Chapter 2

Linear Models for Continuous Data

The starting point in our exploration of statistical models in social research will be the classical linear model. Stops along the way include multiple linear regression, analysis of variance, and analysis of covariance. We will also discuss regression diagnostics and remedies.

2.1 Introduction to Linear Models

Linear models are used to study how a quantitative variable depends on one or more predictors or explanatory variables. The predictors themselves may be quantitative or qualitative.

2.1.1 The Program Effort Data

We will illustrate the use of linear models for continuous data using a small dataset extracted from Mauldin and Berelson (1978) and reproduced in Table 2.1. The data include an index of social setting, an index of family planning effort, and the percent decline in the crude birth rate (CBR)—the number of births per thousand population—between 1965 and 1975, for 20 countries in Latin America and the Caribbean.

The index of social setting combines seven social indicators, namely literacy, school enrollment, life expectancy, infant mortality, percent of males aged 15–64 in the non-agricultural labor force, gross national product *per capita* and percent of population living in urban areas. Higher scores represent higher socio-economic levels.

TABLE 2.1: The Program Effort Data

	Setting	Effort	CBR Decline
Bolivia	46	0	1
Brazil	74	0	10
Chile	89	16	29
Colombia	77	16	25
CostaRica	84	21	29
Cuba	89	15	40
Dominican Rep	68	14	21
Ecuador	70	6	0
El Salvador	60	13	13
Guatemala	55	9	4
Haiti	35	3	0
Honduras	51	7	7
Jamaica	87	23	21
Mexico	83	4	9
Nicaragua	68	0	7
Panama	84	19	22
Paraguay	74	3	6
Peru	73	0	2
Trinidad-Tobago	84	15	29
Venezuela	91	7	11

The index of family planning effort combines 15 different program indicators, including such aspects as the existence of an official family planning policy, the availability of contraceptive methods, and the structure of the family planning program. An index of 0 denotes the absence of a program, 1–9 indicates weak programs, 10–19 represents moderate efforts and 20 or more denotes fairly strong programs.

Figure 2.1 shows scatterplots for all pairs of variables. Note that CBR decline is positively associated with both social setting and family planning effort. Note also that countries with higher socio-economic levels tend to have stronger family planning programs.

In our analysis of these data we will treat the percent decline in the CBR as a continuous response and the indices of social setting and family planning effort as predictors. In a first approach to the data we will treat the predictors as continuous covariates with linear effects. Later we will group

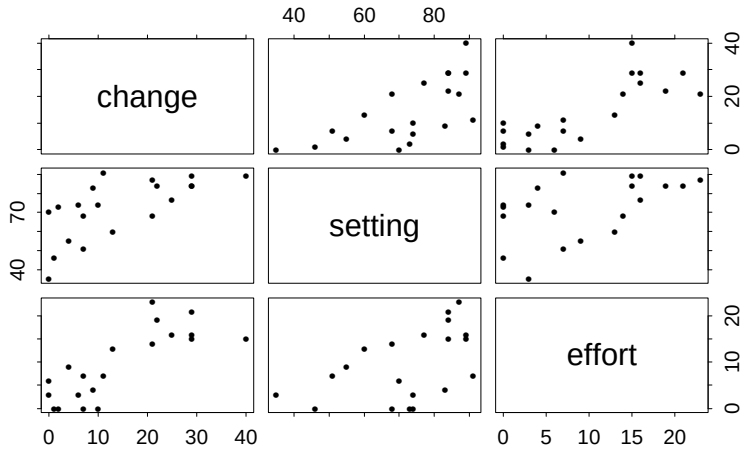

FIGURE 2.1: Scattergrams for the Program Effort Data

them into categories and treat them as discrete factors.

2.1.2 The Random Structure

The first issue we must deal with is that the response will vary even among units with identical values of the covariates. To model this fact we will treat each response y_i as a realization of a random variable Y_i. Conceptually, we view the observed response as only one out of many possible outcomes that we could have observed under identical circumstances, and we describe the possible values in terms of a probability distribution.

For the models in this chapter we will assume that the random variable Y_i has a normal distribution with mean μ_i and variance σ^2, in symbols:

$$Y_i \sim N(\mu_i, \sigma^2).$$

The mean μ_i represents the expected outcome, and the variance σ^2 measures the extent to which an actual observation may deviate from expectation.

Note that the expected value may vary from unit to unit, but the variance is the same for all. In terms of our example, we may expect a larger fertility decline in Cuba than in Haiti, but we don't anticipate that our expectation will be closer to the truth for one country than for the other.

The normal or *Gaussian* distribution (after the mathematician Karl Gauss) has probability density function

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\}. \tag{2.1}$$

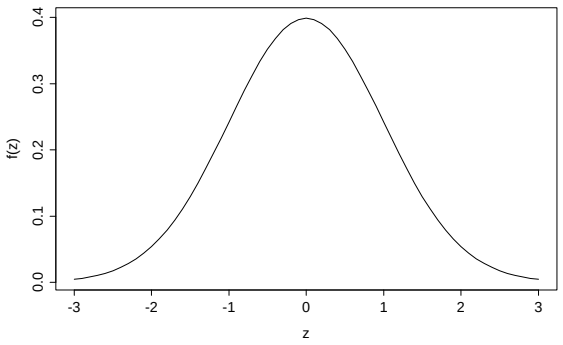

FIGURE 2.2: The Standard Normal Density

The standard density with mean zero and standard deviation one is shown in Figure 2.2.

Most of the probability mass in the normal distribution (in fact, 99.7%) lies within three standard deviations of the mean. In terms of our example, we would be very surprised if fertility in a country declined 3σ more than expected. Of course, we don't know yet what to expect, nor what σ is.

So far we have considered the distribution of one observation. At this point we add the important assumption that the observations are mutually *independent*. This assumption allows us to obtain the joint distribution of the data as a simple product of the individual probability distributions, and underlies the construction of the likelihood function that will be used for estimation and testing. When the observations are independent they are also uncorrelated and their covariance is zero, so $\text{cov}(Y_i, Y_j) = 0$ for $i \neq j$.

It will be convenient to collect the n responses in a column vector $\mathbf{y}$, which we view as a realization of a random vector $\mathbf{Y}$ with mean $\text{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and variance-covariance matrix $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The diagonal elements of $\text{var}(\mathbf{Y})$ are all σ^2 and the off-diagonal elements are all zero, so the n observations are uncorrelated and have the same variance. Under the assumption of normality, $\mathbf{Y}$ has a multivariate normal distribution

$$\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{2.2}$$

with the stated mean and variance.

2.1.3 The Systematic Structure

Let us now turn our attention to the systematic part of the model. Suppose that we have data on p predictors $x_1, \ldots, x_p$ which take values $x_{i1}, \ldots, x_{ip}$

for the i-th unit. We will assume that the expected response depends on these predictors. Specifically, we will assume that μ_i is a *linear* function of the predictors

$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

for some unknown coefficients $\beta_1, \beta_2, \ldots, \beta_p$. The coefficients β_j are called *regression coefficients* and we will devote considerable attention to their interpretation.

This equation may be written more compactly using matrix notation as

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{2.3}$$

where $\mathbf{x}_i'$ is a row vector with the values of the p predictors for the i-th unit and $\boldsymbol{\beta}$ is a column vector containing the p regression coefficients. Even more compactly, we may form a column vector $\boldsymbol{\mu}$ with all the expected responses and then write

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{2.4}$$

where $\mathbf{X}$ is an $n \times p$ matrix containing the values of the p predictors for the n units. The matrix $\mathbf{X}$ is usually called the *model* or design matrix. Matrix notation is not only more compact but, once you get used to it, it is also easier to read than formulas with lots of subscripts.

The expression $\mathbf{X}\boldsymbol{\beta}$ is called the *linear predictor*, and includes many special cases of interest. Later in this chapter we will show how it includes simple and multiple linear regression models, analysis of variance models and analysis of covariance models.

The simplest possible linear model assumes that every unit has the same expected value, so that $\mu_i = \mu$ for all i. This model is often called the *null* model, because it postulates no systematic differences between the units. The null model can be obtained as a special case of Equation 2.3 by setting $p = 1$ and $x_i = 1$ for all i. In terms of our example, this model would expect fertility to decline by the same amount in all countries, and would attribute all observed differences between countries to random variation.

At the other extreme we have a model where every unit has its own expected value μ_i. This model is called the *saturated* model because it has as many parameters in the linear predictor (or linear parameters, for short) as it has observations. The saturated model can be obtained as a special case of Equation 2.3 by setting $p = n$ and letting x_i take the value 1 for unit i and 0 otherwise. In this model the x's are indicator variables for the different units, and there is no random variation left. All observed differences between countries are attributed to their own idiosyncrasies.

Obviously the null and saturated models are not very useful by themselves. Most statistical models of interest lie somewhere in between, and most of this chapter will be devoted to an exploration of the middle ground. Our aim is to capture systematic sources of variation in the linear predictor, and let the error term account for unstructured or random variation.

2.2 Estimation of the Parameters

Consider for now a rather abstract model where $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ for some predictors $\mathbf{x}_i$. How do we estimate the parameters $\boldsymbol{\beta}$ and σ^2?

2.2.1 Estimation of β

The likelihood principle instructs us to pick the values of the parameters that maximize the likelihood, or equivalently, the logarithm of the likelihood function. If the observations are independent, then the likelihood function is a product of normal densities of the form given in Equation 2.1. Taking logarithms we obtain the normal log-likelihood

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum(y_i - \mu_i)^2/\sigma^2, \tag{2.5}$$

where $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$. The most important thing to notice about this expression is that maximizing the log-likelihood with respect to the linear parameters $\boldsymbol{\beta}$ for a fixed value of σ^2 is exactly equivalent to minimizing the sum of squared differences between observed and expected values, or residual sum of squares

$$\text{RSS}(\boldsymbol{\beta}) = \sum(y_i - \mu_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.6}$$

In other words, we need to pick values of $\boldsymbol{\beta}$ that make the fitted values $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ as close as possible to the observed values y_i.

Taking derivatives of the residual sum of squares with respect to $\boldsymbol{\beta}$ and setting the derivative equal to zero leads to the so-called *normal equations* for the maximum-likelihood estimator $\hat{\boldsymbol{\beta}}$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

If the model matrix $\mathbf{X}$ is of full column rank, so that no column is an exact linear combination of the others, then the matrix of cross-products $\mathbf{X}'\mathbf{X}$ is of full rank and can be inverted to solve the normal equations. This gives an explicit formula for the *ordinary least squares* (OLS) or maximum likelihood estimator of the linear parameters:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.7}$$

If $\mathbf{X}$ is not of full column rank one can use generalized inverses, but interpretation of the results is much more straightforward if one simply eliminates redundant columns. Most current statistical packages are smart enough to detect and omit redundancies automatically.

There are several numerical methods for solving the normal equations, including methods that operate on $\mathbf{X}'\mathbf{X}$, such as Gaussian elimination or the Choleski decomposition, and methods that attempt to simplify the calculations by factoring the model matrix $\mathbf{X}$, including Householder reflections, Givens rotations and the Gram-Schmidt orthogonalization. We will not discuss these methods here, assuming that you will trust the calculations to a reliable statistical package. For further details see McCullagh and Nelder (1989, Section 3.8) and the references therein.

The foregoing results were obtained by maximizing the log-likelihood with respect to $\boldsymbol{\beta}$ for a fixed value of σ^2. The result obtained in Equation 2.7 does not depend on σ^2, and is therefore a global maximum.

For the *null* model $\mathbf{X}$ is a vector of ones, $\mathbf{X}'\mathbf{X} = n$ and $\mathbf{X}'\mathbf{y} = \sum \mathbf{y}_i$ are scalars and $\hat{\beta} = \bar{y}$, the sample mean. For our sample data $\bar{y} = 14.3$. Thus, the calculation of a sample mean can be viewed as the simplest case of maximum likelihood estimation in a linear model.

2.2.2 Properties of the Estimator

The least squares estimator $\hat{\boldsymbol{\beta}}$ of Equation 2.7 has several interesting properties. If the model is correct, in the (weak) sense that the expected value of the response Y_i given the predictors $\mathbf{x}_i$ is indeed $\mathbf{x}_i'\boldsymbol{\beta}$, then the OLS estimator is *unbiased*, its expected value equals the true parameter value:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{2.8}$$

It can also be shown that if the observations are uncorrelated and have constant variance σ^2, then the variance-covariance matrix of the OLS estimator is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2. \tag{2.9}$$

This result follows immediately from the fact that $\hat{\boldsymbol{\beta}}$ is a linear function of the data $\mathbf{y}$ (see Equation 2.7), and the assumption that the variance-covariance matrix of the data is $\mathrm{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

A further property of the estimator is that it has minimum variance among all unbiased estimators that are linear functions of the data, i.e.

it is the best linear unbiased estimator (BLUE). Since no other unbiased estimator can have lower variance for a fixed sample size, we say that OLS estimators are fully *efficient*.

Finally, it can be shown that the sampling distribution of the OLS estimator $\hat{\boldsymbol{\beta}}$ in large samples is approximately multivariate normal with the mean and variance given above, i.e.

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2).$$

Applying these results to the *null* model we see that the sample mean $\bar{y}$ is an unbiased estimator of μ, has variance σ^2/n, and is approximately normally distributed in large samples.

All of these results depend only on second-order assumptions concerning the mean, variance and covariance of the observations, namely the assumption that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Of course, $\hat{\boldsymbol{\beta}}$ is also a maximum likelihood estimator under the assumption of normality of the observations. If $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ then the sampling distribution of $\hat{\boldsymbol{\beta}}$ is *exactly* multivariate normal with the indicated mean and variance.

The significance of these results cannot be overstated: the assumption of normality of the observations is required only for inference in small samples. The really important assumption is that the observations are uncorrelated and have constant variance, and this is sufficient for inference in large samples.

2.2.3 Estimation of σ^2

Substituting the OLS estimator of $\boldsymbol{\beta}$ into the log-likelihood in Equation 2.5 gives a profile likelihood for σ^2

$$\log L(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\text{RSS}(\hat{\boldsymbol{\beta}})/\sigma^2.$$

Differentiating this expression with respect to σ^2 (not σ) and setting the derivative to zero leads to the maximum likelihood estimator

$$\hat{\sigma^2} = \text{RSS}(\hat{\boldsymbol{\beta}})/n.$$

This estimator happens to be *biased*, but the bias is easily corrected dividing by $n-p$ instead of n. The situation is exactly analogous to the use of $n-1$ instead of n when estimating a variance. In fact, the estimator of σ^2 for

the *null* model is the sample variance, since $\hat{\beta} = \bar{y}$ and the residual sum of squares is RSS $= \sum(y_i - \bar{y})^2$.

Under the assumption of normality, the ratio RSS$/\sigma^2$ of the residual sum of squares to the true parameter value has a chi-squared distribution with $n - p$ degrees of freedom and is independent of the estimator of the linear parameters. You might be interested to know that using the chi-squared distribution as a likelihood to estimate σ^2 (instead of the normal likelihood to estimate both $\boldsymbol{\beta}$ and σ^2) leads to the unbiased estimator.

For the sample data the RSS for the null model is 2650.2 on 19 d.f. and therefore $\hat{\sigma} = 11.81$, the sample standard deviation.

