CHAPTER 5

The General Regression Situation

In presenting the general regression situation we state many results without proving them. For proofs, the reader could consult, for example, Plackett (1960), Seber (1977), or Rao (1973).

5.1. GENERAL LINEAR REGRESSION

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.1.1}$$

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,
$\boldsymbol{\epsilon}$ is an $(n \times 1)$ vector of errors,

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

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

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}. \tag{5.1.1a}$$

The error sum of squares is then

$$\begin{aligned} \boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{5.1.2}$$

[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. (5.1.2), minimizes $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$. It can be determined by differentiating Eq. (5.1.2) 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 $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ with respect to a vector quantity $\boldsymbol{\beta}$ is equivalent to differentiating $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ 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*

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{5.1.3}$$

Two main cases arise: either Eq. (5.1.3) 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 $\mathbf{b}$). If some of the normal equations depend on others, $\mathbf{X'X}$ is singular, so that $(\mathbf{X'X})^{-1}$ 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 23. If the p normal equations are independent, $\mathbf{X'X}$ is nonsingular, and its inverse exists. In this case the solution of the normal equations can be written

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}. \tag{5.1.4}$$

This solution $\mathbf{b}$ has the following properties:

1. It is an estimate of $\boldsymbol{\beta}$ that minimizes the error sum of squares $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$, *irrespective* of any distribution properties of the errors.
Note: An assumption that the errors $\boldsymbol{\epsilon}$ are normally distributed is *not* required in order to obtain the estimate $\mathbf{b}$ but it *is* required later in order to make tests that 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 $\mathbf{b}$ 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 variance (of *any* linear functions of the Y's that provide unbiased estimates), irrespective of distribution properties of the errors.
Note: Suppose we have an expression $T = l_1Y_1 + l_2Y_2 + \cdots + l_nY_n$, 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 that we can write as $E(T)$ and a variance that we can write as $V(T)$. 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* or the *Gauss–Markov* Theorem. See Jaske (1994).]

A Justification for Using Least Squares

3. If the errors are independent and $\epsilon_i \sim N(0, \sigma^2)$, then $\mathbf{b}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. (In vector terms we can write $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, meaning that $\boldsymbol{\epsilon}$ follows an n-dimensional multivariate normal distribution with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ (where $\mathbf{0}$ denotes a vector consisting entirely of zeros and of the same length as $\boldsymbol{\epsilon}$) and $\mathbf{V}(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$; that is, $\boldsymbol{\epsilon}$ has a variance–covariance matrix whose diagonal elements, $V(\epsilon_i)$, $i = 1, 2, \ldots,$

n, are all σ^2 and whose off-diagonal elements, covariance (ϵ_i, ϵ_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

$$\prod_{i=1}^{n} \frac{1}{\sigma(2\pi)^{1/2}} e^{-\epsilon_i^2/(2\sigma^2)} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/2\sigma^2}. \tag{5.1.5}$$

Thus for a fixed value of σ, maximizing the likelihood function is equivalent to minimizing the quantity $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$. 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:

$$f(\epsilon_i) = (2\sigma)^{-1} e^{-|\epsilon_i|/\sigma} \qquad (-\infty \leq \epsilon_i \leq \infty) \tag{5.1.6}$$

rather than the normal distribution:

$$f(\epsilon_i) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-\epsilon_i^2/2\sigma^2} \tag{5.1.7}$$

as is usually assumed. The double exponential frequency function has a pointed peak of height $1/2\sigma$ at $\epsilon_i = 0$, 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

$$\sum_{i=1}^{n} |\epsilon_i|,$$

the sum of absolute errors, and not the minimization of

$$\sum_{i=1}^{n} \epsilon_i^2,$$

the sum of *squares* of errors.

5.2. LEAST SQUARES PROPERTIES

Assuming that $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathbf{V}(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$, we can proceed with the following steps whether the errors are normally distributed or not.

1. The fitted values are obtained from $\hat{\mathbf{Y}} = \mathbf{Xb}$.
2. The vector of residuals is given by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.

It is true that $\Sigma_{i=1}^{n} e_i \hat{Y}_i = 0$, whatever the linear 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 $\Sigma_{i=1}^{n} e_i = 0$. (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. Thus $\mathbf{e}'\hat{\mathbf{Y}} = 0 = \hat{\mathbf{Y}}'\mathbf{e}$ always, and $\mathbf{e}'\mathbf{1} = 0 = \mathbf{1}'\mathbf{e}$ when the model contains β_0.)

