hat{Y}_i - \bar{Y})^2 = \sum \hat{Y}_i^2 - (\sum Y_i)^2/n$$

and

$$\begin{aligned} \sum \hat{Y}_i^2 = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} &= (\mathbf{Xb})'(\mathbf{Xb}) \\ &= \mathbf{b}'\mathbf{X}'\mathbf{Xb} \\ &= \mathbf{b}'\mathbf{X}'\mathbf{Y} \end{aligned}$$

because $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$ from the normal Eqs. (2.6.3).)

With Distributional Assumptions

The analysis of variance breakup is an algebraic equality (or a geometric one, depending on one's viewpoint—see Section 10.6) only and does not depend on distributive properties of the errors. However, if we assume additionally that $\varepsilon_i \sim N(0, \sigma^2)$ and that the ε_i are independent—that is, that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, we can do the following:

1. Test lack of fit by treating the ratio

$$\left[\frac{\text{SS(lack of fit)}/(n-p-n_e)}{\text{SS(pure error)}/n_e}\right] \quad (2.6.12)$$

as an $F[(n-p-n_e), n_e]$ variate, and comparing its value with $F[(n-p-n_e), n_e, 1-\alpha]$. If there is no lack of fit, $\text{SS(residual)}/(n-p) = \text{MS}_E$ usually called s^2 is an unbiased estimate of σ^2. If lack of fit cannot be tested, use of s^2 as an estimate of σ^2 *implies* an assumption that the model is correct. (If it is not, s^2 will usually be too large since it is a random variable with a

mean *greater* than σ^2. Note carefully, however, that due to sampling fluctuation—since it *is* a random variable—it could also be too small.)

2. Test the overall regression equation (more specifically, test $H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ against H_1: not all $\beta_i = 0$) by treating the mean square ratio

$$\frac{[\mathrm{SS}(R|b_0)/(p-1)]}{s^2} \tag{2.6.13}$$

as an $F(p-1, v)$ variate where $v = n - p$

Suppose we decide on a specified risk level α. The fact that the observed mean-square ratio exceeds $F(p-1, v, 1-\alpha)$ means that a "statistically significant" regression has been obtained; in other words, the proportion of the variation observed in the data, which has been accounted for by the equation, is greater than would be expected by chance in $100(1-\alpha)\%$ similar sets of data with the same values of n and $\mathbf{X}$. This does not necessarily mean that the equation is useful for predictive purposes. Unless the range of values predicted by the fitted equation is considerably greater than the size of the random error, prediction will often be of no value even though a "significant" F-value has been obtained, since the equation will be "fitted to the errors" only.

Work by J. M. Wetz (in a 1964 Ph.D. thesis, "Criteria for judging adequacy of estimation by an approximating response function," written under the direction of Dr. G. E. P. Box at the University of Wisconsin) suggests that in order that an equation should be regarded as a satisfactory predictor (in the sense that the range of response values predicted by the equation is substantial compared with the standard error of the response), the observed F-ratio of (regression mean square)/(residual mean square) should exceed not merely the selected percentage point of the F-distribution, but at least *four times* the selected percentage point. For example, if $p = 11$, $v = 20$, $\alpha = 0.05$, $F(10, 20, 0.95) = 2.35$. Thus the observed F-ratio would have to exceed at least 9.4 for the fitted equation to be rated as a satisfactory prediction tool. For more detailed information, see Appendix 2C.