2.3 Tests of Hypotheses

Consider testing hypotheses about the regression coefficients $\boldsymbol{\beta}$. Sometimes we will be interested in testing the significance of a single coefficient, say β_j, but on other occasions we will want to test the joint significance of several components of $\boldsymbol{\beta}$. In the next few sections we consider tests based on the sampling distribution of the maximum likelihood estimator and likelihood ratio tests.

2.3.1 Wald Tests

Consider first testing the significance of one particular coefficient, say

$$H_0 : \beta_j = 0.$$

The m.l.e. $\hat{\beta}_j$ has a distribution with mean 0 (under H_0) and variance given by the j-th diagonal element of the matrix in Equation 2.9. Thus, we can base our test on the ratio

$$t = \frac{\hat{\beta}_j}{\sqrt{\text{var}(\hat{\beta}_j)}}. \tag{2.10}$$

Note from Equation 2.9 that var$(\hat{\beta}_j)$ depends on σ^2, which is usually unknown. In practice we replace σ^2 by the unbiased estimate based on the residual sum of squares.

Under the assumption of normality of the data, the ratio of the coefficient to its standard error has under H_0 a *Student's t* distribution with $n - p$ degrees of freedom when σ^2 is estimated, and a standard normal distribution if σ^2 is known. This result provides a basis for exact inference in samples of any size.

Under the weaker second-order assumptions concerning the means, variances and covariances of the observations, the ratio has approximately in large samples a standard normal distribution. This result provides a basis for approximate inference in large samples.

Many analysts treat the ratio as a Student's t statistic regardless of the sample size. If normality is suspect one should not conduct the test unless the sample is large, in which case it really makes no difference which distribution is used. If the sample size is moderate, using the t test provides a more conservative procedure. (The Student's t distribution converges to a standard normal as the degrees of freedom increases to ∞. For example the 95% two-tailed critical value is 2.09 for 20 d.f., and 1.98 for 100 d.f., compared to the normal critical value of 1.96.)

The t test can also be used to construct a confidence interval for a coefficient. Specifically, we can state with $100(1-\alpha)\%$ confidence that β_j is between the bounds

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p}\sqrt{\text{var}(\hat{\beta}_j)}, \tag{2.11}$$

where $t_{1-\alpha/2,n-p}$ is the two-sided critical value of Student's t distribution with $n-p$ d.f. for a test of size α.

The Wald test can also be used to test the joint significance of several coefficients. Let us partition the vector of coefficients into two components, say $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ with p_1 and p_2 elements, respectively, and consider the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}.$$

In this case the Wald statistic is given by the quadratic form

$$W = \hat{\boldsymbol{\beta}}_2' \,\text{var}^{-1}(\hat{\boldsymbol{\beta}}_2)\, \hat{\boldsymbol{\beta}}_2,$$

where $\hat{\boldsymbol{\beta}}_2$ is the m.l.e. of $\boldsymbol{\beta}_2$ and $\text{var}(\hat{\boldsymbol{\beta}}_2)$ is its variance-covariance matrix. Note that the variance depends on σ^2 which is usually unknown; in practice we substitute the estimate based on the residual sum of squares.

In the case of a single coefficient $p_2 = 1$ and this formula reduces to the square of the t statistic in Equation 2.10.

Asymptotic theory tells us that under H_0 the large-sample distribution of the m.l.e. is multivariate normal with mean vector $\mathbf{0}$ and variance-covariance matrix $\text{var}(\boldsymbol{\beta}_2)$. Consequently, the large-sample distribution of the quadratic form W is chi-squared with p_2 degrees of freedom. This result holds whether σ^2 is known or estimated.

Under the assumption of normality we have a stronger result. The distribution of W is exactly chi-squared with p_2 degrees of freedom if σ^2 is

known. In the more general case where σ^2 is estimated using a residual sum of squares based on $n-p$ d.f., the distribution of W/p_2 is an F with p_2 and $n-p$ d.f.

Note that as n approaches infinity for fixed p (so $n-p$ approaches infinity), the F distribution times p_2 approaches a chi-squared distribution with p_2 degrees of freedom. Thus, in large samples it makes no difference whether one treats W as chi-squared or W/p_2 as an F statistic. Many analysts treat W/p_2 as F for all sample sizes.

The situation is exactly analogous to the choice between the normal and Student's t distributions in the case of one variable. In fact, a chi-squared with one degree of freedom is the square of a standard normal, and an F with one and v degrees of freedom is the square of a Student's t with v degrees of freedom.

2.3.2 The Likelihood Ratio Test

Consider again testing the joint significance of several coefficients, say

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

as in the previous subsection. Note that we can partition the model matrix into two components $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with p_1 and p_2 predictors, respectively. The hypothesis of interest states that the response does not depend on the last p_2 predictors.

We now build a likelihood ratio test for this hypothesis. The general theory directs us to (1) fit two nested models: a smaller model with the first p_1 predictors in $\mathbf{X}_1$, and a larger model with all p predictors in $\mathbf{X}$; and (2) compare their maximized likelihoods (or log-likelihoods).

Suppose then that we fit the smaller model with the predictors in $\mathbf{X}_1$ only. We proceed by maximizing the log-likelihood of Equation 2.5 for a fixed value of σ^2. The maximized log-likelihood is

$$\max \log \mathrm{L}(\boldsymbol{\beta}_1) = c - \frac{1}{2}\mathrm{RSS}(\mathbf{X}_1)/\sigma^2,$$

where $c = -(n/2)\log(2\pi\sigma^2)$ is a constant depending on π and σ^2 but not on the parameters of interest. In a slight abuse of notation, we have written $\mathrm{RSS}(\mathbf{X}_1)$ for the residual sum of squares after fitting $\mathbf{X}_1$, which is of course a function of the estimate $\hat{\boldsymbol{\beta}}_1$.

Consider now fitting the larger model $X_1 + X_2$ with all predictors. The maximized log-likelihood for a fixed value of σ^2 is

$$\max \log \mathrm{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = c - \frac{1}{2}\mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/\sigma^2,$$

where $\text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$ is the residual sum of squares after fitting $\mathbf{X}_1$ and $\mathbf{X}_2$, itself a function of the estimate $\hat{\boldsymbol{\beta}}$.

To compare these log-likelihoods we calculate minus twice their difference. The constants cancel out and we obtain the likelihood ratio criterion

$$-2\log\lambda = \frac{\text{RSS}(\mathbf{X}_1) - \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)}{\sigma^2}. \tag{2.12}$$

There are two things to note about this criterion. First, we are directed to look at the reduction in the residual sum of squares when we add the predictors in $\mathbf{X}_2$. Basically, these variables are deemed to have a significant effect on the response if including them in the model results in a reduction in the residual sum of squares. Second, the reduction is compared to σ^2, the error variance, which provides a unit of comparison.

To determine if the reduction (in units of σ^2) exceeds what could be expected by chance alone, we compare the criterion to its sampling distribution. Large sample theory tells us that the distribution of the criterion converges to a chi-squared with p_2 d.f. The expected value of a chi-squared distribution with ν degrees of freedom is ν (and the variance is 2ν). Thus, chance alone would lead us to expect a reduction in the RSS of about one σ^2 for each variable added to the model. To conclude that the reduction exceeds what would be expected by chance alone, we usually require an improvement that exceeds the 95-th percentile of the reference distribution.

One slight difficulty with the development so far is that the criterion depends on σ^2, which is not known. In practice, we substitute an estimate of σ^2 based on the residual sum of squares of the *larger* model. Thus, we calculate the criterion in Equation 2.12 using

$$\hat{\sigma^2} = \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/(n-p).$$

The large-sample distribution of the criterion continues to be chi-squared with p_2 degrees of freedom, even if σ^2 has been estimated.

Under the assumption of normality, however, we have a stronger result. The likelihood ratio criterion $-2\log\lambda$ has an *exact* chi-squared distribution with p_2 d.f. if σ^2 is know. In the usual case where σ^2 is estimated, the criterion divided by p_2, namely

$$F = \frac{(\text{RSS}(\mathbf{X}_1) - \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2))/p_2}{\text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/(n-p)}, \tag{2.13}$$

has an exact F distribution with p_2 and $n-p$ d.f.

The numerator of F is the reduction in the residual sum of squares per degree of freedom spent. The denominator is the average residual sum of

squares, a measure of noise in the model. Thus, an F-ratio of one would indicate that the variables in $\mathbf{X}_2$ are just adding noise. A ratio in excess of one would be indicative of signal. We usually reject H_0, and conclude that the variables in $\mathbf{X}_2$ have an effect on the response if the F criterion exceeds the 95-th percentage point of the F distribution with p_2 and $n - p$ degrees of freedom.

A Technical Note: In this section we have built the likelihood ratio test for the linear parameters $\boldsymbol{\beta}$ by treating σ^2 as a nuisance parameter. In other words, we have maximized the log-likelihood with respect to $\boldsymbol{\beta}$ for fixed values of σ^2. You may feel reassured to know that if we had maximized the log-likelihood with respect to both $\boldsymbol{\beta}$ and σ^2 we would have ended up with an equivalent criterion based on a comparison of the *logarithms* of the residual sums of squares of the two models of interest. The approach adopted here leads more directly to the distributional results of interest and is typical of the treatment of scale parameters in generalized linear models.□

2.3.3 Student's t, F and the Anova Table

You may be wondering at this point whether you should use the Wald test, based on the large-sample distribution of the m.l.e., or the likelihood ratio test, based on a comparison of maximized likelihoods (or log-likelihoods). The answer in general is that in large samples the choice does not matter because the two types of tests are asymptotically equivalent.

In linear models, however, we have a much stronger result: the two tests are *identical*. The proof is beyond the scope of these notes, but we will verify it in the context of specific applications. The result is unique to linear models. When we consider logistic or Poisson regression models later in the sequel we will find that the Wald and likelihood ratio tests differ.

At least for linear models, however, we can offer some simple practical advice:

- To test hypotheses about a single coefficient, use the t-test based on the estimator and its standard error, as given in Equation 2.10.

- To test hypotheses about several coefficients, or more generally to compare nested models, use the F-test based on a comparison of RSS's, as given in Equation 2.13.

The calculations leading to an F-test are often set out in an analysis of variance (anova) table, showing how the total sum of squares (the RSS of the null model) can be partitioned into a sum of squares associated with $\mathbf{X}_1$,

a sum of squares *added by* $\mathbf{X}_2$, and a residual sum of squares. The table also shows the degrees of freedom associated with each sum of squares, and the mean square, or ratio of the sum of squares to its d.f.

Table 2.2 shows the usual format. We use ϕ to denote the null model. We also assume that one of the columns of $\mathbf{X}_1$ was the constant, so this block adds only $p_1 - 1$ variables to the null model.

TABLE 2.2: The Hierarchical Anova Table

Source of variation	Sum of squares	Degrees of freedom
$\mathbf{X}_1$	$\text{RSS}(\phi) - \text{RSS}(\mathbf{X}_1)$	$p_1 - 1$
$\mathbf{X}_2$ given $\mathbf{X}_1$	$\text{RSS}(\mathbf{X}_1) - \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$	p_2
Residual	$\text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$	$n - p$
Total	$\text{RSS}(\phi)$	$n - 1$

Sometimes the component associated with the constant is shown explicitly and the bottom line becomes the total (also called 'uncorrected') sum of squares: $\sum y_i^2$. More detailed analysis of variance tables may be obtained by introducing the predictors one at a time, while keeping track of the reduction in residual sum of squares at each step.

Rather than give specific formulas for these cases, we stress here that *all* anova tables can be obtained by calculating differences in RSS's and differences in the number of parameters between nested models. Many examples will be given in the applications that follow. A few descriptive measures of interest, such as simple, partial and multiple correlation coefficients, turn out to be simple functions of these sums of squares, and will be introduced in the context of the applications.

An important point to note before we leave the subject is that the order in which the variables are entered in the anova table (reflecting the order in which they are added to the model) is extremely important. In Table 2.2, we show the effect of adding the predictors in $\mathbf{X}_2$ to a model that already has $\mathbf{X}_1$. This *net* effect of X_2 after allowing for X_1 can be quite different from the *gross* effect of X_2 when considered by itself. The distinction is important and will be stressed in the context of the applications that follow.

2.4 Simple Linear Regression

Let us now turn to applications, modelling the dependence of a continuous response y on a single linear predictor x. In terms of our example, we will study fertility decline as a function of social setting. One can often obtain useful insight into the form of this dependence by plotting the data, as we did in Figure 2.1.

2.4.1 The Regression Model

We start by recognizing that the response will vary even for constant values of the predictor, and model this fact by treating the responses y_i as realizations of random variables

$$Y_i \sim N(\mu_i, \sigma^2) \tag{2.14}$$

with means μ_i depending on the values of the predictor x_i and constant variance σ^2.

The simplest way to express the dependence of the expected response μ_i on the predictor x_i is to assume that it is a linear function, say

$$\mu_i = \alpha + \beta x_i. \tag{2.15}$$

This equation defines a straight line. The parameter α is called the *constant* or intercept, and represents the expected response when $x_i = 0$. (This quantity may not be of direct interest if zero is not in the range of the data.) The parameter β is called the *slope*, and represents the expected increment in the response per unit change in x_i.

You probably have seen the simple linear regression model written with an explicit error term as

$$Y_i = \alpha + \beta x_i + \epsilon_i.$$

Did I forget the error term? Not really. Equation 2.14 defines the random structure of the model, and is equivalent to saying that $Y_i = \mu_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Equation 2.15 defines the systematic structure of the model, stipulating that $\mu_i = \alpha + \beta x_i$. Combining these two statements yields the traditional formulation of the model. Our approach separates more clearly the systematic and random components, and extends more easily to generalized linear models by focusing on the distribution of the response rather than the distribution of the error term.

2.4.2 Estimates and Standard Errors

The simple linear regression model can be obtained as a special case of the general linear model of Section 2.1 by letting the model matrix $\mathbf{X}$ consist of two columns: a column of ones representing the constant and a column with the values of x representing the predictor. Estimates of the parameters, standard errors, and tests of hypotheses can then be obtained from the general results of Sections 2.2 and 2.3.

It may be of interest to note that in simple linear regression the estimates of the constant and slope are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}.$$

The first equation shows that the fitted line goes through the means of the predictor and the response, and the second shows that the estimated slope is simply the ratio of the covariance of x and y to the variance of x.

Fitting this model to the family planning effort data with CBR decline as the response and the index of social setting as a predictor gives a residual sum of squares of 1449.1 on 18 d.f. (20 observations minus two parameters: the constant and slope).

Table 2.3 shows the estimates of the parameters, their standard errors and the corresponding t-ratios.

TABLE 2.3: Estimates for Simple Linear Regression of CBR Decline on Social Setting Score

Parameter	Symbol	Estimate	Std.Error	t-ratio
Constant	α	-22.13	9.642	-2.29
Slope	β	0.5052	0.1308	3.86

We find that, on the average, each additional point in the social setting scale is associated with an additional half a percentage point of CBR decline, measured from a baseline of an expected 22% *increase* in CBR when social setting is zero. (Since the social setting scores range from 35 to 91, the constant is not particularly meaningful in this example.)

The estimated standard error of the slope is 0.13, and the corresponding t-test of 3.86 on 18 d.f. is highly significant. With 95% confidence we estimate that the slope lies between 0.23 and 0.78.