3. $\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ 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 location defined by* $\mathbf{X}_0$. For example, if the model were $Y = \beta_0 + \beta_1 X + \beta_{11}X^2 + \epsilon$, 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

$$V(\hat{Y}_0) = \mathbf{X}_0'\mathbf{V}(\mathbf{b})\mathbf{X}_0 = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2. \tag{5.2.1}$$

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

Source	df	SS	MS
Regression	p	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	$\text{MS}_{\text{Regression}}$
Residual	$n - p$	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$\text{MS}_{\text{Residual}}$
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

$$\text{SS}(b_0) = \frac{(\Sigma\, Y_i)^2}{n} = n\overline{Y}^2 \tag{5.2.2}$$

$$\text{SS}(\text{Regression}|b_0) = \text{SS}(\text{Reg}|b_0) = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{(\Sigma\, Y_i)^2}{n}. \tag{5.2.3}$$

These sums of squares are based on 1 and $p - 1$ degrees of freedom, respectively.

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 predictor variables (though approximate use of "very close" points is sometimes seen in practice). This provides an analysis of variance table as follows. (Note: *lof* = lack of fit, *pe* = pure error).

Source	df	SS	MS	
b_0	1	SS(b_0)		$\text{MS}_{\text{Regression}}$
Regression$|b_0$	$p - 1$	SS(Reg$|b_0$)	MS(Reg$|b_0$)	
Lack of fit	$n - p - n_e$	SS(*lof*)	MS(*lof*)	$\text{MS}_{\text{Residual}}$
Pure error	n_e	SS(*pe*)	MS(*pe*)	
Total	n	$\mathbf{Y}'\mathbf{Y}$		

The R^2 Statistic

The ratio

$$R^2 = \frac{\text{SS}(\text{Reg}|b_0)}{\mathbf{Y}'\mathbf{Y} - \text{SS}(b_0)} = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} \tag{5.2.4}$$

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.*

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 in Section 2.1. A perfect fit to the data for which $\hat{Y}_i = Y_i$, an unlikely event in practice, would give $R^2 = 1$.

If $\hat{Y}_i = \overline{Y}$, that is, if $b_1 = b_2 = \cdots b_{p-1} = 0$ (or if a model $Y = \beta_0 + \epsilon$ 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.

R^2 Can Be Deceptive

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 that 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 100 observations that occur in five groups each of 20 repeats, we have effectively five pieces of information, represented by five mean values, and 95 degrees of freedom for pure error, 19 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 100 observations can be well predicted by a model with only five parameters is not surprising since there are really only five distinct data sites and not 100 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, 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_{n-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

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\overline{Y}^2} = 1 - \frac{\text{RSS}_{n-p}}{\text{CTSS}} \tag{5.2.5}$$

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 is preferred by some workers, is the *adjusted* R^2 defined, in our context, as

$$R_a^2 = 1 - \frac{(\mathrm{RSS}_{n-p})/(n-p)}{(\mathrm{CTSS})/(n-1)} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right). \tag{5.2.6}$$

An "adjustment" has been made for the corresponding degrees of freedom of the two quantities RSS_{n-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 this is all.)

As pointed out by Kennard (1971), 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 15. Apart from this, we do not use adjusted R^2 in this book.

The equivalence of the numerators in Eqs. (5.2.4) and (5.2.5) may be established as follows:

$$\Sigma(\hat{Y}_i - \overline{Y})^2 = \Sigma\,\hat{Y}_i^2 - (\Sigma\,Y_i)^2/n$$

and

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

because $\mathbf{X'Xb} = \mathbf{X'Y}$ from the normal equations.

5.3. LEAST SQUARES PROPERTIES WHEN $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$

The analysis of variance breakup is an algebraic equality (or a geometric one, depending on one's viewpoint—see Chapter 20) only and does not depend on distributive properties of the errors. However, if we assume additionally that $\epsilon_i \sim N(0, \sigma^2)$ and that the ϵ_i are independent—that is, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$—we can do the following.

