Contents

4

Categorial Explanatory Variables

Chapter Preview. Categorical variables allow us to group observations into distinct categories. This chapter shows how to incorporate categorical variables into regression functions using binary variables, thus considerably widening the scope of potential applications for regression analysis. Statistical inference for several coefficients is introduced in this chapter to allow analysts to make decisions about categorical variables (as well as other important applications). Categorical explanatory variables also provide the basis of *ANOVA* models, representations which are equivalent to regression in some circumstances that permit easier interpretation and analysis.

4.1 The Role of Binary Variables

Categorical variables provide labels for observations to denote membership in distinct groups, or categories. A binary variable is a special case of a categorical variable. To illustrate, a binary variable may tell us whether or not someone has health insurance. A categorical variable could tell us whether someone has (i) private individual health insurance, (ii) private group insurance, (iii) public insurance or (iv) no health insurance.

For categorical variables, there may or may not be an ordering of the groups. For health insurance, it is difficult to say which is "larger," private individual versus public health insurance (such as Medicare). However, for education, we may group individuals from a dataset into "low," "intermediate" and "high" years of education. In this case, there is an ordering among groups; this ordering may or may not provide information about the dependent variable. *Factor* is another term used for a (unordered) categorical explanatory variable.

The most direct way of handling categorical variables in regression is through the use of binary variables. A categorical variable with c levels can be represented using c binary variables, one for each category. For example, from a categorical education variable, we could code c=3 binary variables: (1) a variable to indicate low education, (2) one to indicate intermediate education and (3) one to indicate high education. These binary variables are often known as *dummy variables*. In regression analysis with an intercept term, we use only c-1 of these binary variables. The remaining variable enters implicitly through the intercept term.

Through the use of binary variables, we do not make use of the ordering of categories within a factor. Because no assumption is made regarding the ordering of the categories, for the model fit it does not matter which variable is dropped with

regard to the fit of the model. However, it does matter for the interpretation of the regression coefficients. Consider the following example.

Example - Car Prices. Motor Trend's *1993 New Car Buyer's Guide* provides information on 173 new cars, including the price, horsepower and the type of car. Here, we consider LNPRICE, the natural logarithm of the car price, as the response variable of interest. We use HP, the car's horsepower, as a continuous explanatory variable. Presumably, consumers are willing to pay more for more powerful cars, and HP is a standard industry measure of a car's power. We also consider CARCLASS, the car class, where there are $c = 5$ different types of cars. The variable CARCLASS is categorical, where 0 means Convertible, 1 means Coupe, 2 means Hatchback, 4 means Sedan and 5 means Mini-Van.

We begin by summarizing each continuous variable in Table 4.1. The next step is to display the distribution of the continuous variables. Figure 4.1 shows the box plot for logarithmic car prices. Here the box captures the middle 50% of the data and the so-called "whiskers" capture the middle 80%. This figure shows that the dependent variable is approximately symmetric (which is not true of the price before taking logarithms).

Table 4.1. *Summary Statistics of Each Continuous Variable*

	Number	Mean	Median	Standard deviation	Minimum	Maximum
LNPRICE	173	9.80	9.70	0.60	8.81	11.612
HP	173	147	134	60	55	400

To summarize the categorical variable and its relation to the response variable, Table 4.2 provides summary statistics by car type.

Table 4.2. *Summary Statistics of Logarithmic Price By Car Type*

	CARCLASS	Number	Mean	Standard deviation
Convertible	0	10	10.46	0.77
Coupe	1	46	9.89	0.67
Hatchback	2	19	9.29	0.47
Sedan	4	79	9.83	0.53
Mini-Van	5	19	9.63	0.23
All		173	9.80	0.60

Here, we see that most observations are sedans and coupes and that these car types have similar average prices. The mini-vans are priced similarly to the sedans and coupes, yet have relatively little variation about the average price. The convertibles display the highest average price and the highest variability of prices. Many of these

observations can also be seen in Figure 4.2 which is a box plot of logarithmic price by car type. By using the box and whiskers to capture the majority of the data, the viewer can get a quick sense of the distribution.

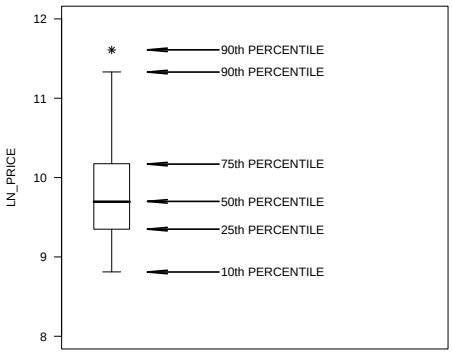

Fig. 4.1. Box plot of car price in logarithmic units. *Source: Motor Trend's 1993 New Car Buyer's Guide*

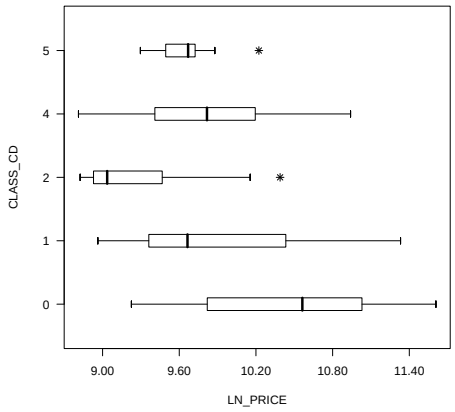

Fig. 4.2. Box plot of Logarithmic price by car type

Both Table 4.2 and Figure 4.2 show that the type of car seems important for explaining price. Is this also true of horsepower? Figure 4.3 shows the answer to be a resounding "Yes!" by exhibiting a strong relationship between LNPRICE and HP. The correlation coefficient turns out to be 87.2%.

Also in Figure 4.3 you will notice that letter coding was used to plot symbols. In this way, we are able to look at the three variables simultaneously. Unfortunately, for this application, adding the letter coding produced little additional information. The letter coding does show that the high priced cars are convertibles and coupes. You should be aware of the potential for using more sophisticated graphing techniques such as letter plots and also realize that they do not always succeed.