Figure 2.3 shows the results in graphical form, plotting observed and fitted values of CBR decline versus social setting. The fitted values are

calculated for any values of the predictor x as $\hat{y} = \hat{\alpha} + \hat{\beta}x$ and lie, of course, in a straight line.

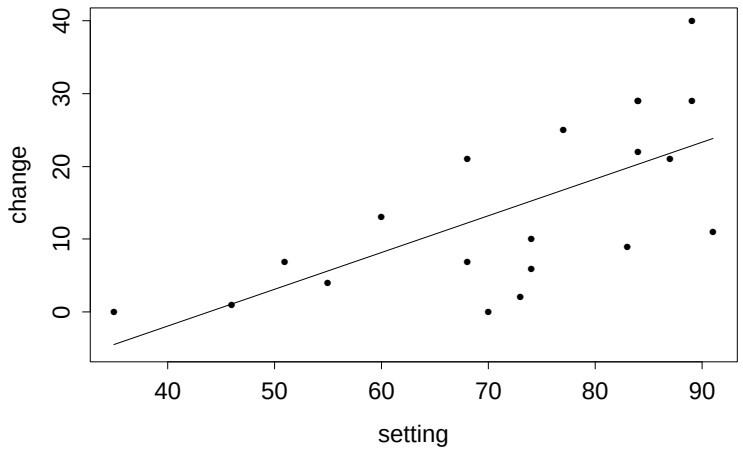

FIGURE 2.3: Linear Regression of CBR Decline on Social Setting

You should verify that the analogous model with family planning effort as a single predictor gives a residual sum of squares of 950.6 on 18 d.f., with constant 2.336 ($\pm$2.662) and slope 1.253 ($\pm$0.2208). Make sure you know how to interpret these estimates.

2.4.3 Anova for Simple Regression

Instead of using a test based on the distribution of the OLS estimator, we could test the significance of the slope by comparing the simple linear regression model with the null model. Note that these models are nested, because we can obtain the null model by setting $\beta = 0$ in the simple linear regression model.

Fitting the null model to the family planning data gives a residual sum of squares of 2650.2 on 19 d.f. Adding a linear effect of social setting reduces the RSS by 1201.1 at the expense of one d.f. This gain can be contrasted with the remaining RSS of 1449.1 on 18 d.f. by constructing an F-test. The calculations are set out in Table 2.4, and lead to an F-statistic of 14.9 on one and 18 d.f.

These results can be used to verify the equivalence of t and F test statistics and critical values. Squaring the observed t-statistic of 3.86 gives the observed F-ratio of 14.9. Squaring the 95% two-sided critical value of the

TABLE 2.4: Analysis of Variance for Simple Regression of CBR Decline on Social Setting Score

Source of variation	Degrees of freedom	Sum of squares	Mean squared	F-ratio
Setting	1	1201.1	1201.1	14.9
Residual	18	1449.1	80.5	
Total	19	2650.2		

Student's t distribution with 18 d.f., which is 2.1, gives the 95% critical value of the F distribution with one and 18 d.f., which is 4.4.

You should verify that the t and F tests for the model with a linear effect of family planning effort are $t = 5.67$ and $F = 32.2$.

2.4.4 Pearson's Correlation Coefficient

A simple summary of the strength of the relationship between the predictor and the response can be obtained by calculating a proportionate reduction in the residual sum of squares as we move from the null model to the model with x. The quantity

$$R^2 = 1 - \frac{\text{RSS}(x)}{\text{RSS}(\phi)}$$

is know as the *coefficient of determination*, and is often described as the proportion of 'variance' explained by the model. (The description is not very accurate because the calculation is based on the RSS not the variance, but it is too well entrenched to attempt to change it.) In our example the RSS was 2650.2 for the null model and 1449.1 for the model with setting, so we have 'explained' 1201.1 points or 45.3% as a linear effect of social setting.

The square root of the proportion of variance explained in a simple linear regression model, with the same sign as the regression coefficient, is *Pearson's linear correlation coefficient*. This measure ranges between -1 and 1, taking these values for perfect inverse and direct relationships, respectively. For the model with CBR decline as a linear function of social setting, Pearson's $r = 0.673$. This coefficient can be calculated directly from the covariance of x and y and their variances, as

$$r = \frac{\sum(y - \bar{y})(x - \bar{x})}{\sqrt{\sum(y - \bar{y})^2 \sum(x - \bar{x})^2}}.$$

There is one additional characterization of Pearson's r that may help in interpretation. Suppose you standardize y by subtracting its mean and dividing by its standard deviation, standardize x in the same fashion, and then regress the standardized y on the standardized x forcing the regression through the origin (i.e. omitting the constant). The resulting estimate of the regression coefficient is Pearson's r. Thus, we can interpret r as the expected change in the response in units of standard deviation associated with a change of one standard deviation in the predictor.

In our example, each standard deviation of increase in social setting is associated with an additional decline in the CBR of 0.673 standard deviations. While the regression coefficient expresses the association in the original units of x and y, Pearson's r expresses the association in units of standard deviation.

You should verify that a linear effect of family planning effort accounts for 64.1% of the variation in CBR decline, so Pearson's $r = 0.801$. Clearly CBR decline is associated more strongly with family planning effort than with social setting.

2.5 Multiple Linear Regression

Let us now study the dependence of a continuous response on two (or more) linear predictors. Returning to our example, we will study fertility decline as a function of both social setting and family planning effort.

2.5.1 The Additive Model

Suppose then that we have a response y and two predictors x_1 and x_2. We will use y_i to denote the value of the response and x_{i1} and x_{i2} to denote the values of the predictors for the i-th unit, where $i = 1, \ldots, n$.

We maintain the assumptions regarding the stochastic component of the model, so y_i is viewed as a realization of $Y_i \sim N(\mu_i, \sigma^2)$, but change the structure of the systematic component. We now assume that the expected response μ_i is a linear function of the two predictors, that is

$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{2.16}$$

This equation defines a plane in three dimensional space (you may want to peek at Figure 2.4 for an example). The parameter α is the constant, representing the expected response when both x_{i1} and x_{i2} are zero. (As before, this value may not be directly interpretable if zero is not in the

range of the predictors.) The parameter β_1 is the slope along the x_1-axis and represents the expected change in the response per unit change in x_1 at constant values of x_2. Similarly, β_2 is the slope along the x_2 axis and represents the expected change in the response per unit change in x_2 while holding x_1 constant.

It is important to note that these interpretations represent abstractions based on the model that we may be unable to observe in the real world. In terms of our example, changes in family planning effort are likely to occur in conjunction with, if not directly as a result of, improvements in social setting. The model, however, provides a useful representation of the data and hopefully approximates the results of comparing countries that differ in family planning effort but have similar socio-economic conditions.

A second important feature of the model is that it is *additive*, in the sense that the effect of each predictor on the response is assumed to be the same for all values of the other predictor. In terms of our example, the model assumes that the effect of family planning effort is exactly the same at every social setting. This assumption may be unrealistic, and later in this section we will introduce a model where the effect of family planning effort is allowed to depend on social setting.

2.5.2 Estimates and Standard Errors

The multiple regression model in 2.16 can be obtained as a special case of the general linear model of Section 2.1 by letting the model matrix $\mathbf{X}$ consist of three columns: a column of ones representing the constant, a column representing the values of x_1, and a column representing the values of x_2. Estimates, standard errors and tests of hypotheses then follow from the general results in Sections 2.2 and 2.3.

Fitting the two-predictor model to our example, with CBR decline as the response and the indices of family planning effort and social setting as linear predictors, gives a residual sum of squares of 694.0 on 17 d.f. (20 observations minus three parameters: the constant and two slopes). Table 2.5 shows the parameter estimates, standard errors and t-ratios.

We find that, on average, the CBR declines an additional 0.27 percentage points for each additional point of improvement in social setting at constant levels of family planning effort. The standard error of this coefficient is 0.11. Since the t ratio exceeds 2.11, the five percent critical value of the t distribution with 17 d.f., we conclude that we have evidence of association between social setting and CBR decline net of family planning effort. A 95% confidence interval for the social setting slope, based on Student's t

TABLE 2.5: Estimates for Multiple Linear Regression of CBR Decline on Social Setting and Family Planning Effort Scores

Parameter	Symbol	Estimate	Std.Error	t-ratio
Constant	α	-14.45	7.094	-2.04
Setting	β_1	0.2706	0.1079	2.51
Effort	β_2	0.9677	0.2250	4.30

distribution with 17 d.f., has bounds 0.04 and 0.50.

Similarly, we find that on average the CBR declines an additional 0.97 percentage points for each additional point of family planning effort at constant social setting. The estimated standard error of this coefficient is 0.23. Since the coefficient is more than four times its standard error, we conclude that there is a significant linear association between family planning effort and CBR decline at any given level of social setting. With 95% confidence we conclude that the additional percent decline in the CBR per extra point of family planning effort lies between 0.49 and 1.44.

The constant is of no direct interest in this example because zero is not in the range of the data; while some countries have a value of zero for the index of family planning effort, the index of social setting ranges from 35 for Haiti to 91 for Venezuela.

The estimate of the residual standard deviation in our example is $\hat{\sigma} = 6.389$. This value, which is rarely reported, provides a measure of the extent to which countries with the same setting and level of effort can experience different declines in the CBR.

Figure 2.4 shows the estimated regression equation $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ evaluated for a grid of values of the two predictors. The grid is confined to the range of the data on setting and effort. The regression plane may be viewed as an infinite set of regression lines. For any fixed value of setting, expected CBR decline is a linear function of effort with slope 0.97. For any fixed value of effort, expected CBR decline is a linear function of setting with slope 0.27.

2.5.3 Gross and Net Effects

It may be instructive to compare the results of the multiple regression analysis, which considered the two predictors simultaneously, with the results of the simple linear regression analyses, which considered the predictors one at a time.

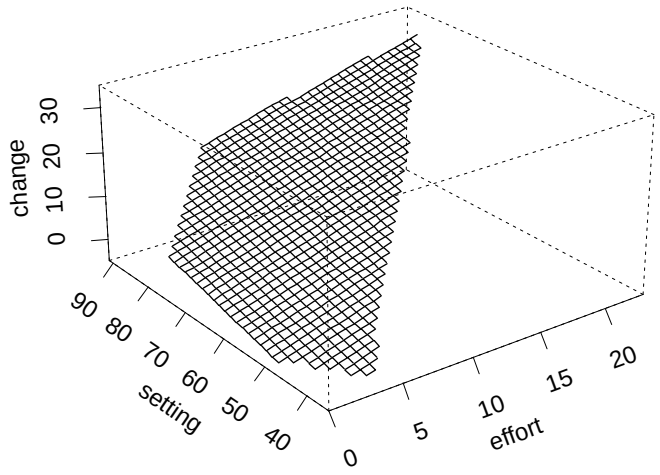

FIGURE 2.4: Multiple Regression of CBR Decline on Social Setting and Family Planning Effort

The coefficients in a simple linear regression represent changes in the response that can be associated with a given predictor, and will be called *gross* effects. In our simple linear regression analysis of CBR decline as a function of family planning effort we found that, on the average, each additional point of family planning effort was associated with an additional 1.25 percentage point of CBR decline. Interpretation of gross effects must be cautious because comparisons involving one factor include, implicitly, other measured and unmeasured factors. In our example, when we compare countries with strong programs with countries with weak programs, we are also comparing implicitly countries with high and low social settings.

The coefficients in a multiple linear regression are more interesting because they represent changes in the response that can be associated with a given predictor for fixed values of other predictors, and will be called *net* effects. In our multiple regression analysis of CBR decline as a function of both family planning effort and social setting, we found that, on the average, each additional point of family planning effort was associated with an additional 0.97 percentage points of CBR decline if we held social setting constant, i.e. if we compared countries with the same social setting. Interpretation of this coefficient as measuring the effect of family planning effort is on somewhat firmer ground than for the gross effect, because the differences have been adjusted for social setting. Caution is in order, however, because there are bound to be other confounding factors that we have not taken into account.

In my view, the closest approximation we have to a true causal effect in social research based on observational data is a net effect in a multiple regression analysis that has controlled for all relevant factors, an ideal that may be approached but probably can never be attained. The alternative is a controlled experiment where units are assigned at random to various treatments, because the nature of the assignment itself guarantees that any ensuing differences, beyond those than can be attributed to chance, must be due to the treatment. In terms of our example, we are unable to randomize the allocation of countries to strong and weak programs. But we can use multiple regression as a tool to adjust the estimated effects for the confounding effects of observed covariates.

TABLE 2.6: Gross and Net Effects of Social Setting and Family Planning Effort on CBR Decline

Predictor	Effect	
	Gross	Net
Setting	0.505	0.271
Effort	1.253	0.968

Gross and net effects may be presented in tabular form as shown in Table 2.6. In our example, the gross effect of family planning effort of 1.25 was reduced to 0.97 after adjustment for social setting, because part of the observed differences between countries with strong and weak programs could be attributed to the fact that the former tend to enjoy higher living standards. Similarly, the gross effect of social setting of 0.51 has been reduced to 0.27 after controlling for family planning effort, because part of the differences between richer and poorer countries could be attributed to the fact that the former tend to have stronger family planning programs.

Note, incidentally, that it is not reasonable to compare either gross or net effects across predictors, because the regression coefficients depend on the units of measurement. I could easily 'increase' the gross effect of family planning effort to 12.5 simply by dividing the scores by ten. One way to circumvent this problem is to standardize the response and all predictors, subtracting their means and dividing by their standard deviations. The regression coefficients for the standardized model (which are sometimes called 'beta' coefficients) are more directly comparable. This solution is particularly appealing when the variables do not have a natural unit of measurement, as is often the case for psychological test scores. On the other hand,

standardized coefficients are heavily dependent on the range of the data; they should not be used, for example, if one has sampled high and low values of one predictor to increase efficiency, because that design would inflate the variance of the predictor and therefore reduce the standardized coefficient.

2.5.4 Anova for Multiple Regression

The basic principles of model comparison outlined earlier may be applied to multiple regression models. I will illustrate the procedures by considering a test for the significance of the entire regression, and a test for the significance of the net effect of one predictor after adjusting for the other.

Consider first the hypothesis that all coefficients other than the constant are zero, i.e.

$$H_0 : \beta_1 = \beta_2 = 0.$$

To test the significance of the entire regression we start with the null model, which had a RSS of 2650.2 on 19 degrees of freedom. Adding the two linear predictors, social setting and family planning effort, reduces the RSS by 1956.2 at the expense of two d.f. Comparing this gain with the remaining RSS of 694.0 on 17 d.f. leads to an F-test of 24.0 on two and 17 d.f. This statistic is highly significant, with a P-value just above 0.00001. Thus, we have clear evidence that CBR decline is associated with social setting and family planning effort. Details of these calculations are shown in Table 2.7

TABLE 2.7: Analysis of Variance for Multiple Regression of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Regression	1956.2	2	978.1	24.0
Residual	694.0	17	40.8	
Total	2650.2	19		

In the above comparison we proceeded directly from the null model to the model with two predictors. A more detailed analysis is possible by adding the predictors one at a time. Recall from Section 2.4 that the model with social setting alone had a RSS of 1449.1 on 18 d.f., which represents a gain of 1201.1 over the null model. In turn, the multiple regression model with both social setting and family planning effort had a RSS of 694.0 on 17 d.f. which represents a gain of 755.1 over the model with social setting alone. These calculation are set out in the *hierarchical* anova shown in Table 2.8.