1. Test lack of fit by treating the ratio

$$\left[\frac{\mathrm{SS(lack\ of\ fit)}/(n-p-n_e)}{\mathrm{SS(pure\ error)}/n_e}\right] \tag{5.3.1}$$

as an $F[(n-p-n_e), n_e]$ variate and by comparing its value with $F[(n-p-n_e), n_e, 1-\alpha]$. If there is no lack of fit, SS(residual)/$(n-p) = \mathrm{MS}_E$, 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 $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ against $H_1\colon$not all $\beta_i = 0$) by treating the mean square ratio

$$\frac{[\mathrm{SS(Reg}|b_0)/(p-1)]}{s^2} \tag{5.3.2}$$

as an $F(p-1, \nu)$ variate, where $\nu = n - p$.

Just Significant Regressions May Not Predict Well

Suppose we decide on a specified risk level α. The fact that the observed mean square ratio exceeds $F(p - 1, \nu, 1 - \alpha)$ means that a "statistically significant" regression has been obtained; in other words, the proportion of the variation in the data which has been accounted for by the fitted equation is deemed greater than would be expected by chance in 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. For more on this, see Section 11.1.

The Distribution of R^2

We see that

$$\begin{aligned} R^2 &= \frac{\text{SS(Regression}|b_0)}{\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2} \\ &= \frac{\text{SS(Regression}|b_0)}{\text{SS(Regression}|b_0) + \text{Residual SS}} \\ &= \frac{\nu_1 F}{\nu_1 F + \nu_2}, \end{aligned} \tag{5.3.3}$$

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. (5.3.2) 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 and (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. (5.3.3) 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.

Properties, Continued

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

$$\hat{Y}_0 \pm t(\nu, 1 - \tfrac{1}{2}\alpha)s_\nu\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}. \tag{5.3.4}$$

4. State that

$$\mathbf{b} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2). \tag{5.3.5}$$

5. Obtain individual $100(1 - \alpha)\%$ confidence intervals for the various parameters separately from the formula

$$b_i \pm t(\nu, 1 - \alpha/2)\text{se}(b_i) \tag{5.3.6}$$

where the "se(b_i)" is the square root of the ith diagonal term of the matrix $(\mathbf{X}'\mathbf{X})^{-1}s^2$. These intervals can be used to define a rectangular block in the space of the β's. This block is *not* a proper joint confidence region for the β's, however; see (6) instead.

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

$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = ps^2F(p, \nu, 1 - \alpha), \tag{5.3.7}$$

where $F(p, \nu, 1 - \alpha)$ is the $1 - \alpha$ point ("upper α-point") of the $F(p, \nu)$ distribution and where s^2 has the same meaning as in (1) above and the model is assumed correct. This equality is the equation of the boundary of an "elliptically shaped" (or "ellipsoidally shaped" more generally) contour in a space that has as many dimensions, p, as there are parameters in $\boldsymbol{\beta}$. (Such regions can also be constructed for a subset of the parameters.)

Comparisons between (5) and (6) are discussed in Sections 5.4 and 5.5.

Bonferroni Limits

A more conservative (wider) set of intervals of the form of (5.3.6) is obtained if we replace $t(\nu, 1 - \alpha/2)$ by $t(\nu, 1 - \alpha/(2p))$. It can be shown that these intervals jointly have a confidence coefficient of at least $1 - \alpha$. These, and other sets of rectangular limits, are described in Nickerson (1994).

5.4. CONFIDENCE INTERVALS VERSUS REGIONS

Figure 5.1 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

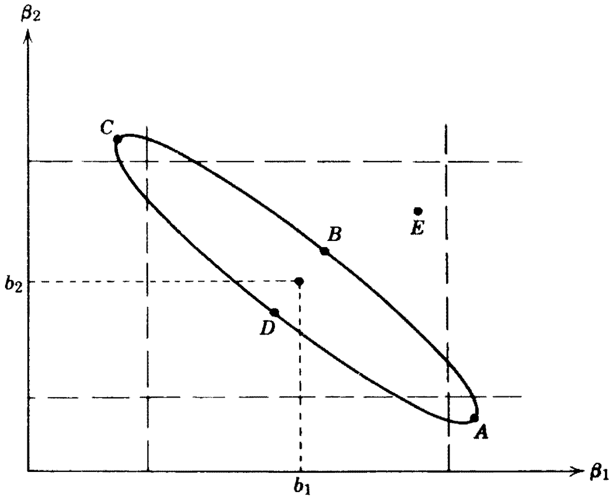

Figure 5.1. Joint and individual confidence statements. The point (b_1, b_2) defined by the least squares estimates is at the center of both ellipse and rectangle.