Are the continuous and categorical variables jointly important determinants of response? To answer this, a regression was run using LNPRICE as the response, HP as an explanatory variable and four binary variables of the car class. Here, we define C to be a binary variable for convertibles so that $C = 1$ if the car is a convertible and $C = 0$ for the other car types. Similarly, define the binary variables K for coupe, H for hatchback, S for sedan and M for mini-van.

Display 4.3 summarizes the results of a regression run using HP, C, K, H and S as explanatory variables. From the ANOVA Table, we see that the independent variables explain a good deal of the price variability. For example, the proportion of variability explained is $R^2 = 48.0304/62.1070 = 77.3\%$. Further, the t-ratios for HP show that it is a significant explanatory variable because $t(b_{HP}) = 21.2$ is very large.

From Table 4.3, we see that the fitted regression equation is

$$\widehat{y} = 8.55326 + 0.008379HP + 0.124C + 0.033K - 0.18H + 0.042S.$$

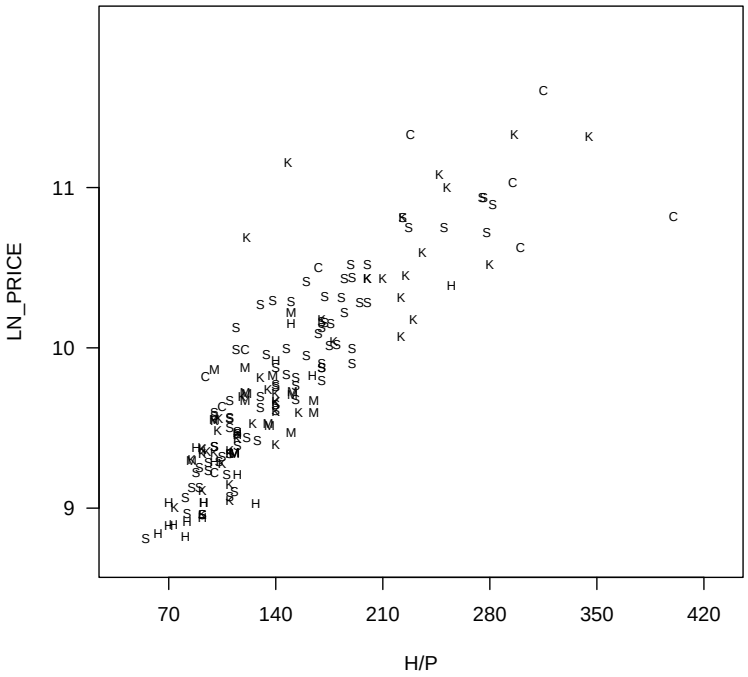

Fig. 4.3. Letter plot of logarithmic prices versus horsepower. Here, the letter codes are 'C' for convertible, 'K' for coupe, 'H' for hatchback, 'S' for sedan and 'M' for mini-van.

ANOVA Table

Source	Sum of Squares	*df*	Mean Square
Regression	48.0304	5	9.6061
Error	14.0765	167	0.0843
Total	62.1070	172	

Coefficient Estimates

Explanatory Variable	Coefficient	Standard Error	t-ratio
Constant	8.55326	0.08383	102.04
HP	0.008379	0.0003954	21.2
C	0.124	0.118	1.05
K	0.033	0.080	0.41
H	−0.180	0.094	−1.90
S	0.042	0.745	0.57

Table 4.3. *ANOVA table and coefficient estimates for Example 4.1*

Thus, for example, for a convertible with HP = 200, we would predict the logarithmic price to be

$$\widehat{y} = 8.55326 + 0.008379(200) + 0.124(1) + 0.033(0) - 0.18(0) + 0.042(0) = 10.35306,$$

which corresponds to $e^{10.35306} = \$31{,}353$. If, however, the car were a mini-van with HP = 200, we would predict the logarithmic price to be

$$\widehat{y} = 8.55326 + 0.008379(200) + 0.124(0) + 0.033(0) - 0.18(0) + 0.042(0) = 10.22906.$$

The difference between these two estimates is 0.124, the coefficient associated with

convertibles. Thus, we may interpret $b_C = 0.124$ to be the estimated expected price difference between a convertible and a mini-van.

For the model fit in the ANOVA table, it does not matter which variable is dropped with regard to the fit of the model. However, it does matter for the interpretation of the regression coefficients. To illustrate, the regression model was re-run with HP as a continuous explanatory variable and C, K, S and M as binary explanatory variables. The analysis of variance table is the same as given in Table 4.3. The coefficient estimates are given in Table 4.4. Unlike Table 4.3, we see that almost all of the t-ratios of the binary variables are now statistically significant, in that they exceed two in absolute value. Does this mean that the car type is now much more important by retaining these four binary variables?

No, the t-ratios in Table 4.4 are for comparing each of the variables with the omitted hatchback variable. The significant t-ratios mean that each car type is priced significantly higher than the hatchback. This is to be expected from our preliminary examination of the data in Table 4.2, that indicates that hatchbacks were the least expensive type of car. Further, Table 4.2 also shows that mini-vans are close to being in the middle of the price range. Thus, when we examined the summary of the regression fit in Table 4.3, we saw that some car types were more highly priced, some lower, but none were significantly different than the mini-van type.

Table 4.4. *Coefficient Estimates of a Regression of Car Price*

Explanatory variable	Coefficient	Standard error	t-ratio
Constant	8.37338	0.07950	104.3
HP	0.008379	0.0003954	21.2
C	0.304	0.121	2.53
K	0.219	0.082	2.62
S	0.222	0.076	2.94
M	0.180	0.094	1.90

All binary variables are significantly different from the hatchbacks, the omitted binary variable

4.2 Statistical Inference for Several Coefficients

In many applications, it is useful to examine several regression coefficients at the same time. For example, we have already seen the regression function in equation (3.5) expressed as a linear combination of regression coefficients. As another example, when assessing the effect of a categorical variable with c levels, we need to say something jointly about the $c-1$ binary variables that enter the regression equation. To do this, Section 4.2.1 introduces the idea of *sets* of regression coefficients using matrix algebra. Section 4.2.2 shows applications in the context of hypothesis testing and Section 4.2.3 presents other inference applications.