TABLE 2.8: Hierarchical Analysis of Variance for Multiple Regression of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting	1201.1	1	1201.1	29.4
Effort|Setting	755.1	1	755.1	18.5
Residual	694.0	17	40.8	
Total	2650.2	19		

Note the following features of this table. First, adding the sums of squares and d.f.'s in the first two rows agrees with the results in the previous table; thus, we have further decomposed the sum of squares associated with the regression into a term attributed to social setting and a term added by family planning effort.

Second, the notation Effort|Setting emphasizes that we have considered first the contribution of setting and then the additional contribution of effort once setting is accounted for. The order we used seemed more natural for the problem at hand. An alternative decomposition would introduce effort first and then social setting. The corresponding hierarchical anova table is left as an exercise.

Third, the F-test for the additional contribution of family planning effort over and above social setting (which is $F = 18.5$ from Table 2.8) coincides with the test for the coefficient of effort based on the estimate and its standard error (which is $t = 4.3$ from Table 2.5), since $4.3^2 = 18.5$. In both cases we are testing the hypothesis

$$H_0 : \beta_2 = 0$$

that the *net* effect of effort given setting is zero. Keep in mind that dividing estimates by standard errors tests the hypothesis that the variable in question has no effect *after* adjusting for all other variables. It is perfectly possible to find that two predictors are jointly significant while neither exceeds twice its standard error. This occurs when the predictors are highly correlated and either could account for (most of) the effects of the other.

2.5.5 Partial and Multiple Correlations

A descriptive measure of how much we have advanced in our understanding of the response is given by the proportion of variance explained, which was

first introduced in Section 2.4. In our case the two predictors have reduced the RSS from 2650.2 to 694.0, explaining 73.8%.

The square root of the proportion of variance explained is the *multiple correlation coefficient*, and measures the linear correlation between the response in one hand and all the predictors on the other. In our case $R = 0.859$. This value can also be calculated directly as Pearson's linear correlation between the response y and the fitted values $\hat{y}$.

An alternative construction of R is of some interest. Suppose we want to measure the correlation between a single variable y and a set of variables (a vector) $\mathbf{x}$. One approach reduces the problem to calculating Pearson's r between two single variables, y and a linear combination $z = \mathbf{c}'\mathbf{x}$ of the variables in $\mathbf{x}$, and then taking the maximum over all possible vectors of coefficients $\mathbf{c}$. Amazingly, the resulting maximum is R and the coefficients $\mathbf{c}$ are proportional to the estimated regression coefficients.

We can also calculate proportions of variance explained based on the hierarchical anova tables. Looking at Table 2.8, we note that setting explained 1201.1 of the total 2650.2, or 45.3%, while effort explained 755.1 of the same 2650.2, or 28.5%, for a total of 1956.2 out of 2650.2, or 73.8%. In a sense this calculation is not fair because setting is introduced before effort. An alternative calculation may focus on how much the second variable explains not out of the total, but out of the variation left unexplained by the first variable. In this light, effort explained 755.1 of the 1449.1 left unexplained by social setting, or 52.1%.

The square root of the proportion of variation explained by the second variable out of the amount left unexplained by the first is called the *partial correlation coefficient*, and measures the linear correlation between y and x_2 after adjusting for x_1. In our example, the linear correlation between CBR decline and effort after controlling for setting is 0.722.

The following calculation may be useful in interpreting this coefficient. First regress y on x_1 and calculate the residuals, or differences between observed and fitted values. Then regress x_2 on x_1 and calculate the residuals. Finally, calculate Pearson's r between the two sets of residuals. The result is the partial correlation coefficient, which can thus be seen to measure the simple linear correlation between y and x_2 after removing the linear effects of x_1.

Partial correlation coefficients involving three variables can be calculated directly from the pairwise simple correlations. Let us index the response y as variable 0 and the predictors x_1 and x_2 as variables 1 and 2. Then the

partial correlation between variables 0 and 2 adjusting for 1 is

$$r_{02.1} = \frac{r_{02} - r_{01}r_{12}}{\sqrt{1 - r_{01}^2}\sqrt{1 - r_{12}^2}},$$

where r_{ij} denotes Pearson's linear correlation between variables i and j. The formulation given above is more general, because it can be used to compute the partial correlation between two variables (the response and one predictor) adjusting for any number of additional variables.

TABLE 2.9: Simple and Partial Correlations of CBR Decline with Social Setting and Family Planning Effort

Predictor	Correlation	
	Simple	Partial
Setting	0.673	0.519
Effort	0.801	0.722

Simple and partial correlation coefficients can be compared in much the same vein as we compared gross and net effects earlier. Table 2.9 summarizes the simple and partial correlations between CBR decline on the one hand and social setting and family planning effort on the other. Note that the effect of effort is more pronounced and more resilient to adjustment than the effect of setting.

2.5.6 More Complicated Models

So far we have considered four models for the family planning effort data: the null model (ϕ), the one-variate models involving either setting (x_1) or effort (x_2), and the additive model involving setting and effort ($x_1 + x_2$).

More complicated models may be obtained by considering higher order polynomial terms in either variable. Thus, we might consider adding the squares x_1^2 or x_2^2 to capture *non-linearities* in the effects of setting or effort. The squared terms are often highly correlated with the original variables, and on certain datasets this may cause numerical problems in estimation. A simple solution is to reduce the correlation by centering the variables before squaring them, using x_1 and $(x_1 - \bar{x}_1)^2$ instead of x_1 and x_1^2. The correlation can be eliminated entirely, often in the context of designed experiments, by using orthogonal polynomials.

We could also consider adding the cross-product term x_1x_2 to capture a form of *interaction* between setting and effort. In this model the linear predictor would be

$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}. \tag{2.17}$$

This is simply a linear model where the model matrix $\mathbf{X}$ has a column of ones for the constant, a column with the values of x_1, a column with the values of x_2, and a column with the products x_1x_2. This is equivalent to creating a new variable, say x_3, which happens to be the product of the other two.

An important feature of this model is that the effect of any given variable now depends on the value of the other. To see this point consider fixing x_1 and viewing the expected response μ as a function of x_2 for this fixed value of x_1. Rearranging terms in Equation 2.17 we find that μ is a linear function of x_2:

$$\mu_i = (\alpha + \beta_1 x_{i1}) + (\beta_2 + \beta_3 x_{i1}) x_{i2},$$

with both constant and slope depending on x_1. Specifically, the effect of x_2 on the response is itself a linear function of x_1; it starts from a baseline effect of β_2 when x_1 is zero, and has an additional effect of β_3 units for each unit increase in x_1.

The extensions considered here help emphasize a very important point about model building: the columns of the model matrix are not necessarily the predictors of interest, but can be any functions of them, including linear, quadratic or cross-product terms, or other transformations.

Are any of these refinements necessary for our example? To find out, fit the more elaborate models and see if you can obtain significant reductions of the residual sum of squares.

2.6 One-Way Analysis of Variance

We now consider models where the predictors are categorical variables or *factors* with a discrete number of levels. To illustrate the use of these models we will group the index of social setting (and later the index of family planning effort) into discrete categories.

2.6.1 The One-Way Layout

Table 2.10 shows the percent decline in the CBR for the 20 countries in our illustrative dataset, classified according to the index of social setting in three

categories: low (under 70 points), medium (70–79) and high (80 or more).

TABLE 2.10: CBR Decline by Levels of Social Setting

Setting	Percent decline in CBR
Low	1, 0, 7, 21, 13, 4, 7
Medium	10, 6, 2, 0, 25
High	9, 11, 29, 29, 40, 21, 22, 29

It will be convenient to modify our notation to reflect the one-way layout of the data explicitly. Let k denote the number of groups or levels of the factor, n_i denote the number of observations in group i, and let y_{ij} denote the response for the j-th unit in the i-th group, for $j = 1, \ldots, n_i$, and $i = 1, \ldots, k$. In our example $k = 3$ and y_{ij} is the CBR decline in the j-th country in the i-th category of social setting, with $i = 1, 2, 3; j = 1, \ldots, n_i; n_1 = 7, n_2 = 5$ and $n_3 = 8$).

2.6.2 The One-Factor Model

As usual, we treat y_{ij} as a realization of a random variable $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$, where the variance is the same for all observations. In terms of the systematic structure of the model, we assume that

$$\mu_{ij} = \mu + \alpha_i, \tag{2.18}$$

where μ plays the role of the constant and α_i represents the effect of level i of the factor.

Before we proceed further, it is important to note that the model as written is not identified. We have essentially k groups but have introduced $k+1$ linear parameters. The solution is to introduce a constraint, and there are several ways in which we could proceed.

One approach is to set $\mu = 0$ (or simply drop μ). If we do this, the α_i's become *cell means*, with α_i representing the expected response in group i. While simple and attractive, this approach does not generalize well to models with more than one factor.

Our preferred alternative is to set one of the α_i's to zero. Conventionally we set $\alpha_1 = 0$, but any of the groups could be chosen as the *reference cell* or level. In this approach μ becomes the expected response in the reference cell, and α_i becomes the effect of level i of the factor, compared to the reference level.

A third alternative is to require the group effects to add-up to zero, so $\sum \alpha_i = 0$. In this case μ represents some sort of overall expected response, and α_i measures the extent to which responses at level i of the factor deviate from the overall mean. Some statistics texts refer to this constraint as the 'usual' restrictions, but I think the reference cell method is now used more widely in social research.

A variant of the 'usual' restrictions is to require a weighted sum of the effects to add up to zero, so $\sum w_i\alpha_i = 0$. The weights are often taken to be the number of observations in each group, so $w_i = n_i$. In this case μ is a weighted average representing the expected response, and α_i is, as before, the extent to which responses at level i of the factor deviate from the overall mean.

Each of these parameterizations can easily be translated into one of the others, so the choice can rest on practical considerations. The reference cell method is easy to implement in a regression context and the resulting parameters have a clear interpretation.

2.6.3 Estimates and Standard Errors

The model in Equation 2.18 is a special case of the generalized linear model, where the design matrix $\mathbf{X}$ has $k+1$ columns: a column of ones representing the constant, and k columns of indicator variables, say $x_1, \ldots, x_k$, where x_i takes the value one for observations at level i of the factor and the value zero otherwise.

Note that the model matrix as defined so far is rank deficient, because the first column is the sum of the last k. Hence the need for constraints. The cell means approach is equivalent to dropping the constant, and the reference cell method is equivalent to dropping one of the indicator or dummy variables representing the levels of the factor. Both approaches are easily implemented. The other two approaches, which set to zero either the unweighted or weighted sum of the effects, are best implemented using Lagrange multipliers and will not be considered here.

Parameter estimates, standard errors and t ratios can then be obtained from the general results of Sections 2.2 and 2.3. You may be interested to know that the estimates of the regression coefficients in the one-way layout are simple functions of the cell means. Using the reference cell method,

$$\hat{\mu} = \bar{y}_1 \qquad \text{and} \qquad \hat{\alpha_i} = \bar{y}_i - \bar{y}_1 \text{ for } i > 1,$$

where $\bar{y}_i$ is the average of the responses at level i of the factor.

Table 2.11 shows the estimates for our sample data. We expect a CBR decline of almost 8% in countries with low social setting (the reference cell). Increasing social setting to medium or high is associated with additional declines of one and 16 percentage points, respectively, compared to low setting.

TABLE 2.11: Estimates for One-Way Anova Model of CBR Decline by Levels of Social Setting

Parameter	Symbol	Estimate	Std. Error	t-ratio
Low	μ	7.571	3.498	2.16
Medium (vs. low)	α_2	1.029	5.420	0.19
High (vs. low)	α_3	16.179	4.790	3.38

Looking at the t ratios we see that the difference between medium and low setting is not significant, so we accept $H_0 : \alpha_2 = 0$, whereas the difference between high and low setting, with a t-ratio of 3.38 on 17 d.f. and a two-sided P-value of 0.004, is highly significant, so we reject $H_0 : \alpha_3 = 0$. These t-ratios test the significance of two particular contrasts: medium vs. low and high vs. low. In the next subsection we consider an overall test of the significance of social setting.

2.6.4 The One-Way Anova Table

Fitting the model with social setting treated as a factor reduces the RSS from 2650.2 (for the null model) to 1456.4, a gain of 1193.8 at the expense of two degrees of freedom (the two α's). We can contrast this gain with the remaining RSS of 1456.4 on 17 d.f. The calculations are laid out in Table 2.12, and lead to an F-test of 6.97 on 2 and 17 d.f., which has a P-value of 0.006. We therefore reject the hypothesis $H_0 : \alpha_2 = \alpha_3 = 0$ of no setting effects, and conclude that the expected response depends on social setting.

TABLE 2.12: Analysis of Variance for One-Factor Model of CBR Decline by Levels of Social Setting

Source of variation	Sum of squares	Degrees of Freedom	Mean squared	F-ratio
Setting	1193.8	2	596.9	6.97
Residual	1456.4	17	85.7	
Total	2650.2	19		

Having established that social setting has an effect on CBR decline, we can inspect the parameter estimates and t-ratios to learn more about the nature of the effect. As noted earlier, the difference between high and low settings is significant, while that between medium and low is not.

It is instructive to calculate the Wald test for this example. Let $\boldsymbol{\alpha} = (\alpha_2, \alpha_3)'$ denote the two setting effects. The estimate and its variance-covariance matrix, calculated using the general results of Section 2.2, are

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} 1.029 \\ 16.179 \end{pmatrix} \qquad \text{and} \qquad \widehat{\text{var}}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} 29.373 & 12.239 \\ 12.239 & 22.948 \end{pmatrix}.$$

The Wald statistic is

$$W = \hat{\boldsymbol{\alpha}}' \, \widehat{\text{var}}^{-1}(\hat{\boldsymbol{\alpha}}) \, \hat{\boldsymbol{\alpha}} = 13.94,$$

and has approximately a chi-squared distribution with two d.f. Under the assumption of normality, however, we can divide by two to obtain $F = 6.97$, which has an F distribution with two and 17 d.f., and coincides with the test based on the reduction in the residual sum of squares, as shown in Table 2.12.

2.6.5 The Correlation Ratio

Note from Table 2.12 that the model treating social setting as a factor with three levels has reduced the RSS by 1456.6 out of 2650.2, thereby explaining 45.1%. The square root of the proportion of variance explained by a discrete factor is called the *correlation ratio*, and is often denoted η. In our example $\hat{\eta} = 0.672$.

If the factor has only two categories the resulting coefficient is called the *point-biserial correlation*, a measure often used in psychometrics to correlate a test score (a continuous variable) with the answer to a dichotomous item (correct or incorrect). Note that both measures are identical in construction to Pearson's correlation coefficient. The difference in terminology reflects whether the predictor is a continuous variable with a linear effect or a discrete variable with two or more than two categories.

2.7 Two-Way Analysis of Variance

We now consider models involving two factors with discrete levels. We illustrate using the sample data with both social setting and family planning effort grouped into categories. Key issues involve the concepts of main effects and interactions.