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 that 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. In practice, even for two parameters, it is rarely drawn.

If some knowledge of the ellipsoidal region were desired, it would be possible to find the coordinates of the points at the ends of the major axes of the region. (In Figure 5.1 these would be the points A, B, C, and D.) This would involve obtaining the confidence contour and reducing it to canonical form. This also is not difficult, but we do not discuss it, because it is rarely done. The major point to be made here is 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 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 5.1 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 5.2.

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

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

where $\overline{Y}, \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ 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.

See Exercise M in "Exercises for Chapters 5 and 6."

Moral

We shall nearly always look at individual confidence intervals that form a "rectangular brick" in the number of dimensions defined by the number of parameters. This brick is not a correct joint confidence region, which, in general, is a "difficult to see and appreciate" ellipsoidal shape. Knowledge of the correlations between the parameter estimates could be helpful in relating the brick and the ellipse, if the effort were thought to be worthwhile.

5.5. MORE ON CONFIDENCE INTERVALS VERSUS REGIONS

We now discuss a one-number calculation that can be useful in comparing a confidence interval block with an ellipsoidal joint confidence region; see Figures 5.1 and 5.2. In

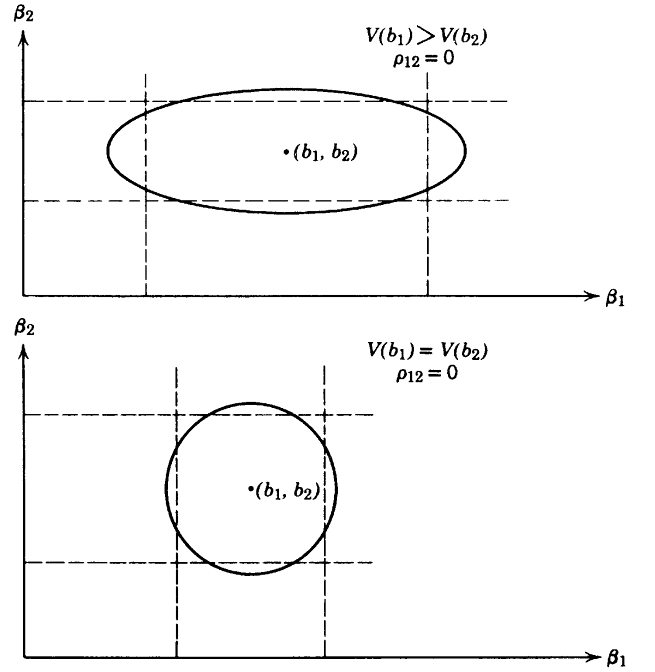

Figure 5.2. Examples of situations where individual confidence intervals combine well to approximate a joint confidence region for two parameters.

this section, we number the model parameters as $\beta_1, \beta_2, \ldots, \beta_p$ (rather than $\beta_0, \beta_1, \ldots, \beta_{p-1}$) to simplify the notation slightly. We first rewrite (5.3.6) and (5.3.7) in the forms

$$b_i \pm t(\nu, 1 - \alpha/2)(V_{ii}s^2)^{1/2}, \qquad i = 1, 2, \ldots, p, \tag{5.5.1}$$

where

$$(V_{ij}) = \mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}, \qquad i, j = 1, 2, \ldots, p,$$

and

$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = ps^2F(p, \nu, 1 - \theta). \tag{5.5.2}$$

In our discussion in Section 5.3, we took $\theta = \alpha$, but this choice is not necessary. In Exercise M in "Exercises for Chapters 5 and 6," we argue that $\alpha = 0.05$ and $\theta = 1 - (1 - \alpha)^2 = 0.10$, approximately, might be appropriate. Such a choice might be sensible if there were no correlations between the estimates b_i. This is unlikely unless the regression follows up a carefully designed experiment. If all the b_i *were* mutually uncorrelated, however, **X'X** would be a diagonal matrix, and the major axes of the ellipse would be parallel to the sides of the rectangular block. Suppose α is given; it is often chosen as 0.05. Then we could chooseθ in such a way that, *when all the b_i are uncorrelated*, the rectangular block and the ellipsoid are of the same size. In general, the volume of the rectangular region is