4.2.1 Sets of Regression Coefficients

Recall that our regression coefficients are specified by $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, a $(k+1) \times 1$ vector. It will be convenient to express linear combinations of the regres-

sion coefficients using the notation $\mathbf{C}\boldsymbol{\beta}$, where $\mathbf{C}$ is a $p \times (k+1)$ matrix that is user-specified (depending on the application). To demonstrate the broad variety of applications in which sets of regression coefficients can be used, we now present a series of special cases.

Some applications involve estimating $\mathbf{C}\boldsymbol{\beta}$. Others involve testing whether $\mathbf{C}\boldsymbol{\beta}$ equals a specific known value (denoted as $\mathbf{d}$). We call $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ the *general linear hypothesis*.

Special Case 1 - One Regression Coefficient. In Section 3.4, we investigated the importance of a single coefficient, say β_j. We may express this coefficient as $\mathbf{C}\boldsymbol{\beta}$ by choosing $p = 1$ and $\mathbf{C}$ to be a $1 \times (k+1)$ vector with a one in the $(j+1)st$ column and zeros otherwise. These choices result in

$$\mathbf{C}\boldsymbol{\beta} = (0 \ ... \ 0 \ 1 \ 0 \ ... \ 0) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \beta_j.$$

Special Case 2 - Regression Function. Here, we choose $p = 1$ and $\mathbf{C}$ to be a $1 \times (k+1)$ vector representing the transpose of a set of explanatory variables. These choices result in

$$\mathbf{C}\boldsymbol{\beta} = (x_0, x_1, \ldots, x_k) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_k x_k = \mathrm{E}\ y.$$

Special Case 3 - Linear Combination of Regression Coefficients. When $p = 1$, we use the convention that lower-case bold letters are vectors and let $\mathbf{C} = \mathbf{c}' = (c_0, \ldots, c_k)'$. In this case, $\mathbf{C}\boldsymbol{\beta}$ is a generic linear combination of regression coefficients

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + \ldots + c_k\beta_k.$$

Special Case 4 - Testing Equality of Regression Coefficients. Suppose that the interest is in testing $H_0 : \beta_1 = \beta_2$. For this purpose, let $p = 1$, $\mathbf{c}' = (0, 1, -1, 0, \ldots, 0)$, and $\mathbf{d}$=0. With these choices, we have

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} = (0, 1, -1, 0, \ldots, 0) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \beta_1 - \beta_2 = 0,$$

so that $H_0 : \beta_1 = \beta_2$.

Special Case 5 - Adequacy of the Model. It is customary in regression analysis to present a test of whether or not *any* of the explanatory variables are useful for explaining the response. Formally, this is a test of the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$. Note that, as a convention, one does not test whether or not the intercept is zero. To test this using the general linear hypothesis, we choose $p = k$, $\mathbf{d} = (0 \ ... \ 0)'$ to be a $k \times 1$ vector of zeros and $\mathbf{C}$ to be a $k \times (k+1)$

matrix such that

$$\mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{d}.$$

Special Case 6 - Testing Portions of the Model. Suppose that we are interested in comparing a *full* regression function

$$\mathrm{E}\ y = \beta_0 + \beta_1 x_1 ... + \beta_k x_k + \beta_{k+1} x_{k+1} + ... + \beta_{k+p} x_{k+p}$$

to a *reduced* regression function,

$$\mathrm{E}\ y = \beta_0 + \beta_1 x_1 ... + \beta_k x_k.$$

Beginning with the full regression, we see that if the null hypothesis $H_0 : \beta_{k+1} = ... = \beta_{k+p} = 0$ holds, then we arrive at the reduced regression. To illustrate, the variables $x_{k+1}, \ldots, x_{k+p}$ may refer to several binary variables representing a categorial variable and our interest is in whether the categorial variable is important. To test the importance of the categorical variable, we want to see whether the binary variables $x_{k+1}, \ldots, x_{k+p}$ *jointly* affect the dependent variables.

To test this using the general linear hypothesis, we choose $\mathbf{d}$ and $\mathbf{C}$ such that

$$\mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \\ \beta_{k+1} \\ \vdots \\ \beta_{k+p} \end{pmatrix} = \begin{pmatrix} \beta_{k+1} \\ \vdots \\ \beta_{k+p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{d}.$$

The additional variables do not need to be the last p in your regression run. Dropping $x_{k+1}, ..., x_{k+p}$ is for notational convenience only. From a list of $k+p$ variables $x_1, ..., x_{k+p}$, you may drop any p that you deem appropriate.

4.2.2 The General Linear Hypothesis

To recap, the general linear hypothesis can be stated as $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$. Here, $\mathbf{C}$ is a $p \times (k+1)$ matrix, $\mathbf{d}$ is a $p \times 1$ vector and both $\mathbf{C}$ and $\mathbf{d}$ are user specified and depend on the application at hand. Although $k+1$ is the number of regression coefficients, p is the number of restrictions under H_0 on these coefficients. (For those readers with knowledge of advanced matrix algebra, p is the rank of $\mathbf{C}$.) This null hypothesis is tested against the alternative $H_a : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$. This may be obvious, but we do require $p \leq k+1$ because we can not test more constraints than free parameters.