2.7.1 The Two-Way Layout

Table 2.13 shows the CBR decline in our 20 countries classified according to two criteria: social setting, with categories low (under 70), medium (70–79) and high (80+), and family planning effort, with categories weak (0–4), moderate (5–14) and strong (15+). In our example both setting and effort are factors with three levels. Note that there are no countries with strong programs in low social settings.

TABLE 2.13: CBR Decline by Levels of Social Setting and Levels of Family Planning Effort

Setting	Effort		
	Weak	Moderate	Strong
Low	1,0,7	21,13,4,7	–
Medium	10,6,2	0	25
High	9	11	29,29,40,21,22,29

We will modify our notation to reflect the two-way layout of the data. Let n_{ij} denote the number of observations in the (i, j)-th cell of the table, i.e. in row i and column j, and let y_{ijk} denote the response of the k-th unit in that cell, for $k = 1, \ldots, n_{ij}$. In our example y_{ijk} is the CBR decline of the k-th country in the i-th category of setting and the j-th category of effort.

2.7.2 The Two-Factor Additive Model

Once again, we treat the response as a realization of a random variable $Y_{ijk} \sim N(\mu_{ijk}, \sigma^2)$. In terms of the systematic component of the model, we will assume that

$$\mu_{ijk} = \mu + \alpha_i + \beta_j \tag{2.19}$$

In this formulation μ represents a baseline value, α_i represents the effect of the i-th level of the row factor and β_j represents the effect of the j-th level of the column factor. Before we proceed further we must note that the model is not identified as stated. You could add a constant δ to each of the α_i's (or to each of the β_j's) and subtract it from μ without altering any of the expected responses. Clearly we need two constraints to identify the model.

Our preferred approach relies on the reference cell method, and sets to zero the effects for the first cell (in the top-left corner of the table), so that $\alpha_1 = \beta_1 = 0$. The best way to understand the meaning of the remaining

parameters is to study Table 2.14, showing the expected response for each combination of levels of row and column factors having three levels each.

TABLE 2.14: The Two-Factor Additive Model

Row	Column		
	1	2	3
1	μ	$\mu+\beta_2$	$\mu+\beta_3$
2	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2$	$\mu+\alpha_2+\beta_3$
3	$\mu+\alpha_3$	$\mu+\alpha_3+\beta_2$	$\mu+\alpha_3+\beta_3$

In this formulation of the model μ represents the expected response in the reference cell, α_i represents the effect of level i of the row factor (compared to level 1) for any fixed level of the column factor, and β_j represents the effect of level j of the column factor (compared to level 1) for any fixed level of the row factor.

Note that the model is *additive*, in the sense that the effect of each factor is the same at all levels of the other factor. To see this point consider moving from the first to the second row. The response increases by α_2 if we move down the first column, but also if we move down the second or third columns.

2.7.3 Estimates and Standard Errors

The model in Equation 2.19 is a special case of the general linear model, where the model matrix $\mathbf{X}$ has a column of ones representing the constant, and two sets of dummy or indicator variables representing the levels of the row and column factors, respectively. This matrix is not of full column rank because the row (as well as the column) dummies add to the constant. Clearly we need two constraints and we choose to drop the dummy variables corresponding to the first row and to the first column. Table 2.15 shows the resulting parameter estimates, standard errors and t-ratios for our example.

Thus, we expect a CBR decline of 5.4% in countries with low setting and weak programs. In countries with medium or high social setting we expect CBR declines of 1.7 percentage points *less* and 2.4 percentage points more, respectively, than in countries with low setting and the same level of effort. Finally, in countries with moderate or strong programs we expect CBR declines of 3.8 and 20.7 percentage points more than in countries with weak programs and the same level of social setting.

It appears from a cursory examination of the t-ratios in Table 2.15 that the only significant effect is the difference between strong and weak pro-

TABLE 2.15: Parameter Estimates for Two-Factor Additive Model of CBR Decline by Social Setting and Family Planning Effort

Parameter		Symbol	Estimate	Std. Error	t-ratio
Baseline	low/weak	μ	5.379	3.105	1.73
Setting	medium	α_2	−1.681	3.855	−0.44
	high	α_3	2.388	4.457	0.54
Effort	moderate	β_2	3.836	3.575	1.07
	strong	β_3	20.672	4.339	4.76

grams. Bear in mind, however, that the table only shows the comparisons that are explicit in the chosen parameterization. In this example it turns out that the difference between strong and moderate programs is also significant. (This test can be calculated from the variance-covariance matrix of the estimates, or by fitting the model with strong programs as the reference cell, so the medium-strong comparison becomes one of the parameters.) Questions of significance for factors with more than two-levels are best addressed by using the F-test discussed below.

2.7.4 The Hierarchical Anova Table

Fitting the two-factor additive model results in a residual sum of squares of 574.4 on 15 d.f., and represents an improvement over the null model of 2075.8 at the expense of four d.f. We can further decompose this gain as an improvement of 1193.8 on 2 d.f. due to social setting (from Section 2.6) and a gain of 882.0, also on 2 d.f., due to effort given setting. These calculations are set out in Table 2.16, which also shows the corresponding mean squares and F-ratios.

TABLE 2.16: Hierarchical Anova for Two-Factor Additive Model of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting	1193.8	2	596.9	15.6
Effort|Setting	882.0	2	441.0	11.5
Residual	574.4	15	38.3	
Total	2650.2	19		

We can combine the sum of squares for setting and for effort given setting to construct a test for the overall significance of the regression. This results in an F-ratio of 13.6 on four and 15 d.f., and is highly significant. The second of the F-ratios shown in Table 2.16, which is 11.5 on two and 15 d.f., is a test for the *net* effect of family planning effort after accounting for social setting, and is highly significant. (The first of the F-ratios in the table, 15.6 on two and 15 d.f., is not in common use but is shown for completeness; it can be interpreted as an alternative test for the *gross* effect of setting, which combines the same numerator as the test in the previous section with a more refined denominator that takes into account the effect of effort.)

There is an alternative decomposition of the regression sum of squares into an improvement of 2040.0 on two d.f. due to effort and a further gain of 35.8 on two d.f. due to setting given effort. The latter can be contrasted with the error sum of squares of 574.4 on 15 d.f. to obtain a test of the *net* effect of setting given effort. This test would address the question of whether socio-economic conditions have an effect on fertility decline after we have accounted for family planning effort.

2.7.5 Partial and Multiple Correlation Ratios

The sums of squares described above can be turned into proportions of variance explained using the now-familiar calculations. For example the two factors together explain 2075.8 out of 2650.2, or 78.3% of the variation in CBR decline.

The square root of this proportion, 0.885 in the example, is the *multiple correlation ratio*; it is analogous to (and in fact is often called) the multiple correlation coefficient. We use the word 'ratio' to emphasize the categorical nature of the predictors and to note that it generalizes to more than one factor the correlation ratio introduced in Section 2.4.

We can also calculate the proportion of variance explained by one of the factors out of the amount left unexplained by the other. In our example effort explained 882.0 out of the 1456.6 that setting had left unexplained, or 60.6%. The square root of this proportion, 0.778, is called the *partial correlation ratio*, and can be interpreted as a measure of correlation between a discrete factor and a continuous variable after adjustment for another factor.

2.7.6 Fitted Means and Standardization

Parameter estimates from the additive model can be translated into *fitted means* using Equation 2.19 evaluated at the estimates. The body of Table 2.17 shows these values for our illustrative example. Note that we are able to estimate the expected CBR decline for strong programs in low social settings although there is no country in our dataset with that particular combination of attributes. Such extrapolation relies on the additive nature of the model and should be interpreted with caution. Comparison of observed and fitted values can yield useful insights into the adequacy of the model, a topic that will be pursued in more detail when we discuss regression diagnostics later in this chapter.

TABLE 2.17: Fitted Means Based on Two-Factor Additive Model of CBR Decline by Social Setting and Family Planning Effort

Setting	Effort			All
	Weak	Moderate	Strong	
Low	5.38	9.22	26.05	13.77
Medium	3.70	7.54	24.37	12.08
High	7.77	11.60	28.44	16.15
All	5.91	9.75	26.59	14.30

Table 2.17 also shows column (and row) means, representing expected CBR declines by effort (and setting) after adjusting for the other factor. The column means are calculated as weighted averages of the cell means in each column, with weights given by the total number of countries in each category of setting. In symbols

$$\hat{\mu}_{.j} = \sum n_{i.}\hat{\mu}_{ij}/n,$$

where we have used a dot as a subscript placeholder so $n_{i.}$ is the number of observations in row i and $\mu_{.j}$ is the mean for column j.

The resulting estimates may be interpreted as *standardized* means; they estimate the CBR decline that would be expected at each level of effort if those countries had the same distribution of social setting as the total sample. (The column means can also be calculated by using the fitted model to predict CBR decline for each observation with the dummies representing social setting held fixed at their sample averages and all other terms kept as observed. This construction helps reinforce their interpretation in terms of predicted CBR decline at various levels of effort adjusted for setting.)

TABLE 2.18: CBR Decline by Family Planning Effort
Before and After Adjustment for Social Setting

Effort	CBR Decline	
	Unadjusted	Adjusted
Weak	5.00	5.91
Moderate	9.33	9.75
Strong	27.86	26.59

Standardized means may be useful in presenting the results of a regression analysis to a non-technical audience, as done in Table 2.18. The column labelled unadjusted shows the observed mean CBR decline by level of effort. The difference of 23 points between strong and weak programs may be due to program effort, but could also reflect differences in social setting. The column labelled adjusted corrects for compositional differences in social setting using the additive model. The difference of 21 points may be interpreted as an effect of program effort net of social setting.

2.7.7 Multiple Classification Analysis

Multiple Classification Analysis (MCA), a technique that has enjoyed some popularity in Sociology, turns out to be just another name for the two factor additive model discussed in this section (and more generally, for multi-factor additive models). A nice feature of MCA, however, is a tradition of presenting the results of the analysis in a table that contains

- the gross effect of each of the factors, calculated using a one-factor model under the 'usual' restrictions, together with the corresponding correlation ratios (called 'eta' coefficients), and
- the net effect of each factor, calculated using a two-factor additive model under the 'usual' restrictions, together with the corresponding partial correlation ratios (unfortunately called 'beta' coefficients).

Table 2.19 shows a multiple classification analysis of the program effort data that follows directly from the results obtained so far. Estimates for the additive model under the usual restrictions can be obtained from Table 2.18 as differences between the row and column means and the overall mean.

The overall expected decline in the CBR is 14.3%. The effects of low, medium and high setting are substantially reduced after adjustment for ef-

TABLE 2.19: Multiple Classification Analysis of CBR Decline by Social Setting and Family Planning Effort

Factor	Category	Gross Effect	Eta	Net Effect	Beta
Setting	Low	−6.73		−0.54	
	Medium	−5.70		−2.22	
	High	9.45		1.85	
			0.67		0.24
Effort	Weak	−9.30		−8.39	
	Moderate	−4.97		−4.55	
	Strong	13.56		12.29	
			0.88		0.78
Total		14.30		14.30	

fort, an attenuation reflected in the reduction of the correlation ratio from 0.67 to 0.24. On the other hand, the effects of weak, moderate and strong programs are slightly reduced after adjustment for social setting, as can be seen from correlation ratios of 0.88 and 0.78 before and after adjustment. The analysis indicates that the effects of effort are more pronounced and more resilient to adjustment than the effects of social setting.

2.7.8 The Model With Interactions

The analysis so far has rested on the assumption of additivity. We now consider a more general model for the effects of two discrete factors on a continuous response which allows for more general effects

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \tag{2.20}$$

In this formulation the first three terms should be familiar: μ is a constant, and α_i and β_j are the *main* effects of levels i of the row factor and j of the column factor.

The new term $(\alpha\beta)_{ij}$ is an *interaction* effect. It represents the effect of the *combination* of levels i and j of the row and column factors. (The notation $(\alpha\beta)$ should be understood as a single symbol, not a product; we could have used γ_{ij} to denote the interaction, but the notation $(\alpha\beta)_{ij}$ is more suggestive and reminds us that the term involves a combined effect.)

One difficulty with the model as defined so far is that it is grossly overparameterized. If the row and column factors have R and C levels, respectively,

we have only RC possible cells but have introduced $1+R+C+RC$ parameters. Our preferred solution is an extension of the reference cell method, and sets to zero all parameters involving the first row or the first column in the two-way layout, so that $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$. The best way to understand the meaning of the remaining parameters is to study Table 2.20, which shows the structure of the means in a three by three layout.

TABLE 2.20: The Two-Factor Model With Interactions

Row	Column		
	1	2	3
1	μ	$\mu+\beta_2$	$\mu+\beta_3$
2	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2+(\alpha\beta)_{22}$	$\mu+\alpha_2+\beta_3+(\alpha\beta)_{23}$
3	$\mu+\alpha_3$	$\mu+\alpha_3+\beta_2+(\alpha\beta)_{32}$	$\mu+\alpha_3+\beta_3+(\alpha\beta)_{33}$

Here μ is the expected response in the reference cell, just as before. The main effects are now more specialized: α_i is the effect of level i of the row factor, compared to level one, when the column factor is at level one, and β_j is the effect of level j of the column factor, compared to level one, when the row factor is at level one. The interaction term $(\alpha\beta)_{ij}$ is the additional effect of level i of the row factor, compared to level one, when the column factor is at level j rather than one. This term can also be interpreted as the additional effect of level j of the column factor, compared to level one, when the row factor is at level i rather than one.

The key feature of this model is that the effect of a factor now depends on the levels of the other. For example the effect of level two of the row factor, compared to level one, is α_2 in the first column, $\alpha_2+(\alpha\beta)_{22}$ in the second column, and $\alpha_2+(\alpha\beta)_{23}$ in the third column.

The resulting model is a special case of the general lineal model where the model matrix $\mathbf{X}$ has a column of ones to represent the constant, a set of $R-1$ dummy variables representing the row effects, a set of $C-1$ dummy variables representing the column effects, and a set of $(R-1)(C-1)$ dummy variables representing the interactions.

The easiest way to calculate the interaction dummies is as products of the row and column dummies. If r_i takes the value one for observations in row i and zero otherwise, and c_j takes the value one for observations in column $j1$ and zero otherwise, then the product $r_i c_j$ takes the value one for observations that are in row i *and* column j, and is zero for all others.

In order to fit this model to the program effort data we need to introduce one additional constraint because the cell corresponding to strong programs

in low settings is empty. As a result, we cannot distinguish β_3 from $\beta_3 + (\alpha\beta)_{23}$. A simple solution is to set $(\alpha\beta)_{23} = 0$. This constraint is easily implemented by dropping the corresponding dummy, which would be r_2c_3 in the above notation.

The final model has eight parameters: the constant, two setting effects, two effort effects, and three (rather than four) interaction terms.

TABLE 2.21: Anova for Two-Factor Model with Interaction Effect for CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting	1193.8	2	596.9	15.5
Effort|Setting	882.0	2	441.0	11.5
Interaction	113.6	3	37.9	1.0
Residual	460.8	12	38.4	
Total	2650.2	19		

Fitting the model gives a RSS of 460.8 on 12 d.f. Combining this result with the anova for the additive model leads to the hierarchical anova in Table 2.21. The F-test for the interaction is one on three and 12 d.f. and is clearly not significant. Thus, we have no evidence to contradict the assumption of additivity. We conclude that the effect of effort is the same at all social settings. Calculation and interpretation of the parameter estimates is left as an exercise.