$$R = 2^p t^p s^p (V_{11}V_{22} \cdots V_{pp})^{1/2}, \tag{5.5.3}$$

where $t = t(\nu, 1 - \alpha/2)$. The volume of the ellipsoidal region is given by a constant (depending on the dimension p) times the product of the semi-axial lengths. It can be shown that this volume is

$$E \equiv \frac{\pi^{p/2}}{\Gamma(p/2+1)}(ps^2F)^{p/2}c_p^{1/2}(V_{11}V_{22}\cdots V_{pp})^{1/2}, \tag{5.5.4}$$

where $\pi = 3.14159$, and where the gamma functions required satisfy $\Gamma(u) = (u-1)\Gamma(u-1)$, $\Gamma(1) = 1$, $\Gamma(\frac{1}{2}) = \pi^{1/2}$. The quantity c_p is defined as the determinant of a normalized form of $(\mathbf{X}'\mathbf{X})^{-1}$, namely, of $\{V_{ij}/(V_{ii}V_{jj})^{1/2}\}$. Thus c_p is simply the determinant of the correlation matrix of $b_1, b_2, \ldots, b_p$.

The ratio of the volumes of the two regions is, in general,

$$\frac{E}{R} = \frac{\pi^{p/2}}{\Gamma(p/2+1)}\,\frac{p^{p/2}F^{p/2}c_p^{1/2}}{2^p t^p}. \tag{5.5.5}$$

We now link α and θ in the following manner. We specify that when $c_p = 1$, that is, when the b_i are uncorrelated, $E = R$, and so their ratio is 1. this implies that, in the case where the major axes of the ellipsoid are parallel to the axes of the β's, we would wish to link α and θ so that the ellipsoid and its approximating rectangular block have equal volume. This requires that

$$\begin{aligned} F(p,\nu,\theta) &= \left\{\frac{\Gamma(p/2+1)}{\pi^{p/2}}\,\frac{2^p t^p}{p^{p/2}}\right\}^{2/p} \\ &= \frac{4\{\Gamma(p/2+1)\}^{2/p}}{\pi p}\{t(\nu, 1-\alpha/2)\}^2 \\ &= \frac{4\{\Gamma(p/2+1)\}^{2/p}}{\pi p}\{F(1,\nu,\alpha)\}. \end{aligned} \tag{5.5.6}$$

Note, as a check, that when $p = 1$, we get the obvious $\theta = \alpha$, for the one-parameter case. When $n \geq 10$, and $\alpha = 0.05$, the θ values vary little for given p. For $p = 2$, for example, the case most often depicted, and for uncorrelated b_i values, we obtain an elliptical confidence region of size equal to the rectangular region based on separate 95% confidence intervals if we use a 91.3% confidence ellipse (approximately). Similar approximate results for some other cases are 88.4% for $p = 3$, 86.0% for $p = 4$, 83.9% for $p = 5$, and 82.1% for $p = 6$.

The above calculations lead us to an easy way to assess how well the rectangular block can represent the correct ellipsoidal region in a regression for any value of p. Using any selected linked values α and θ that satisfy (5.5.6) we see that the right-hand side of (5.5.5) reduces to $c_p^{1/2}$. This value gives the ratio of the volume of the ellipsoidal confidence region compared to the volume of the rectangular block. Note that $0 \leq c_p^{1/2} \leq 1$, the zero corresponding to linear dependence in the $\mathbf{X}$-columns and the 1 to an orthogonal set of $\mathbf{X}$-columns. A relative volume calculation can be made from (5.5.5) even if an ellipsoid other than the one satisfying (5.5.6) is selected, of course. Note that our calculations can also be applied to the slightly different but closely related suggestions for confidence regions made by Weisberg (1985, pp. 97–99, Eqs. (4.2.6) and (4.2.7)). Further information can also be found, if needed, by a canonical analysis of (5.5.2) using the value of θ linked to α.