To understand the basis for the testing procedure, we first recall some of the basic properties of the regression coefficient estimators described in Section 3.3. Now, however, our goal is to understand properties of the linear combinations of regression coefficients specified by $\mathbf{C}\boldsymbol{\beta}$. An obvious estimator of this quantity is $\mathbf{Cb}$. It is easy to see that $\mathbf{Cb}$ is an unbiased estimator of $\mathbf{C}\boldsymbol{\beta}$, because $\mathrm{E}\ \mathbf{Cb} = \mathbf{C}\mathrm{E}\ \mathbf{b} = \mathbf{C}\boldsymbol{\beta}$. Moreover, the variance is $\mathrm{Var}\left(\mathbf{Cb}\right) = \mathbf{C}\mathrm{Var}\left(\mathbf{b}\right)\mathbf{C}' = \sigma^2\mathbf{C}\left(\mathbf{XX}\right)^{-1}\mathbf{C}'$. To assess the

difference between $\mathbf{d}$, the hypothesized value of $\mathbf{C}\boldsymbol{\beta}$, and its estimated value, $\mathbf{Cb}$, we use the following statistic

$$F - ratio = \frac{(\mathbf{Cb} - \mathbf{d})' \left(\mathbf{C} (\mathbf{XX})^{-1} \mathbf{C}'\right)^{-1} (\mathbf{Cb} - \mathbf{d})}{ps^2_{full}}. \tag{4.1}$$

Here, s^2_{full} is the mean square error from the full regression model. Using the theory of linear models, it can be checked that the statistic F-ratio has an F-distribution with numerator degrees of freedom $df_1 = p$ and denominator degrees of freedom $df_2 = n - (k + 1)$ (see Goldberger, 1991, for a proof). Both the statistic and the theoretical distribution are named for R. A. Fisher, a renowned scientist and statistician who did much to advance statistics as a science in the early half of the twentieth century.

Like the normal and the t-distribution, the F-distribution is a continuous distribution. The F-distribution is the sampling distribution for the F-ratio and is proportional to the ratio of two sum of squares, each of which is positive or zero. Thus, unlike the normal distribution and the t-distribution, the F-distribution takes on only nonnegative values. Recall that the t-distribution is indexed by a single degree of freedom parameter. The F-distribution is indexed by two degree of freedom parameters: one for the numerator, df_1, and one for the denominator, df_2.

The test statistic in equation (4.1) is complex in form. Fortunately, there is an alternative that is simpler to implement and to interpret; this alternative is based on the *extra sum of squares principle*.

Procedure for Testing the General Linear Hypothesis.

(i) Run the full regression and get the error sum of squares and mean square error, which we label as $(Error\ SS)_{full}$ and s^2_{full}, respectively.

(ii) Consider the model assuming the null hypothesis is true. Run a regression with this model and get the error sum of squares, which we label $(Error\ SS)_{reduced}$.

(iii) Calculate

$$F - ratio = \frac{(Error\ SS)_{reduced} - (Error\ SS)_{full}}{ps^2_{full}}. \tag{4.2}$$

(iv) Reject the null hypothesis in favor of the alternative if the F -ratio exceeds an F-value. The F-value is a percentile from the F-distribution with $df_1 = p$ and $df_2 = n-(k+1)$ degrees of freedom. The percentile is one minus the significance level of the test. Following our notation with the t-distribution, we denote this percentile as $F_{p,n-(k+1),1-\alpha}$, where α is the significance level.

To understand the extra sum of squares principle, recall that the error sum of squares for the full model is determined to be the minimum value of

$$SS(b^*_0, ..., b^*_k) = \sum_{i=1}^{n} \left(y_i - (b^*_0 + ... + b^*_k x_{i,k})\right)^2.$$

Here, $SS(b^*_0, ..., b^*_k)$ is a function of $b^*_0, ..., b^*_k$ and $(Error\ SS)_{full}$ is the minimum over

all possible values of $b_0^*, ..., b_k^*$. Similarly, $(Error\ SS)_{reduced}$ is the minimum error sum of squares under the constraints in the null hypothesis. Because there are fewer possibilities under the null hypothesis, we have that

$$(Error\ SS)_{full} \leq (Error\ SS)_{reduced}. \tag{4.3}$$

To illustrate, consider our first special case where $H_0 : \beta_j = 0$. In this case, the difference between the full and reduced models amounts to dropping a variable. A consequence of equation (4.3) is that, when adding variables to a regression model, the error sum of squares never goes up (and, in fact, usually goes down). Thus, adding variables to a regression model always increases R^2, the coefficient of determination.

How large a decrease in the error sum of squares is statistically significant? Intuitively, one can view the F-ratio as the difference in the error sum of squares divided by the number of constraints, $((Error\ SS)_{reduced} - (Error\ SS)_{full})/p$, and then rescaled by the best estimate of the variance term, the s^2, from the full model. Under the null hypothesis, this statistic follows an F-distribution and we may compare the test statistic to this distribution to see if it is unusually large.

Using the relationship $Regression\ SS = Total\ SS - Error\ SS$, we can re-express the difference in the error sum of squares as

$$(Error\ SS)_{reduced} - (Error\ SS)_{full} = (Regression\ SS)_{full} - (Regression\ SS)_{reduced}.$$

This difference is known as a *Type III Sum of Squares*. When testing the importance of a set of explanatory variables, $x_{k+1}, ..., x_{k+p}$, in the presence of $x_1, ..., x_k$, you will find that many statistical software packages compute this quantity directly in a single regression run. The advantage of this is it allows the analyst to perform an F-test using a single regression run, instead of two regression runs as in our four-step procedure described above.

Example. Before discussing the logic and the implications of the F-test, let's illustrate the use of it. Consider our taxpayer example, described in Example 3.3. This illustrates our Special Case 3 of the general linear hypothesis – testing portions of the model. Suppose that we are examining a 1990 sample of $n = 65$ returns prepared by a local branch office of a national tax preparation service. We are working with the full regression model in equation (3.9) and wish to compare it to the reduced regression model in equation (3.10). Thus, the number of variables that we consider dropping is $p = 3$.

(i) We begin by running the full regression and get $(Error\ SS)_{full} = 401.61$ and $s^2_{full} = 7.046$.

(ii) We next run the reduced regression model to get $(Error\ SS)_{reduced} = 504.88$.

(iii) We calculate the test statistic

$$F - ratio = \frac{504.88 - 401.61}{3(7.046)} = 4.886.$$

(iv) Using a 5% level of significance, it turns out that the 95th percentile from an F-distribution with $df_1 = 3$ and $df_2 = 57$ is approximately F-value ≈ 2.766. Thus, we reject the null hypothesis $H_0 : \beta_{14} = \beta_{24} = \beta_{34} = 0$. This suggests

that it is important to have separate regression functions for married and single filers.