THE DISTRIBUTION OF R^2. We see that

$$\begin{aligned} R^2 &= \frac{\mathrm{SS(Regression}|b_0)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\ &= \frac{\mathrm{SS(Regression}|b_0)}{\mathrm{SS(Regression}|b_0) + \mathrm{Residual\ SS}} \\ &= \frac{v_1 F}{v_1 F + v_2}. \end{aligned} \tag{2.6.13a}$$

where the quantity

$$F = \frac{\text{SS(Regression}\,|\,b_0)/\nu_1}{\text{Residual SS}/\nu_2}$$

is our usual F statistic for testing overall regression given b_0, that is, for testing the null hypothesis H_0: that all the β's (excluding β_0) are zero against the alternative hypothesis H_1: that at least one of the β's (excluding β_0) is not zero. The value of β_0 is irrelevant to the test. To correspond to Eq. (2.6.13) we can set $\nu_1 = p - 1, \nu_2 = n - p$. Under H_0, F is distributed as an $F(\nu_1, \nu_2)$ variable. A statistical theorem tells us that R^2 follows a $\beta(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)$ distribution, called the beta distribution with (here) degrees of freedom $\frac{1}{2}\nu_1$ and $\frac{1}{2}\nu_2$. We shall not discuss the beta distribution at all but, clearly, if we had appropriate tables we could test H_0 against H_1 using R^2. The result would be *exactly* equivalent to that of our standard F-test, the significance point for R^2 being obtained from Eq. (2.6.13a) with $F(p - 1, n - p, 1 - \alpha)$ substituted for F. For this reason, and because tables of the beta distribution are not as universally available as those of F, a test on R^2 is rarely done.

3. If we use an estimate s_v^2 for σ^2, $100(1 - \alpha)\%$ confidence limits for the mean value of Y at $\mathbf{X}_0$ are obtained from

$$\hat{Y}_0 \pm t(v, 1 - \tfrac{1}{2}\alpha)s_v\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}.$$

4. State that

$$\mathbf{b} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2). \tag{2.6.14}$$

5. Obtain a joint $100(1 - \alpha)\%$ confidence region for *all* the parameters $\boldsymbol{\beta}$ from the equation

$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \leq ps^2F(p, v, 1 - \alpha) \tag{2.6.15}$$

where $F(p, v, 1 - \alpha)$ is the $1 - \alpha$ point ("upper α-point") of the $F(p, v)$ distribution and where s^2 has the same meaning as in (1) above and the model is assumed correct. In general this will be useful only when p is small, say 2, 3, or 4, unless care is taken to present the information in a form in which it can be readily understood. The inequality above provides the equation of an "elliptically shaped" contour in a space which has as many dimensions, p, as there are parameters in $\boldsymbol{\beta}$. We can obtain individual confidence intervals for the various parameters separately from the formula

$$b_i \pm t(v, 1 - \tfrac{1}{2}\alpha)(\text{estimated s.d.}(b_i))$$

where the "estimated s.d.(b_i)" is the square root of the ith diagonal term of the matrix $(\mathbf{X}'\mathbf{X})^{-1}s^2$. (For a calculation of this type when there are two parameters β_0 and β_1, see Eq. (2.3.1), and after replacement of σ^2 by s^2, see Section 1.4.) Separate confidence intervals of this type appear in our

printouts and are often useful. We de-emphasize them, however, for the following reason. Figure 2.1 illustrates a possible situation that may arise when two parameters are considered. The joint 95% confidence region for the true parameters, β_1 and β_2, is shown as a long thin ellipse and encloses values (β_1, β_2) which the data regard as *jointly* reasonable for the parameters. It takes into account the correlation between the estimates b_1 and b_2. The individual 95% confidence intervals for β_1 and β_2 separately are appropriate for specifying ranges for the individual parameters irrespective of the value of the other parameter. If an attempt is made to interpret these intervals simultaneously—that is (wrongly), regard the rectangle which they define as a joint confidence region—then, for example, it may be thought that the coordinates of the point E provide reasonable values for (β_1, β_2). The joint confidence region, however, clearly indicates that such a point is not reasonable. When only two parameters are involved, construction of the confidence ellipse is not difficult. When more parameters are involved the calculations are not difficult to handle in a computer but interpretation is difficult. One possible solution is to find the coordinates of the points at the ends of the major axes of the region. (In Figure 2.1 these would be the points A, B, C, and D.) This would involve obtaining the confidence contour and reducing it to canonical form. This is not difficult and is done via methods related to those in Sections 6.9 and 6.10. However, we confine ourselves to pointing out the moral that the "joint" message of individual confidence intervals should be regarded with caution, and attention should be paid both to the relative sizes of the $V(b_i)$ and to the sizes of the covariances of b_i and