2.7.9 Factors or Variates?

In our analysis of CBR decline we treated social setting and family planning effort as continuous *variates* with linear effects in Sections 2.4 and 2.5, and as discrete *factors* in Sections 2.6 and 2.7.

The fundamental difference between the two approaches hinges on the assumption of linearity. When we treat a predictor as a continuous variate we assume a linear effect. If the assumption is reasonable we attain a parsimonious fit, but if it is not reasonable we are forced to introduce transformations or higher-order polynomial terms, resulting in models which are often harder to interpret.

A reasonable alternative in these cases is to model the predictor as a discrete factor, an approach that allows arbitrary changes in the response from one category to another. This approach has the advantage of a simpler

and more direct interpretation, but by grouping the predictor into categories we are not making full use of the information in the data.

In our example we found that social setting explained 45% of the variation in CBR declines when treated as a variate and 45% when treated as a factor with three levels. Both approaches give the same result, suggesting that the assumption of linearity of setting effects is reasonable.

On the other hand family planning effort explained 64% when treated as a variate and 77% when treated as a factor with three levels. The difference suggests that we might be better off grouping effort into three categories. The reason, of course, is that the effect of effort is non-linear: CBR decline changes little as we move from weak to moderate programs, but raises steeply for strong programs.

2.8 Analysis of Covariance Models

We now consider models where some of the predictors are continuous variates and some are discrete factors. We continue to use the family planning program data, but this time we treat social setting as a variate and program effort as a factor.

2.8.1 The Data and Notation

Table 2.22 shows the effort data classified into three groups, corresponding to weak (0–4), moderate (5–14) and strong (15+) programs. For each group we list the values of social setting and CBR decline.

TABLE 2.22: Social Setting Scores and CBR Percent Declines by Levels of Family Planning Effort

Family Planning Effort					
Weak		Moderate		Strong	
Setting	Change	Setting	Change	Setting	Change
46	1	68	21	89	29
74	10	70	0	77	25
35	0	60	13	84	29
83	9	55	4	89	40
68	7	51	7	87	21
74	6	91	11	84	22
72	2			84	29

As usual, we modify our notation to reflect the structure of the data. Let k denote the number of groups, or levels of the discrete factor, n_i the number of observations in group i, y_{ij} the value of the response and x_{ij} the value of the predictor for the j-th unit in the i-th group, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$.

2.8.2 The Additive Model

We keep the random structure of our model, treating y_{ij} as a realization of a random variable $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$. To express the dependence of the expected response μ_{ij} on a discrete factor we have used an anova model of the form $\mu_{ij} = \mu + \alpha_i$, whereas to model the effect of a continuous predictor we have used a regression model of the form $\mu_{ij} = \alpha + \beta x_{ij}$. Combining these two models we obtain the additive analysis of covariance model

$$\mu_{ij} = \mu + \alpha_i + \beta x_{ij}. \tag{2.21}$$

This model defines a series of straight-line regressions, one for each level of the discrete factor (you may want to peek at Figure 2.5). These lines have different intercepts $\mu + \alpha_i$, but a common slope β, so they are *parallel*. The common slope β represents the effects of the continuous variate at any level of the factor, and the differences in intercept α_i represent the effects of the discrete factor at any given value of the covariate.

The model as defined in Equation 2.21 is not identified: we could add a constant δ to each α_i and subtract it from μ without changing any of the expected values. To solve this problem we set $\alpha_1 = 0$, so μ becomes the intercept for the reference cell, and α_i becomes the difference in intercepts between levels i and one of the factor.

The analysis of covariance model may be obtained as a special case of the general linear model by letting the model matrix $\mathbf{X}$ have a column of ones representing the constant, a set of k dummy variables representing the levels of the discrete factor, and a column with the values of the continuous variate. The model is not of full column rank because the dummies add up to the constant, so we drop one of them, obtaining the reference cell parametrization. Estimation and testing then follows form the general results in Sections 2.2 and 2.3.

Table 2.23 shows the parameter estimates, standard errors and t-ratios after fitting the model to the program effort data with setting as a variate and effort as a factor with three levels.

The results show that each point in the social setting scale is associated with a further 0.17 percentage points of CBR decline at any given level of

TABLE 2.23: Parameter Estimates for Analysis of Covariance Model of CBR Decline by Social Setting and Family Planning Effort

Parameter		Symbol	Estimate	Std.Error	t-ratio
Constant		μ	-5.954	7.166	-0.83
Effort	moderate	α_2	4.144	3.191	1.30
	strong	α_3	19.448	3.729	5.21
Setting	(linear)	β	0.1693	0.1056	1.60

effort. Countries with moderate and strong programs show additional CBR declines of 19 and 4 percentage points, respectively, compared to countries with weak programs at the same social setting.

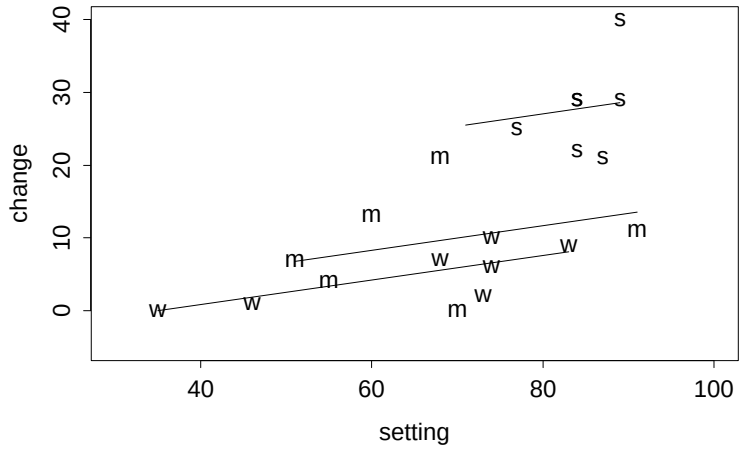

FIGURE 2.5: Analysis of Covariance Model for CBR Decline by Social Setting Score and Level of Program Effort

Figure 2.5 depicts the analysis of covariance model in graphical form. We have plotted CBR decline as a function of social setting using the letters w, m and s for weak, moderate and strong programs, respectively. The figure also shows the fitted lines for the three types of programs. The vertical distances between the lines represent the effects of program effort at any given social setting. The common slope represents the effect of setting at any given level of effort.

2.8.3 The Hierarchical Anova Table

Fitting the analysis of covariance model to our data gives a RSS of 525.7 on 16 d.f. (20 observations minus four parameters: the constant, two intercepts and one slope). Combining this result with the RSS's for the null model and for the model of Section 2.4 with a linear effect of setting, leads to the hierarchical analysis of variance shown in Table 2.24.

TABLE 2.24: Hierarchical Anova for Analysis of Covariance Model of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting (linear)	1201.1	1	1201.1	36.5
Effort|Setting	923.4	2	461.7	14.1
Residual	525.7	16	32.9	
Total	2650.2	19		

The most interesting statistic in this table is the F-test for the net effect of program effort, which is 14.1 on two and 16 d.f. and is highly significant, so we reject the hypothesis $H_0 : \alpha_2 = \alpha_3 = 0$ of no program effects. Looking at the t-ratios in Table 2.23 we see that the difference between strong and weak programs is significant, while that between moderate and weak programs is not, confirming our earlier conclusions. The difference between strong and moderate programs, which is not shown in the table, is also significant.

From these results we can calculate proportions of variance explained in the usual fashion. In this example setting explains 45.3% of the variation in CBR declines and program effort explains an additional 34.5%, representing 63.7% of what remained unexplained, for a total of 80.1%. You should be able to translate these numbers into simple, partial and multiple correlation coefficients or ratios.

2.8.4 Gross and Net Effects

The estimated net effects of setting and effort based on the analysis of covariance model may be compared with the estimated gross effects based on the simple linear regression model for setting and the one-way analysis of variance model for effort. The results are presented in a format analogous to multiple classification analysis in Table 2.25, where we have used the reference cell method rather than the 'usual' restrictions.

TABLE 2.25: Gross and Net Effects of Social Setting Score and Level of Family Planning Effort on CBR Decline

Predictor	Category	Effect	
		Gross	Net
Setting	(linear)	0.505	0.169
Effort	Weak	–	–
	Moderate	4.33	4.14
	Strong	22.86	19.45

The effect of social setting is reduced substantially after adjusting for program effort. On the other hand, the effects of program effort, measured by comparing strong and moderate programs with weak ones, are hardly changed after adjustment for social setting.

If interest centers on the effects of program effort, it may be instructive to calculate CBR declines by categories of program effort unadjusted and adjusted for linear effects of setting. To obtain adjusted means we use the fitted model to predict CBR decline with program effort set at the observed values but social setting set at the sample mean, which is 72.1 points. Thus, we calculate expected CBR decline at level i of effort holding setting constant at the mean as $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}\,72.1$. The results are shown in Table 2.26.

TABLE 2.26: CBR Decline by Family Planning Effort Before and After Linear Adjustment for Social Setting

Effort	CBR Decline	
	Unadjusted	Adjusted
Weak	5.00	6.25
Moderate	9.33	10.40
Strong	27.86	25.70

Thus, countries with strong program show on average a 28% decline in the CBR, but these countries tend to have high social settings. If we adjusted linearly for this advantage, we would expect them to show only a 26% decline. Clearly, adjusting for social setting does not change things very much.

Note that the analysis in this sections parallels the results in Section 2.7. The only difference is the treatment of social setting as a discrete factor with

three levels or as a continuous variate with a linear effect.

2.8.5 The Assumption of Parallelism

In order to test the assumption of equal slopes in the analysis of covariance model we consider a more general model where

$$\mu_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i)x_{ij}. \tag{2.22}$$

In this formulation each of the k groups has its own intercept $\mu + \alpha_i$ and its own slope $\beta + \gamma_i$.

As usual, this model is overparametrized and we introduce the reference cell restrictions, setting $\alpha_1 = \gamma_1 = 0$. As a result, μ is the constant and β is the slope for the reference cell, α_i and γ_i are the differences in intercept and slope, respectively, between level i and level one of the discrete factor. (An alternative is to drop μ and β, so that α_i is the constant and γ_i is the slope for group i. The reference cell method, however, extends more easily to models with more than one discrete factor.)

The parameter α_i may be interpreted as the effect of level i of the factor, compared to level one, when the covariate is zero. (This value will not be of interest if zero is not in the range of the data.) On the other hand, β is the expected increase in the response per unit increment in the variate when the factor is at level one. The parameter γ_i is the additional expected increase in the response per unit increment in the variate when the factor is at level i rather than one. Also, the product $\gamma_i x$ is the additional effect of level i of the factor when the covariate has value x rather than zero.

Before fitting this model to the program effort data we take the precaution of centering social setting by subtracting its mean. This simple transformation simplifies interpretation of the intercepts, since a value of zero represents the mean setting and is therefore definitely in the range of the data. The resulting parameter estimates, standard errors and t-ratios are shown in Table 2.27.

The effect of setting is practically the same for countries with weak and moderate programs, but appears to be more pronounced in countries with strong programs. Note that the slope is 0.18 for weak programs but increases to 0.64 for strong programs. Equivalently, the effect of strong programs compared to weak ones seems to be somewhat more pronounced at higher levels of social setting. For example strong programs show 13 percentage points more CBR decline than weak programs at average levels of setting, but the difference increases to 18 percentage points if setting is 10 points

TABLE 2.27: Parameter Estimates for Ancova Model with Different Slopes for CBR Decline by Social Setting and Family Planning Effort
(Social setting centered around its mean)

Parameter		Symbol	Estimate	Std.Error	t-ratio
Constant		μ	6.356	2.477	2.57
Effort	moderate	α_2	3.584	3.662	0.98
	strong	α_3	13.333	8.209	1.62
Setting	(linear)	β	0.1836	0.1397	1.31
Setting ×	moderate	γ_2	−0.0868	0.2326	−0.37
Effort	strong	γ_3	0.4567	0.6039	0.46

above the mean. However, the t ratios suggest that none of these interactions is significant.

To test the hypothesis of parallelism (or no interaction) we need to consider the joint significance of the two coefficients representing differences in slopes, i.e. we need to test $H_0 : \gamma_2 = \gamma_3 = 0$. This is easily done comparing the model of this subsection, which has a RSS of 497.1 on 14 d.f., with the parallel lines model of the previous subsection, which had a RSS of 525.7 on 16 d.f. The calculations are set out in Table 2.28.

TABLE 2.28: Hierarchical Anova for Model with Different Slopes of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting (linear)	1201.1	1	1201.1	33.8
Effort (intercepts)	923.4	2	461.7	13.0
Setting × Effort (slopes)	28.6	2	14.3	0.4
Residual	497.1	14	35.5	
Total	2650.2	19		

The test for parallelism gives an F-ratio of 0.4 on two and 14 d.f., and is clearly not significant. We therefore accept the hypothesis of parallelism and conclude that we have no evidence of an interaction between program effort and social setting.

2.9 Regression Diagnostics

The process of statistical modeling involves three distinct stages: formulating a model, fitting the model to data, and checking the model. Often, the third stage suggests a reformulation of the model that leads to a repetition of the entire cycle and, one hopes, an improved model. In this section we discuss techniques that can be used to check the model.

2.9.1 Fitted Values and Residuals

The raw materials of model checking are the *residuals* r_i defined as the differences between observed and fitted values

$$r_i = y_i - \hat{y}_i, \tag{2.23}$$

where y_i is the observed response and $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is the fitted value for the i-th unit.

The fitted values may be written in matrix notation as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Using Equation 2.7 for the m.l.e. of $\boldsymbol{\beta}$, we can write the fitted values as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ where

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

The matrix $\mathbf{H}$ is called the *hat* matrix because it maps y into y-hat. From these results one can show that the fitted values have mean $E(\hat{\mathbf{y}}) = \mu$ and variance-covariance matrix $\text{var}(\hat{\mathbf{y}}) = \mathbf{H}\sigma^2$.

The residuals may be written in matrix notation as $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, where $\mathbf{y}$ is the vector of responses and $\hat{\mathbf{y}}$ is the vector of fitted values. Since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, we can write the raw residuals as $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. It is then a simple matter to verify that under the usual second-order assumptions, the residuals have expected value $\mathbf{0}$ and variance-covariance matrix $\text{var}(\mathbf{r}) = (\mathbf{I} - \mathbf{H})\sigma^2$. In particular, the variance of the i-th residual is

$$\text{var}(r_i) = (1 - h_{ii})\sigma^2, \tag{2.24}$$

where h_{ii} is the i-th diagonal element of the hat matrix.

This result shows that the residuals may have different variances even when the original observations all have the same variance σ^2, because the precision of the fitted values depends on the pattern of covariate values.

For models with a constant it can be shown that the value of h_{ii} is always between $1/n$ and $1/r$, where n is the total number of observations and r is the number of replicates of the i-th observation (the number of units with

the same covariate values as the i-th unit). In simple linear regression with a constant and a predictor x we have

$$h_{ii} = 1/n + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2},$$

so that h_{ii} has a minimum of $1/n$ at the mean of x. Thus, the variance of the fitted values is smallest for observations near the mean and increases towards the extremes, as you might have expected. Perhaps less intuitively, this implies that the variance of the residuals is greatest near the mean and decreases as one moves towards either extreme.