Example, $p = 2$. Consider the straight line steam-data fit in Chapter 1. Take $\alpha = 0.05$. For this fit, $p = 2$, $\nu = 23$, the right-hand side of (5.5.6) is 2.7241, $1 - \theta = 0.9132$, which we round to 0.913. The correlation between intercept and slope is $r = V_{12}/(V_{11}V_{22})^{1/2} = (-0.0073535)/[(0.4267941)(0.0001398)]^{1/2} = -(0.90628)^{1/2} = -0.952$. Thus $c_p^{1/2} = (1 - r^2)^{1/2} = 0.306$ and so the 91.3% ellipse covers about 30.6% of the area of the rectangle. Moreover, the high negative correlation indicates a diagonal upper-left-

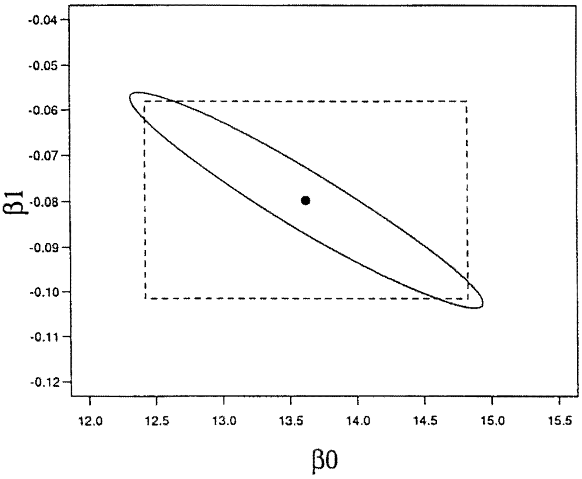

Figure 5.3. Individual 95% confidence bands and a 91.3% joint confidence region for the steam data (Appendix 1A, Y and X_8).

to-lower-right-lying ellipse. Figure 5.3 shows that these simple calculations describe the situation well. (If the ellipse in Figure 5.3 were replaced by the 95% ellipse, which surrounds the 91.3% ellipse and juts out somewhat more at the upper-left and lower-right extremes, the area covered would increase to about 38.4% of the area of the rectangle.)

Example, p = 4. Consider (see Appendix 1A) the steam data again, in particular, the planar fit of the response variable onto predictors X_5, X_6, and X_8 and an intercept. We have $p = 4$, $\nu = 21$ and the sides of the rectangular t-block for $\alpha = 0.05$ are $-13.02 \leq \beta_0 \leq -7.08$, $0.075 \leq \beta_5 \leq 0.729$, $0.114 \leq \beta_6 \leq 0.284$, and $-0.089 \leq \beta_8 \leq -0.074$. The predictors are very highly correlated, however, and $c_p^{1/2} = 0.000977$. Thus the ellipsoid defined by $\theta = 0.14$ $(1 - \theta = 0.86)$ has a volume of only about 0.1% of the rectangular block. The latter thus gives a totally misleading impression; the ellipsoid is an extremely long thin one. (Even if variable X_5 is dropped, the ellipsoid still represents only about 4.3% of the three-dimensional rectangular t-block.)

Conclusion. The value of $c_p^{1/2}$, where c_p is the determinant of the correlation matrix of the b_i, is a useful calculation to display in regression problems. It provides the ratio of the volume of the (correct) ellipsoidal joint confidence region for the β's to the (wrong but easily obtained) rectangular t-block region, for linked α and θ values that achieve equal volumes in the uncorrelated case. Supplementary information can also be obtained from a canonical reduction of the ellipse's equation. (See, for example, Box and Draper, 1987, pp. 332–372.) The eigenvectors will give the exact orientations of the major axes of the ellipsoid with respect to the β-axes. These orientations depend on the correlations between the various pairs of b's. In particular, if $c_p = 1$, the axes of the ellipse are aligned exactly with the β-axes. Diagrams such as Figures 5.1–5.3 are not needed, once their nature is understood.

When *F*-Test and *t*-Tests Conflict

Occasionally one finds a practical regression problem where an overall F-test for regression given b_0 is significant, but all the t-tests for individual hypotheses $H_0: \beta_i = 0$

are not significant. Largey and Spencer (1996) discuss how such occurrences are related to diagrams of the form of Figure 5.3. (The reverse case, a nonsignificant F but significant t-values, is possible but even rarer.)

References

Box and Draper (1987); Draper and Guttman (1995); Largey and Spencer (1996); Weisberg (1985); Willan and Watts (1978).

APPENDIX 5A. SELECTED USEFUL MATRIX RESULTS

For a more comprehensive list of results see, for example, Graybill (1961) or Rao (1973).

1. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$, etc. If $\mathbf{M}' = \mathbf{M}$, both are symmetric.

2. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

3. A square matrix $\mathbf{C}$ is said to be *orthogonal* if $\mathbf{C}'\mathbf{C} = \mathbf{I}$. Then $\mathbf{C}' = \mathbf{C}^{-1}$.