To illustrate the test for the adequacy of the model, consider the data summarized in the ANOVA Table **??** and assume a significance level at 5 percent. From Table **??**, the F-ratio is 0.2874 / 0.0102 = 28.18. With $df_1 = 2$ and $df_2 = 33$, we have that the F-value is approximately 3.30. This leads us to reject the notion that the MILES and FOOTAGE variables are not useful in understanding rent per square foot, reaffirming what we learned in the graphical and correlation analysis. Any other result would be surprising.

Some Special Cases

The general linear hypothesis test is available whenever you can express one model as a subset of another. For this reason, it useful to think of it as a device for comparing "smaller" to "larger" models. However, the smaller model must be a subset of the larger model. For example, the general linear hypothesis test cannot be used to compare the regression functions E $y = \beta_0 + \beta_7 x_7$ versus E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. This is because the former, smaller function is not a subset of the latter, larger function.

The general linear hypothesis can be used in many instances, although its use is not always necessary. For example, suppose that we wish to test $H_0 : \beta_k = 0$. We can already seen that this null hypothesis can be examined using the t-ratio test. In this special case, it turns out that $(t - ratio)^2 = F - ratio$. Thus, these tests are equivalent for testing $H_0 : \beta_k = 0$ versus $H_a : \beta_k \neq 0$. The F-test has the advantage that it works for more than one predictor whereas the t-test has the advantage that one can consider one-sided alternatives. Thus, both tests are considered useful.

Dividing the numerator and denominator of equation (4.2) by $Total\ SS$, the test statistic can also be written as:

$$F - ratio = \frac{\left(R^2_{full} - R^2_{reduced}\right)/p}{\left(1 - R^2_{full}\right)/(n - (k+1))}. \tag{4.4}$$

The interpretation of this expression is that the F-ratio measures the drop in the coefficient of determination, R^2.

The expression in equation (4.2) is particularly useful for testing the adequacy of the model, our Special Case 5. In this case, $p = 0$, and the regression sum of squares under the reduced model is zero. Thus, we have

$$F - ratio = \frac{((Regression\ SS)_{full})/k}{s^2_{full}} = \frac{(Regression\ MS)_{full}}{(Error\ SS)_{full}}.$$

This test statistic is a regular feature of the ANOVA table for many statistical packages.

Again, dividing by $Total\ SS$, we may write

$$F - ratio = \frac{R^2}{1 - R^2}\frac{n - (k+1)}{k}.$$

Because both F-ratio and R^2 are measures of model fit, it seems intuitively plausible

that they be related in some fashion. A consequence of this relationship is the fact that as R^2 increases, so does the F-ratio and vice versa. The F-ratio is used because its sampling distribution is known under a null hypothesis so we can make statements about statistical significance. The R^2 measure is used because of the easy interpretations associated with it.

4.2.3 Estimating and Predicting Several Coefficients

Estimating Linear Combinations of Regression Coefficients

In some applications, the main interest is to estimate a linear combination of regression coefficients. To illustrate, recall in Example 3.6 that we developed a regression function for an individual's charitable contributions (y) in terms of their wages (x). In this function, there was an abrupt discontinuity at $x = 55,500$. To model this, we defined the binary variable z to be zero if $x < 55,500$ and to be one if $x \geq 55,500$ and the regression function $\mathrm{E}\ y = \beta_0 + \beta_1 x + \beta_2 z(x - 55,500)$. Thus, the marginal expected change in contributions per dollar wage change for wages in excess of 55,500 is $\partial\left(\mathrm{E}\ y\right)/\partial x = \beta_1 + \beta_2$.

To estimate $\beta_1 + \beta_2$, a reasonable estimator is $b_1 + b_2$ which is readily available from standard regression software. In addition, we would also like to compute standard errors for $b_1 + b_2$ to be used, for example, in determining a confidence interval for $\beta_1 + \beta_2$. However, b_1 and b_2 are typically correlated so that the calculation of the standard error of $b_1 + b_2$ requires estimation of the covariance between b_1 and b_2.

Estimating $\beta_1 + \beta_2$ is an example of our Special Case 3 that considers linear combinations of regression coefficients of the form $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + ... + c_k\beta_k$. For our charitable contribution's example, we would choose $c_1 = c_2 = 1$ and other c's equal to zero.

To estimate $\mathbf{c}'\boldsymbol{\beta}$, we replace the vector of parameters by the vector of estimators and use $\mathbf{c}'\mathbf{b}$. To assess the reliability of this estimator, as in Section 4.2.2, we have that $\mathrm{Var}\left(\mathbf{c}'\mathbf{b}\right) = \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$. Thus, we may define the estimated standard deviation, or standard error, of $\mathbf{c}'\mathbf{b}$ to be

$$se\left(\mathbf{c}'\mathbf{b}\right) = s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}.$$

With this quantity, a $100(1-\alpha)\%$ confidence interval for $\mathbf{c}'\boldsymbol{\beta}$ is

$$\mathbf{c}'\mathbf{b} \pm t_{n-(k+1),1-\alpha/2}\ se(\mathbf{c}'\mathbf{b}). \tag{4.5}$$

The confidence interval in equation (4.5) is valid under Assumptions F1-F5; see Goldberger (1991) for a proof. If we choose $\mathbf{c}$ to have a "1" in the $(j+1)st$ row and zeros otherwise, then $\mathbf{c}'\boldsymbol{\beta} = \beta_j$, $\mathbf{c}'\mathbf{b} = b_j$ and

$$se(b_j) = s\sqrt{(j+1)st\ diagonal\ element\ of\ (\mathbf{X}'\mathbf{X})^{-1}}.$$

Thus, (4.5) generalizes the confidence interval for individual regression coefficients introduced in Section 3.4's equation (**??**).

Another important application of equation (4.5) is the choice of $\mathbf{c}$ corresponding to a set of explanatory variables of interest, say, $\mathbf{x}^* = (1, x_1^*, x_2^*, ..., x_k^*)'$. These may correspond to an observation within the data set or to a point outside the available data. The parameter of interest, $\mathbf{c}'\boldsymbol{\beta} = \mathbf{x}^{*\prime}\boldsymbol{\beta}$, is the expected response or

the regression function at that point. Then, $\mathbf{x}^{*\prime}\mathbf{b}$ provides a point estimator and equation (4.5) provides the corresponding confidence interval.