Table 2.29 shows raw residuals (and other quantities to be discussed below) for the covariance analysis model fitted to the program effort data. Note that the model underestimates the decline of fertility in both Cuba and the Dominican Republic by a little bit more than eleven percentage points. At the other end of the scale, the model overestimates fertility change in Ecuador by ten percentage points.

2.9.2 Standardized Residuals

When we compare residuals for different observations we should take into account the fact that their variances may differ. A simple way to allow for this fact is to divide the raw residual by an estimate of its standard deviation, calculating the *standardized* (or internally studentized) residual

$$s_i = \frac{r_i}{\sqrt{1 - h_{ii}}\hat{\sigma}}, \tag{2.25}$$

where $\hat{\sigma}$ is the estimate of the standard deviation based on the residual sum of squares.

Standardized residuals are useful in detecting anomalous observations or *outliers*. In general, any observation with a standardized residual greater than two in absolute value should be considered worthy of further scrutiny although, as we shall see below, such observations are not necessarily outliers.

Returning to Table 2.29, we see that the residuals for both Cuba and the Dominican Republic exceed two in absolute value, whereas the residual for Ecuador does not. Standardizing the residuals helps assess their magnitude relative to the precision of the estimated regression.

2.9.3 Jack-knifed Residuals

One difficulty with standardized residuals is that they depend on an estimate of the standard deviation that may itself be affected by outliers, which may

TABLE 2.29: Regression Diagnostics for Analysis of Covariance Model of CBR Decline by Social Setting and Program Effort

Country	Residual			Leverage	Cook's
	r_i	s_i	t_i	h_{ii}	D_i
Bolivia	−0.83	−0.17	−0.16	0.262	0.0025
Brazil	3.43	0.66	0.65	0.172	0.0225
Chile	0.44	0.08	0.08	0.149	0.0003
Colombia	−1.53	−0.29	−0.28	0.164	0.0042
Costa Rica	1.29	0.24	0.24	0.143	0.0025
Cuba	11.44	**2.16**	**2.49**	0.149	**0.2043**
Dominican Rep.	11.30	**2.16**	**2.49**	0.168	**0.2363**
Ecuador	−10.04	−1.93	**−2.13**	0.173	**0.1932**
El Salvador	4.65	0.90	0.89	0.178	0.0435
Guatemala	−3.50	−0.69	−0.67	0.206	0.0306
Haiti	0.03	0.01	0.01	**0.442**	0.0000
Honduras	0.18	0.04	0.03	0.241	0.0001
Jamaica	−7.22	−1.36	−1.40	0.144	0.0782
Mexico	0.90	0.18	0.18	0.256	0.0029
Nicaragua	1.44	0.27	0.26	0.147	0.0032
Panama	−5.71	−1.08	−1.08	0.143	0.0484
Paraguay	−0.57	−0.11	−0.11	0.172	0.0006
Peru	−4.40	−0.84	−0.83	0.166	0.0352
Trinidad-Tobago	1.29	0.24	0.24	0.143	0.0025
Venezuela	−2.59	−0.58	−0.56	0.381	0.0510

thereby escape detection.

A solution to this problem is to standardize the i-th residual using an estimate of the error variance obtained by *omitting* the i-th observation. The result is the so-called *jack-knifed* (or externally studentized, or sometimes just studentized) residual

$$t_i = \frac{r_i}{\sqrt{1-h_{ii}}\hat{\sigma}_{(i)}}, \tag{2.26}$$

where $\hat{\sigma}_{(i)}$ denotes the estimate of the standard deviation obtained by fitting the model without the i-th observation, and is based on a RSS with $n-p-1$ d.f. Note that the fitted value and the hat matrix are still based on the model with all observations.

You may wonder what would happen if we omitted the i-th observation not just for purposes of standardizing the residual, but also when estimating the residual itself. Let $\hat{\boldsymbol{\beta}}_{(i)}$ denote the estimate of the regression coefficients obtained by omitting the i-th observation. We can combine this estimate with the covariate values of the i-th observation to calculate a predicted response $\hat{y}_{(i)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$ based on the rest of the data. The difference between observed and predicted responses is sometimes called a *predictive* residual

$$y_i - \hat{y}_{(i)}.$$

Consider now standardizing this residual, dividing by an estimate of its standard deviation. Since the i-th unit was not included in the regression, y_i and $\hat{y}_{(i)}$ are independent. The variance of the predictive residual is

$$\text{var}(y_i - \hat{y}_{(i)}) = (1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i)\sigma^2,$$

where $\mathbf{X}_{(i)}$ is the model matrix without the i-th row. This variance is estimated replacing the unknown σ^2 by $\hat{\sigma}^2_{(i)}$, the estimate based on the RSS of the model omitting the i-th observation. We are now in a position to calculate a standardized predictive residual

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\sqrt{\hat{\text{var}}(y_i - \hat{y}_{(i)})}}. \tag{2.27}$$

The result turns out to be exactly the same as the jack-knifed residual in Equation 2.26 and provides an alternative characterization of this statistic.

At first sight it might appear that jack-knifed residuals require a lot of calculation, as we would need to fit the model omitting each observation in turn. It turns out, however, that there are simple updating formulas that allow direct calculation of regression coefficients and RSS's after omitting one observation (see Weisberg, 1985, p. 293). These formulas can be used to show that the jack-knifed residual t_i is a simple function of the standardized residual s_i

$$t_i = s_i\sqrt{\frac{n-p-1}{n-p-s_i^2}}.$$

Note that t_i is a monotonic function of s_i, so ranking observations by their standardized residuals is equivalent to ordering them by their jack-knifed residuals.

The jack-knifed residuals on Table 2.29 make Cuba and the D.R. stand out more clearly, and suggest that Ecuador may also be an outlier.

2.9.4 A Test For Outliers

The jack-knifed residual can also be motivated as a formal test for outliers. Suppose we start from the model $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ and add a dummy variable to allow a location shift for the i-th observation, leading to the model

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma z_i,$$

where z_i is a dummy variable that takes the value one for the i-th observation and zero otherwise. In this model γ represents the extent to which the i-th response differs from what would be expected on the basis of its covariate values $\mathbf{x}_i$ and the regression coefficients $\boldsymbol{\beta}$. A formal test of the hypothesis

$$H_0 : \gamma = 0$$

can therefore be interpreted as a test that the i-th observation follows the same model as the rest of the data (i.e. is not an outlier).

The Wald test for this hypothesis would divide the estimate of γ by its standard error. Remarkably, the resulting t-ratio,

$$t_i = \frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}}$$

on $n - p - 1$ d.f., is none other than the jack-knifed residual.

This result should not be surprising in light of the previous developments. By letting the i-th observation have its own parameter γ, we are in effect estimating $\boldsymbol{\beta}$ from the rest of the data. The estimate of γ measures the difference between the response and what would be expected from the rest of the data, and coincides with the predictive residual.

In interpreting the jack-knifed residual as a test for outliers one should be careful with levels of significance. If the suspect observation had been picked in advance then the test would be valid. If the suspect observation has been selected after looking at the data, however, the nominal significance level is not valid, because we have implicitly conducted more than one test. Note that if you conduct a series of tests at the 5% level, you would expect one in twenty to be significant by chance alone.

A very simple procedure to control the overall significance level when you plan to conduct k tests is to use a significance level of α/k for each one. A basic result in probability theory known as the *Bonferroni* inequality guarantees that the overall significance level will not exceed α. Unfortunately, the procedure is conservative, and the true significance level could be considerably less than α.

For the program effort data the jack-knifed residuals have $20-4-1=15$ d.f. To allow for the fact that we are testing 20 of them, we should use a significance level of $0.05/20 = 0.0025$ instead of 0.05. The corresponding two-sided critical value of the Student's t distribution is $t_{.99875,15} = 3.62$, which is substantially higher than the standard critical value $t_{.975,15} = 2.13$. The residuals for Cuba, the D.R. and Ecuador do not exceed this more stringent criterion, so we have no evidence that these countries depart systematically from the model.

2.9.5 Influence and Leverage

Let us return for a moment to the diagonal elements of the hat matrix. Note from Equation 2.24 that the variance of the residual is the product of $1-h_{ii}$ and σ^2. As h_{ii} approaches one the variance of the residual approaches zero, indicating that the fitted value $\hat{y}_i$ is forced to come close to the observed value y_i. In view of this result, the quantity h_{ii} has been called the *leverage* or potential influence of the i-th observation. Observations with high leverage require special attention, as the fit may be overly dependent upon them.

An observation is usually considered to have high leverage if h_{ii} exceeds $2p/n$, where p is the number of predictors, including the constant, and n is the number of observations. This tolerance is not entirely arbitrary. The trace or sum of diagonal elements of $\mathbf{H}$ is p, and thus the average leverage is p/n. An observation is influential if it has more than twice the mean leverage.

Table 2.29 shows leverage values for the analysis of covariance model fitted to the program effort data. With 20 observations and four parameters, we would consider values of h_{ii} exceeding 0.4 as indicative of high leverage. The only country that exceeds this tolerance is Haiti, but Venezuela comes close. Haiti has high leverage because it is found rather isolated at the low end of the social setting scale. Venezuela is rather unique in having high social setting but only moderate program effort.

2.9.6 Actual Influence and Cook's Distance

Potential influence is not the same as actual influence, since it is always possible that the fitted value $\hat{y}_i$ would have come close to the observed value y_i anyway. Cook proposed a measure of influence based on the extent to which parameter estimates would change if one omitted the i-th observation. We define *Cook's Distance* as the standardized difference between $\hat{\boldsymbol{\beta}}_{(i)}$, the estimate obtained by omitting the i-th observation, and $\hat{\boldsymbol{\beta}}$, the estimate

obtained using all the data

$$D_i = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \, \hat{\text{var}}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})/p. \tag{2.28}$$

It can be shown that Cook's distance is also the Euclidian distance (or sum of squared differences) between the fitted values $\hat{\mathbf{y}}_{(i)}$ obtained by omitting the i-th observation and the fitted values $\hat{\mathbf{y}}$ based on all the data, so that

$$D_i = \sum_{j=1}^{n} (\hat{y}_{(i)j} - \hat{y}_j)^2/(p\hat{\sigma}^2). \tag{2.29}$$

This result follows readily from Equation 2.28 if you note that $\text{var}^{-1}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{X}/\sigma^2$ and $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$.

It would appear from this definition that calculation of Cook's distance requires a lot of work, but the regression updating formulas mentioned earlier simplify the task considerably. In fact, D_i turns out to be a simple function of the standardized residual s_i and the leverage h_{ii},

$$D_i = s_i^2 \frac{h_{ii}}{(1 - h_{ii})p}.$$

Thus, Cook's distance D_i combines residuals and leverages in a single measure of influence.

Values of D_i near one are usually considered indicative of excessive influence. To provide some motivation for this rule of thumb, note from Equation 2.28 that Cook's distance has the form W/p, where W is formally identical to the Wald statistic that one would use to test H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ if one hypothesized the value $\hat{\boldsymbol{\beta}}_{(i)}$. Recalling that W/p has an F distribution, we see that Cook's distance is equivalent to the F statistic for testing this hypothesis. A value of one is close to the median of the F distribution for a large range of values of the d.f. An observation has excessive influence if deleting it would move this F statistic from zero to the median, which is equivalent to moving the point estimate to the edge of a 50% confidence region. In such cases it may be wise to repeat the analysis without the influential observation and examine which estimates change as a result.

Table 2.29 shows Cook's distance for the analysis of covariance model fitted to the program effort data. The D.R., Cuba and Ecuador have the largest indices, but none of them is close to one. To investigate the exact nature of the D.R.'s influence, I fitted the model excluding this country. The main result is that the parameter representing the difference between moderate and weak programs is reduced from 4.14 to 1.89. Thus, a large part

of the evidence pointing to a difference between moderate and weak programs comes from the D.R., which happens to be a country with substantial fertility decline and only moderate program effort. Note that the difference was not significant anyway, so no conclusions would be affected.

Note also from Table 2.29 that Haiti, which had high leverage or potential influence, turned out to have no actual influence on the fit. Omitting this country would not alter the parameter estimates at all.

2.9.7 Residual Plots

One of the most useful diagnostic tools available to the analyst is the residual plot, a simple scatterplot of the residuals r_i versus the fitted values $\hat{y}_i$. Alternatively, one may plot the standardized residuals s_i or the jack-knifed residuals t_i versus the fitted values. In all three cases we expect basically a rectangular cloud with no discernible trend or pattern. Figure 2.6 shows a plot of jack-knifed residuals for the analysis of covariance model fitted to the program effort data. Some of the symptoms that you should be alert for

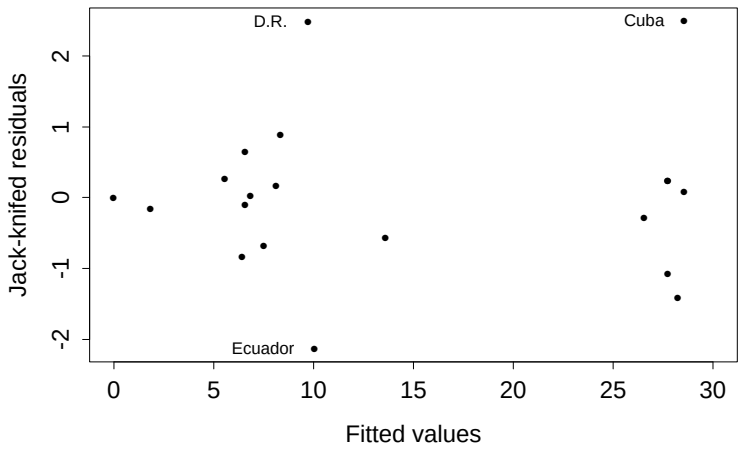

FIGURE 2.6: Residual Plot for Analysis of Covariance Model of CBR Decline by Social Setting and Program Effort

when inspecting residual plots include the following:

- Any trend in the plot, such as a tendency for negative residuals at small $\hat{y}_i$ and positive residuals at large $\hat{y}_i$. Such a trend would indicate non-linearities in the data. Possible remedies include transforming the response or introducing polynomial terms on the predictors.

- Non-constant spread of the residuals, such as a tendency for more clustered residuals for small $\hat{y}_i$ and more dispersed residuals for large $\hat{y}_i$. This type of symptom results in a cloud shaped like a megaphone, and indicates heteroscedasticity or non-constant variance. The usual remedy is a transformation of the response.

For examples of residual plots see Weisberg (1985) or Draper and Smith (1966).

2.9.8 The Q-Q Plot

A second type of diagnostic aid is the probability plot, a graph of the residuals versus the expected order statistics of the standard normal distribution. This graph is also called a *Q-Q Plot* because it plots quantiles of the data versus quantiles of a distribution. The Q-Q plot may be constructed using raw, standardized or jack-knifed residuals, although I recommend the latter.

The first step in constructing a Q-Q plot is to order the residuals from smallest to largest, so $r_{(i)}$ is the i-th smallest residual. The quantity $r_{(i)}$ is called an *order statistic*. The smallest value is the first order statistic and the largest out of n is the n-th order statistic.