4. A square matrix $\mathbf{M}$ is said to be idempotent if $\mathbf{MM} = \mathbf{M}$. This can be written as $\mathbf{M}^2 = \mathbf{M}$, also.

5. If $\mathbf{M}$ is symmetric and idempotent,

$$(\mathbf{I} - 2\mathbf{M})'(\mathbf{I} - 2\mathbf{M}) = \mathbf{I}.$$

Thus any matrix of the form $\mathbf{I} - 2\mathbf{M}$, where M is symmetric and idempotent, is orthogonal.

6. Trace $(\mathbf{AB})$ = trace $(\mathbf{BA})$, where trace denotes the sum of the diagonal elements of a square matrix. ($\mathbf{A} = p \times q$, $\mathbf{B} = q \times p$, say.)

7. If

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \text{and if} \quad \begin{cases} \mathbf{P} = \mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}, \\ \mathbf{Q} = \mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}, \end{cases}$$

then

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{P}^{-1} & -\mathbf{A}^{-1}\mathbf{BQ}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CP}^{-1} & \mathbf{Q}^{-1} \end{bmatrix}$$

assuming all matrices shown inverted are nonsingular. Alternatively,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BQ}^{-1}\mathbf{CA}^{-1} & -\mathbf{A}^{-1}\mathbf{BQ}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{CA}^{-1} & \mathbf{Q}^{-1} \end{bmatrix}.$$

If $\mathbf{M}$ is symmetric, set $\mathbf{C} = \mathbf{B}'$.

8. If $\mathbf{E}$ is $n \times p$, and $\mathbf{F}$ is $p \times n$, then

$$(\mathbf{I}_n + \mathbf{EF})^{-1} = \mathbf{I}_n - \mathbf{E}(\mathbf{I}_p + \mathbf{FE})^{-1}\mathbf{F}.$$

This is especially useful when p is much smaller than n.

Special Case 1. If $\mathbf{X}$ is $n \times p$

$$\begin{aligned}(\mathbf{I}_n + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')^{-1} &= \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{I}_p + \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{X}' \\ &= \mathbf{I}_n - \tfrac{1}{2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.\end{aligned}$$

Thus $(\mathbf{I}_n + \mathbf{H})^{-1} = \mathbf{I}_n - \frac{1}{2}\mathbf{H}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

Special Case 2. If $\mathbf{A}$ is $n \times n$, and $\mathbf{u}$, $\mathbf{v}$ are $n \times 1$ vectors, then

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = (\mathbf{I} + \mathbf{A}^{-1}\mathbf{u}\mathbf{v}')^{-1}\,\mathbf{A}^{-1} = \mathbf{A}^{-1} - (\mathbf{A}^{-1} - (\mathbf{A}^{-1}\mathbf{u})(\mathbf{v}'\mathbf{A}^{-1})/\{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}\}.$$

This enables inversion of $\mathbf{A} + \mathbf{u}\mathbf{v}$ from knowledge of $\mathbf{A}^{-1}$. (Set $\mathbf{E} = \mathbf{A}^{-1}\mathbf{u}$, $\mathbf{F} = \mathbf{v}'$.)

9. If $\mathbf{A}$ is $p \times p$, $\mathbf{B}$ is $p \times q$, $\mathbf{C}$ is $q \times p$, and $\mathbf{D}$ is $q \times q$, then

$$\begin{vmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{vmatrix} = |\mathbf{A}||\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}||\mathbf{D}|$$

Proof. Premultiply the original matrix by

$$\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ -\mathbf{C}\mathbf{A}^{-1} & I_q\end{bmatrix}$$

to give a matrix equation; then take determinants of both sides.

Special Case 1. Set $\mathbf{C} = -\mathbf{B}'$, $\mathbf{D} = \mathbf{I}$ and we obtain the result

$$|\mathbf{A}||\mathbf{I} + \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{A} + \mathbf{B}\mathbf{B}'|.$$

Special Case 2. Set $\mathbf{C} = \mathbf{B}'$ if the partitioned matrix is symmetric.

A useful reference for some special inverse matrices is Roy and Sarhan (1956).

EXERCISES

Exercises for Chapter 5 are located in the Section "Exercises for Chapters 5 and 6" at the end of Chapter 6.