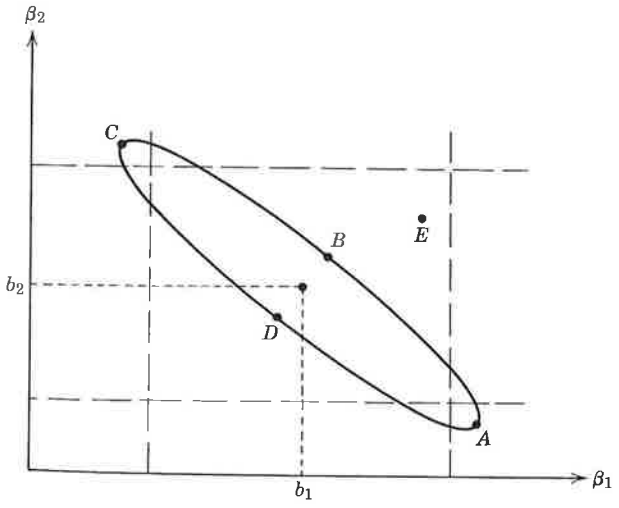

Figure 2.1 Joint and individual confidence statements.

b_j. When b_i and b_j have variances of different sizes and the correlation between b_i and b_j, namely

$$\rho_{ij} = \frac{\text{cov}(b_i, b_j)}{[V(b_i)V(b_j)]^{1/2}}$$

is not small, the situation illustrated in Figure 2.1 occurs. If ρ_{ij} is close to zero then the rectangular region defined by individual confidence intervals will approximate to the correct joint confidence region, though the joint region is correct. The elongation of the region will depend on the relative sizes of $V(b_i)$ and $V(b_j)$. Some examples are shown in Figure 2.2.

Note. If the model is written originally, and fitted, in the alternative form

$$E(Y - \bar{Y}) = \beta_1(X_1 - \bar{X}_1) + \beta_2(X_2 - \bar{X}_2) + \cdots + \beta_k(X_k - \bar{X}_k)$$

where $\bar{Y}, \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ are the observed means of the actual data, then joint confidence intervals can be obtained that do not involve β_0, which sometimes is of little interest.

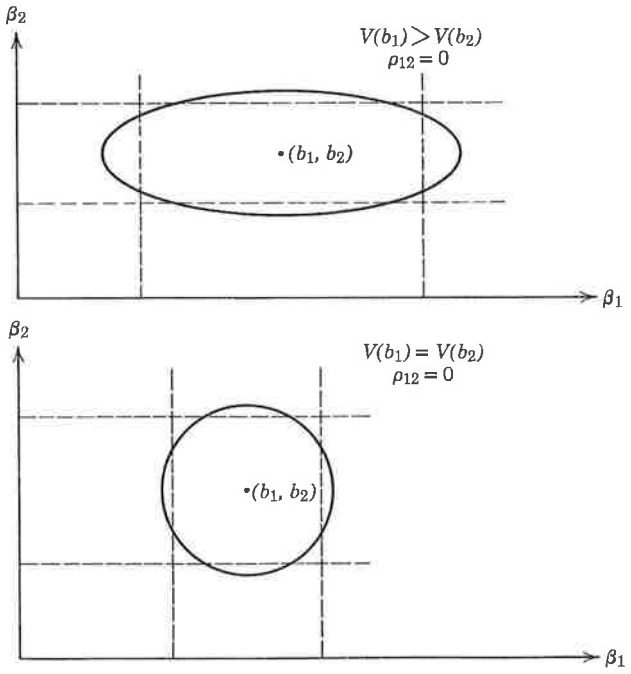

Figure 2.2 Examples of situations where individual confidence intervals combine well to approximate a joint confidence region for two parameters.