Prediction Intervals

Prediction is an inferential goal that is closely related to estimating the regression function at a point. Suppose that, when considering charitable contributions, we know an individual's wages (and thus whether wages are in excess of 55,500) and wish to predict the amount of charitable contributions. In general, we assume that the set of explanatory variables $\mathbf{x}^*$ is known and wish to predict the corresponding response y^*. This new response follows the assumptions as described in Section 3.2. Specifically, the expected response is E $y^* = \mathbf{x}^{*\prime}\boldsymbol{\beta}$, $\mathbf{x}^*$ is nonstochastic, Var $y^* = \sigma^2$, y^* is independent of $\{y_1, ..., y_n\}$ and is normally distributed. Under these assumptions, a 100(1-α)% prediction interval for y^* is

$$\mathbf{x}^{*\prime}\mathbf{b} \pm t_{n-(k+1),1-\alpha/2}\ s\sqrt{1 + \mathbf{x}^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^*}. \tag{4.6}$$

Equation (4.6) generalizes the prediction interval for introduced in Section 2.4.

4.3 One Factor ANOVA Model

To establish notation for the one factor ANOVA model, we now consider the following example.

Example. Hospital Charges. We now study the impact of various predictors on hospital charges in the state of Wisconsin. Identifying predictors of hospital charges can provide direction for hospitals, government, insurers and consumers in controlling these factors that in turn leads to better control of hospital costs. The data for the year 1989 were obtained from the Office of Health Care Information, Wisconsin's Department of Health and Human Services. Cross sectional data are used, which details the 20 diagnosis related group (DRG) discharge costs for hospitals in the state of Wisconsin, broken down into nine major health service areas and three types of payer (Fee for service, HMO, and other). Even though there are 540 potential DRG, area and payer combinations ($20 \times 9 \times 3 = 540$), only 526 combinations were actually realized in the 1989 data set. Other predictor variables included the logarithm of the total number of discharges (NO DSCHG) and total number of hospital beds (NUM BEDS) for each combination. The response variable is the logarithm of total hospital charges per number of discharges (CHG_NUM).

As before, we use the symbol y to denote the response variable. Not surprisingly, it turns out that the diagnosis-related group (DRG) is an important determinant of costs. In this section, we focus our analysis on this categorical variable. Thus, we use the notation y_{ij} to mean the ith observation of the jth DRG. For this data set, j may be 1, 2, ..., or 20. For the jth DRG, we assume there are n_j observations. There are $n = n_1 + n_2 + \ldots + n_c$ observations. The data are:

Data for DRG 1	y_{11}	y_{21}	...	$y_{n1,1}$
Data for DRG 2	y_{12}	y_{22}	...	$y_{n2,1}$
.	.	.	...	.
Data for DRG c	y_{1c}	y_{2c}	...	$y_{nc,c}$

where $c = 20$ is the number of levels of the DRG factor. Because we do not assume an ordering of the levels, any system of ordering of the DRGs is fine. Because each level of a factor can be arranged in a single row (or column), another term for this type of data is a one way classification. Thus, a *one way model* is another term for a one factor model.

Summarizing the Data: Hospital Charges Case Study

An important summary measure of each level of the factor is the sample average. We use

$$\overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

to denote the average from the jth DRG.

To get an idea of cost by level of the factor, Figure 4.4 is a scatter plot of $\{y_{ij}\}$ versus $\{\overline{y}_j\}$. This plot illustrates several features of the data. These are:

1. First, it is clear that the average cost varies by type of DRG. For example, it turns out that *angina pectoris*, chest pains, normal newborns and chemotherapy are relatively inexpensive diagnosis-related groups. On the other hand, major joint and limb reattachment and psychoses are expensive DRGs.

2. We see that the variability is about the same for each DRG. Note that we have controlled for the frequency by working on a per discharge basis. Further, working in logarithmic units evens out the variability.

3. As emphasized by Levin, Sarlin and Webne-Behrman (1989), when the horizontal and vertical axes are on the same scale, the data are centered about a 45 degree line. This aids in interpreting the graph. In particular, the scatter plot makes it easy to identify the outlier for the group with average cost about 8.4. For this particular combination of medical treatment, health service area and type of payer, there were only two patients discharged in 1989, compared to an average of 509 discharges. Thus, although unusual, this point represents a relatively small amount of information about hospital costs and should not have an undue influence in driving the model selection.

Model Assumptions and Analysis

In this section, the mean μ_j is allowed to vary by the level of the factor, denoted by j. We can express this model as

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j, \qquad j = 1, \ldots, c.$$

This is short-hand notation for $n_1 + n_2 + \ldots + n_c = n$ equations, one for each observation. The random errors $\{\varepsilon_{ij}\}$ are assumed to be a random sample from an unknown population of errors. Because we assume the expected value of each error is zero, we have $\mathrm{E}\, y_{ij} = \mu_j$. Thus, we interpret μ_j to be the expected value of the response y_{ij}. Similarly, because we assume that the random errors have variance σ^2, we have $\mathrm{Var}\, y_{ij} = \sigma^2$. Thus, we interpret σ^2 to be the true, unknown variance of the response. This variance is assumed to be common over all factor levels.