The next step is to imagine taking a sample of size n from a standard normal distribution and calculating the order statistics, say $z_{(i)}$. The expected values of these order statistics are sometimes called *rankits*. A useful approximation to the i-th rankit in a sample of size n is given by

$$E(\mathbf{z}_{(i)}) \approx \Phi^{-1}[(i - 3/8)/(n + 1/4)]$$

where Φ^{-1} denotes the inverse of the standard normal distribution function. An alternative approximation proposed by Filliben (1975) uses $\Phi^{-1}[(i - 0.3175)/(n+0.365)]$ except for the first and last rankits, which are estimated as $\Phi^{-1}(1 - 0.5^{1/n})$ and $\Phi^{-1}(0.5^{1/n})$, respectively. The two approximations give very similar results.

If the observations come from a normal distribution we would expect the observed order statistics to be reasonably close to the rankits or expected order statistics. In particular, if we plot the order statistics versus the rankits we should get approximately a straight line.

Figure 2.7 shows a Q-Q plot of the jack-knifed residuals from the analysis of covariance model fitted to the program effort data. The plot comes very close to a straight line, except possibly for the upper tail, where we find a couple of residuals somewhat larger than expected. In general, Q-Q plots showing curvature indicate skew distributions, with downward concavity corresponding to negative skewness (long tail to the left) and upward

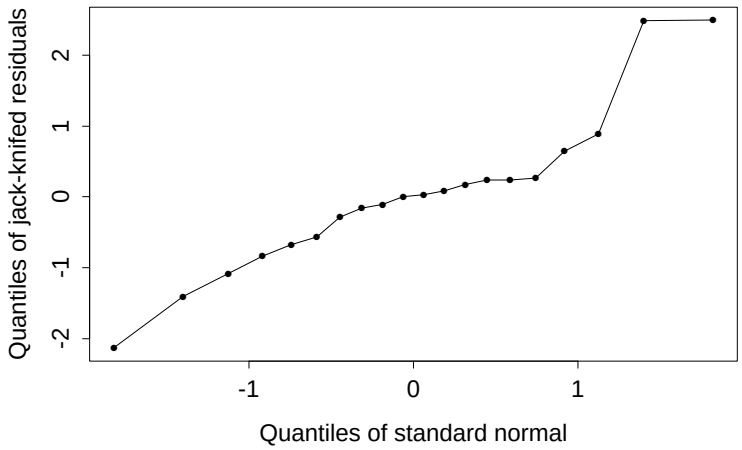

FIGURE 2.7: Q-Q Plot of Residuals From Analysis of Covariance Model of CBR Decline by Social Setting and Program Effort

concavity indicating positive skewness. On the other hand, S-shaped Q-Q plots indicate heavy tails, or an excess of extreme values, relative to the normal distribution.

Filliben (1975) has proposed a test of normality based on the linear correlation between the observed order statistics and the rankits and has published a table of critical values. The 5% points of the distribution of r for $n = 10(10)100$ are shown below. You would reject the hypothesis of normality if the correlation is *less* than the critical value. Note than to accept normality we require a very high correlation coefficient.

n	10	20	30	40	50	60	70	80	90	100
r	.917	.950	.964	.972	.977	.980	.982	.984	.985	.987

The Filliben test is closely related to the Shapiro-Francia approximation to the Shapiro-Wilk test of normality. These tests are often used with standardized or jack-knifed residuals, although the fact that the residuals are correlated affects the significance levels to an unknown extent. For the program effort data in Figure 2.7 the Filliben correlation is a respectable 0.966. Since this value exceeds the critical value of 0.950 for 20 observations, we conclude that we have no evidence against the assumption of normally distributed residuals.

2.10 Transforming the Data

We now consider what to do if the regression diagnostics discussed in the previous section indicate that the model is not adequate. The usual solutions involve transforming the response, transforming the predictors, or both.

2.10.1 Transforming the Response

The response is often transformed to achieve linearity and homoscedasticity or constant variance. Examples of *variance stabilizing* transformations are the square root, which tends to work well for counts, and the arc-sine transformation, which is often appropriate when the response is a proportion. These two solutions have fallen out of fashion as generalized linear models designed specifically to deal with counts and proportions have increased in popularity. My recommendation in these two cases is to abandon the linear model in favor of better alternatives such as Poisson regression and logistic regression.

Transformations to achieve linearity, or *linearizing* transformations, are still useful. The most popular of them is the logarithm, which is specially useful when one expects effects to be proportional to the response. To fix ideas consider a model with a single predictor x, and suppose the response is expected to increase 100ρ percent for each point of increase in x. Suppose further that the error term, denoted U, is multiplicative. The model can then be written as

$$Y = \gamma(1+\rho)^x U.$$

Taking logs on both sides of the equation, we obtain a linear model for the transformed response

$$\log Y = \alpha + \beta x + \epsilon,$$

where the constant is $\alpha = \log\gamma$, the slope is $\beta = \log(1+\rho)$ and the error term is $\epsilon = \log U$. The usual assumption of normal errors is equivalent to assuming that U has a log-normal distribution. In this example taking logs has transformed a relatively complicated multiplicative model to a familiar linear form.

This development shows, incidentally, how to interpret the slope in a linear regression model when the response is in the log scale. Solving for ρ in terms of β, we see that a unit increase in x is associated with an increase of $100(e^\beta - 1)$ percent in y. If β is small, $e^\beta - 1 \approx \beta$, so the coefficient can be interpreted directly as a relative effect. For $|\beta| < 0.10$ the absolute error of the approximation is less than 0.005 or half a percentage point. Thus, a coefficient of 0.10 can be interpreted as a 10% effect on the response.

A general problem with transformations is that the two aims of achieving linearity and constant variance may be in conflict. In generalized linear models the two aims are separated more clearly, as we will see later in the sequel.

2.10.2 Box-Cox Transformations

Box and Cox (1964) have proposed a family of transformations that can be used with non-negative responses and which includes as special cases all the transformations in common use, including reciprocals, logarithms and square roots.

The basic idea is to work with the power transformation

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

and assume that $y^{(\lambda)}$ follows a normal linear model with parameters $\boldsymbol{\beta}$ and σ^2 for some value of λ. Note that this transformation is essentially y^λ for $\lambda \neq 0$ and $\log(y)$ for $\lambda = 0$, but has been scaled to be continuous at $\lambda = 0$. Useful values of λ are often found to be in the range $(-2, 2)$. Except for scaling factors, -1 is the reciprocal, 0 is the logarithm, 1/2 is the square root, 1 is the identity and 2 is the square.

Given a value of λ, we can estimate the linear model parameters $\boldsymbol{\beta}$ and σ^2 as usual, except that we work with the transformed response $y^{(\lambda)}$ instead of y. To select an appropriate transformation we need to try values of λ in a suitable range. Unfortunately, the resulting models cannot be compared in terms of their residual sums of squares because these are in different units. We therefore use a likelihood criterion.

Starting from the normal distribution of the transformed response $y^{(\lambda)}$, we can change variables to obtain the distribution of y. The resulting log-likelihood is

$$\log L(\boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum (y_i^{(\lambda)} - \mu_i)^2/\sigma^2 + (\lambda - 1)\sum \log(y_i),$$

where the last term comes from the Jacobian of the transformation, which has derivative $y^{\lambda-1}$ for all λ. The other two terms are the usual normal likelihood, showing that we can estimate $\boldsymbol{\beta}$ and σ^2 for any fixed value of λ by regressing the transformed response $y^{(\lambda)}$ on the x's. Substituting the m.l.e.'s of $\boldsymbol{\beta}$ and σ^2 we obtain the concentrated or profile log-likelihood

$$\log L(\lambda) = c - \frac{n}{2}\log \mathrm{RSS}(y^{(\lambda)}) + (\lambda - 1)\sum \log(y_i),$$

where $c = \frac{n}{2}\log(2\pi/n) - \frac{n}{2}$ is a constant not involving λ.

Calculation of the profile log-likelihood can be simplified slightly by working with the alternative transformation

$$z^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \tilde{y}^{\lambda-1}} & \lambda \neq 0 \\ \log(y)\tilde{y} & \lambda = 0, \end{cases}$$

where $\tilde{y}$ is the geometric mean of the original response, best calculated as $\tilde{y} = \exp(\sum \log(y_i)/n)$. The profile log-likelihood can then be written as

$$\log L(\lambda) = c - \frac{n}{2}\log \mathrm{RSS}(z^{(\lambda)}), \tag{2.30}$$

where $\mathrm{RSS}(z^{(\lambda)})$ is the RSS after regressing $z^{(\lambda)}$ on the x's. Using this alternative transformation the models for different values of λ can be compared directly in terms of their RSS's.

In practice we evaluate this profile log-likelihood for a range of possible values of λ. Rather than selecting the exact maximum, one often rounds to a value such as -1, 0, 1/2, 1 or 2, particularly if the profile log-likelihood is relatively flat around the maximum.

More formally, let $\hat{\lambda}$ denote the value that maximizes the profile likelihood. We can test the hypothesis H_0: $\lambda = \lambda_0$ for any fixed value λ_0 by calculating the likelihood ratio criterion

$$\chi^2 = 2(\log L(\hat{\lambda}) - \log L(\lambda_0)),$$

which has approximately in large samples a chi-squared distribution with one d.f. We can also define a likelihood-based confidence interval for λ as the set of values that would be a accepted by the above test, i.e. the set of values for which twice the log-likelihood is within $\chi^2_{1-\alpha,1}$ of twice the maximum log-likelihood. Identifying these values requires a numerical search procedure.

Box-Cox transformations are designed for non-negative responses, but can be applied to data that have occassional zero or negative values by adding a constant α to the response before applying the power transformation. Although α could be estimated, in practice one often uses a small value such as a half or one (depending, obviously, on the scale of the response).

Let us apply this procedure to the program effort data. Since two countries show no decline in the CBR, we add 0.5 to all responses before transforming them. Figure 2.8 shows the profile log-likelihood as a function of λ for values in the range $(-1, 2)$. Note that $\lambda = 1$ is not a bad choice, indicating that the model in the original scale is reasonable. A slightly better choice appears to be $\lambda = 0.5$, which is equivalent to using $\sqrt{y + 0.5}$ as the

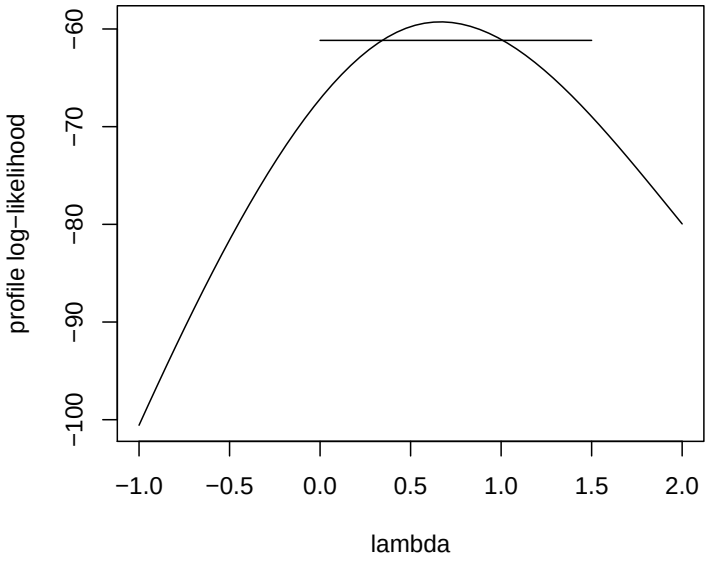

FIGURE 2.8: Profile Log-likelihood for Box-Cox Transformations for Ancova Model of CBR Decline by Setting and Effort

response. Fitting this model leads to small changes in the significance of the coefficients of setting and strong programs, but does not materially alter any of the conclusions.

More formally, we note that the profile log-likelihood for $\lambda = 1$ is -61.07. The maximum is attained at $\lambda = 0.67$ and is -59.25. Twice the difference between these values gives a chi-squared statistic of 3.65 on one degree of freedom, which is below the 5% critical value of 3.84. Thus, there is no evidence that we need to transform the response. A more detailed search shows that a 95% confidence interval for λ goes from 0.34 to 1.01. The horizontal line in Figure 2.8, at a height of -61.17, identifies the limits of the likelihood-based confidence interval.

2.10.3 The Atkinson Score

The Box-Cox procedure requires fitting a series of linear models, one for each trial value of λ. Atkinson (1985) has proposed a simpler procedure that gives

a quick indication of whether a transformation of the response is required at all. In practical terms, this technique involves adding to the model an auxiliary variable a defined as

$$a_i = y_i \left(\log(y_i/\tilde{y}) - 1\right), \tag{2.31}$$

where $\tilde{y}$ is the geometric mean of y, as in the previous subsection. Let γ denote the coefficient of a in the expanded model. If the estimate of γ is significant, then a Box-Cox transformation is indicated. A preliminary estimate of the value of λ is $1 - \hat{\gamma}$.

To see why this procedure works suppose the true model is

$$\mathbf{z}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where we have used the scale-independent version of the Box-Cox transformation. Expanding the left-hand-side using a first-order Taylor series approximation around $\lambda = 1$ gives

$$z^{(\lambda)} \approx z^{(1)} + (\lambda - 1) \left.\frac{dz^{(\lambda)}}{d\lambda}\right|_{\lambda=1}.$$

The derivative evaluated at $\lambda = 1$ is $a + \log \tilde{y} + 1$, where a is given by Equation 2.31. The second term does not depend on λ, so it can be absorbed into the constant. Note also that $z^{(1)} = y - 1$. Using these results we can rewrite the model as

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta} + (1 - \lambda)\mathbf{a} + \boldsymbol{\epsilon}.$$

Thus, to a first-order approximation the coefficient of the ancillary variable is $1 - \lambda$.

For the program effort data, adding the auxiliary variable a (calculated using CBR+1/2 to avoid taking the logarithm of zero) to the analysis of covariance model gives a coefficient of 0.59, suggesting a Box-Cox transformation with $\lambda = 0.41$. This value is reasonably close to the square root transformation suggested by the profile log-likelihood. The associated t-statistic is significant at the two percent level, but the more precise likelihood ratio criterion of the previous section, though borderline, was not significant. In conclusion, we do not have strong evidence of a need to transform the response.

2.10.4 Transforming the Predictors

The Atkinson procedure is similar in spirit to a procedure first suggested by Box and Tidwell (1962) to check whether one of the predictors needs to be

transformed. Specifically, to test whether one should use a transformation x^λ of a continuous predictor x, these authors suggest adding the auxiliary covariate

$$a_i = x_i \log(x_i)$$

to a model that already has x.

Let $\hat{\gamma}$ denote the estimated coefficient of the auxiliary variate $x \log(x)$ in the expanded model. This coefficient can be tested using the usual t statistic with $n - p$ d.f. If the test is significant, it indicates a need to transform the predictor. A preliminary estimate of the appropriate transformation is given by $\hat{\lambda} = \hat{\gamma}/\hat{\beta}+1$, where $\hat{\beta}$ is the estimated coefficient of x in the original model with x but not $x \log(x)$.

We can apply this technique to the program effort data by calculating a new variable equal to the product of setting and its logarithm, and adding it to the covariance analysis model with setting and effort. The estimated coefficient is -0.030 with a standard error of 0.728, so there is no need to transform setting. Note, incidentally, that the effect of setting is not significant in this model.