To estimate the parameters $\{\mu_j\}$, as with regression we use the *method of least*

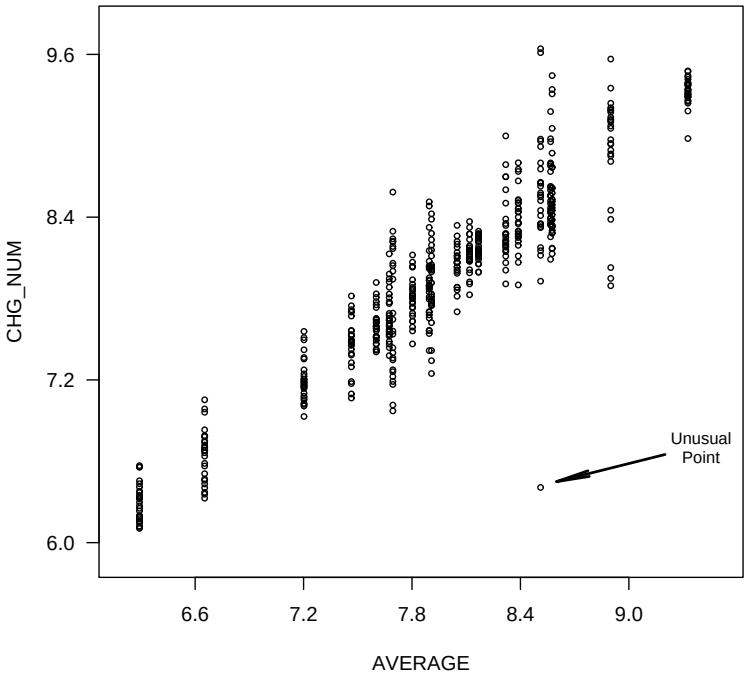

Fig. 4.4. Scatter plot of responses versus average response over diagnosis-related group (DRG). *Source: Wisconsin Department of Health and Human Services.*

squares, introduced in Section 3.1. That is, let $\hat{\mu}_j$ be a candidate estimate of μ_j. The quantity

$$SS(\hat{\mu}_1, \ldots, \hat{\mu}_c) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \hat{\mu}_j)^2$$

represents the sum of squared deviations of the responses from these candidate estimates. From straight-forward algebra, the value of $\hat{\mu}_j$ that minimizes this sum of squares is $\bar{y}_j$. Thus, $\bar{y}_j$ is the *least squares estimate* of μ_j.

To understand how reliable the estimates are, we can partition the variability as in the regression case, presented in Sections 3.3 and 4.3. The minimum sum of squared deviations is called the *error sum of squares* and is defined to be

$$\text{Error SS} = SS(\bar{y}_1, \ldots, \bar{y}_c) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

The total variation in the data set is summarized by the *total sum of squares*, Total $SS = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$. The difference, called the *factor sum of squares*, can be expressed as:

$$\text{Factor SS} = \text{Total SS} - \text{Error SS}$$

$$= \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 - \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\bar{y}_j-\bar{y})^2$$

$$= \sum_{j=1}^{c} n_j(\bar{y}_j-\bar{y})^2$$

The last two equalities follow from algebra manipulation. The Factor SS plays the same role as the Regression SS in Chapters 2 and 3. The variability decomposition is summarized in the following analysis of variance (ANOVA) table.

Table 4.5. *ANOVA Table for One Factor Model*

Source	Sum of Square	*df*	Mean Square
Factor	Factor SS	$c-1$	Factor MS
Error	Error SS	$n-c$	Error MS
Total	Total SS	$n-1$	

The conventions for this table are the same as in the regression case. That is, the mean squares (MS) column is defined by the sum of squares (SS) column divided by the degrees of freedom (df) column. Thus, $Factor\ MS \equiv (Factor\ SS)/(c-1)$ and $Error\ MS \equiv (Error\ SS)/(n-c)$. We use

$$s^2 = \text{Error MS} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j}\hat{e}_{ij}^2}{n-c}$$

to be our estimate of σ^2, where $\hat{e}_{ij} = y_{ij} - \bar{y}_j$ is the residual. The variability in the ANOVA table is often summarized by $R^2 = (Factor\ SS)/(Total\ SS)$, the coefficient of determination, or its adjusted version, $R_a^2 = 1 - s^2/s_y^2$, where $s_y^2 = (Total\ SS)/(n-1)$.

To make a formal decision as to whether the differences among machines are real, we introduce a test of hypothesis in the one factor model framework. The null, or working, hypothesis, is no difference among the levels of the factors, denoted by H_0: $\mu_1 = \mu_2 = \ldots = \mu_c$. This notation states that the null hypothesis is equality of the means. The alternative hypothesis is that at least some of the means differ from one another. As in regression, we examine the test statistic F-ratio = (Factor MS)/(Error MS). The procedure is to reject the null hypothesis in favor of the alternative if F-ratio > F-value. Here, F-value is a percentile from the F-distribution with $df_1 = c-1$ and $df_2 = n-c$ degrees of freedom. The percentile is one minus significance level of the test.

To interpret this test, recall that under H_0, we have equality of the means so that all means μ_j are equal to one another and are equal to, say, μ. The sample averages are approximations to the true means. Thus, under H_0, we expect the sample means to be close to one number, $\bar{y}$. To examine their separation, we look at squared differences, $(\bar{y}_j - \bar{y})^2$. To give levels with more observations greater weight and look at all separations together, we examine

$$\sum_{j=1}^{c} n_j(\bar{y}_j - \bar{y})^2 = \text{Factor SS}.$$

The larger that Factor SS is, the less likely we will be to believe in the null hypothesis H_0. Dividing Factor SS by $(c-1)$ and by s^2 = Error MS is the right standardization so that we can compare to the reference distribution, the F-distribution.

As another example, consider the Hospital Charges example. From Figure 4.4 , it seems clear that costs differ by DRG. To make a formal statement using our test of hypothesis machinery, some straightforward calculations yield:

Table 4.6. *ANOVA Table for Hospital Charges*

Source	Sum of Squares	*df*	Mean Square
DRG	260.09	19	13.69
Error	36.54	506	0.0722
Total	296.63	525	

From this table, we note that DRGs have explained $R^2 = 260.09/296.63 = 0.877$, or 87.7%, of the variability. The "typical" error is $s = (\text{Error MS})^{1/2} = 0.27$. To conduct the test of the null hypothesis, H_0: $\mu_1 = \mu_2 = \ldots = \mu_{20}$, we have F-ratio $= 13.69/0.0722 = 189.6$. From the F-table, with $df_1 = 19$, df_2 = infinity and, at the 5% level of significance, we have F-value = 1.590. Because F-ratio ¿ F-value, we reject the null hypothesis in favor of the alternative, that there is some difference among costs of different Diagnosis Related Groups.

Although comforting, this hypothesis test does not really tell us anything that is not clearly evident in Figure 4.4 . To supplement this information, it is useful to give estimates, and ranges of reliability, of the cost summary measures. To this end, we use $\bar{y}_j$ as our *point estimate* of the parameter μ_j. To provide a range of reliability, the corresponding interval estimate is

$$\bar{y}_j \pm (t - value)\frac{s}{\sqrt{n_j}}.$$

Here, the t-value is a percentile from the t-distribution with $n - c$ degrees. The percentile is 1 – (1 – confidence level) / 2.

To illustrate, we consider costs for the psychoses DRG, the highest cost of the medical treatment groups. This was the $j = 10$th DRG, and we have $\bar{y}_{10} = 9.3267$ and $n_{10} = 26$. Thus, a 95% confidence interval for μ_{10} is

$$9.3267 \pm \ (1.96)(0.27)/(26)^{1/2} = 9.3267 \pm \ 0.1038, \text{ or } (9.2229, 9.4305).$$

Note that these estimates are in natural logarithmic units. In dollars per discharge, our point estimate is $e^{9.3267} = \$11,234$ and our 95% confidence interval is $(e^{9.229}, e^{9.4305})$, or (\$10,188, \$12,463).

Link with Regression and Reparameterization

An important feature of the one factor ANOVA tests of hypotheses and confidence intervals is the ease of computation. Although the sum of squares appear complex, it is important to note that *no matrix calculations are required*. Rather, all of the calculations can be done through averages and sums of squares. This been an important consideration historically, before the age of readily available desktop computing. Further, it also provides for direct interpretation of the results.

In this subsection, we show how a one factor ANOVA model can be rewritten as a regression model. Using the regression formulation, we already have introduced many of the important statistical inference results. For example, with this rewriting we will be able to show that the test of equality of means is a special case of the regression test of model adequacy. Thus, the justifications of the tests and intervals estimates done in the regression case need not be repeated in the ANOVA context. Further, additional inference techniques in Chapters 5 and 6 will be available for both the regression and ANOVA models.

To this end, for a categorical variable with c levels, define c binary variables, x_1, $x_2, \ldots, x_c$. Here, x_j indicates whether or not an observation falls in the jth level. With these variables, we can rewrite our one factor ANOVA model $y = \mu_j + \varepsilon$ as

$$y = \mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_c x_c + \varepsilon. \tag{4.7}$$

The regression model in equation (4.7) includes c independent variables but does not include an intercept term, β_0. To include an intercept term, define $\tau_j = \mu_j - \mu$, where μ is an, as yet, unspecified parameter. Because each observation must fall into one of the c categories, we have $x_1 + x_2 + \ldots + x_c = 1$ for each observation. Thus, using $\mu_j = \tau_j + \mu$ in equation (4.7), we have

$$y = \mu + \tau_1 x_1 + \tau_2 x_2 + \ldots + \tau_c x_c + \varepsilon. \tag{4.8}$$

Thus, we have re-written the model into what appears to be our usual regression format, as in equation (4.8).

We use the τ in lieu of β for historical reasons. ANOVA models were invented by R.A. Fisher in connection with agricultural experiments. Here, the typical set-up is to apply several *treatments* to plots of land in order to quantify crop yield responses. Thus, the Greek "t", τ, suggests the word treatment, another term used to described levels of the factor of interest.

A simpler version of equation (4.8) can be given when we identify the level of the factor. That is, if we know an observation falls in the jth level, then only x_j is one and the other x's are 0. Thus, a simpler expression for equation (4.8) is

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}.$$

Comparing equations (4.7) and (4.8), we see that the number of parameters has increased by one. That is, in equation (4.7), there are c parameters, $\mu_1, \ldots, \mu_c$, even though in equation (4.8) there are $c + 1$ parameters, μ and $\tau_1, \ldots, \tau_c$. The model in equation (4.8) is said to be *overparameterized*. To make these two expressions equivalent, we now present two ways of *restricting* the movement of the parameters in (4.8).

The first type of restriction, usually done in the regression context, is to require

one of the τ's to be zero. This amounts to *dropping* one of the explanatory variables. For example, we might use

$$y = \mu + \tau_1 x_1 + \tau_2 x_2 + \ldots + \tau_{c-1} x_{c-1} + \varepsilon, \tag{4.9}$$

dropping x_c. With this formulation, it is easy to fit the model in equation (4.9) using regression statistical software routines because one only needs to run the regression with $c-1$ explanatory variables. However, one needs to be careful with the interpretation of parameters. To equate the models in (4.7) and (4.8), we need to define $\mu \equiv \mu_c$ and $\tau_j = \mu_j - \mu_c$ for $j = 1, 2, \ldots, c-1$. That is, the regression intercept term is the mean level of the category dropped, and each regression coefficient is the difference between a mean level and the mean level dropped. It is not necessary to drop the last level c, and indeed, one could drop any level. However, the interpretation of the parameters does depend on the variable dropped. With this restriction, the fitted values are $\hat{\mu} = \hat{\mu}_c = \bar{y}_c$ and $\hat{\tau}_j = \hat{\mu}_j - \hat{\mu}_c = \bar{y}_j - \bar{y}_c$. Recall that the carat ($\hat{}$), or "hat", stands for an estimated, or fitted, value.

The second type of restriction, from the ANOVA context, is to interpret μ as a mean for the entire population. To this end, the usual requirement is $\mu \equiv (1/n)\sum_{j=1}^{c} n_j \mu_j$, that is, μ is a weighted average of means. With this definition, we interpret $\tau_j = \mu_j - \mu$ as treatment differences between a mean level and the population mean. Another way of expressing this restriction is $\sum_{j=1}^{c} n_j \tau_j = 0$, that is, the (weighted) sum of treatment differences is zero. The disadvantage of this restriction is that it is not readily implementable with a regression routine, and a special routine is needed. The advantage is that there is a symmetry in the definitions of the parameters. There is no need to worry about which variable is being dropped from the equation, an important consideration. With this restriction, the fitted values are

$$\hat{\mu} = (1/n)\sum_{j=1}^{c} n_j \hat{\mu}_j = (1/n)\sum_{j=1}^{c} n_j \bar{y}_j = \bar{y} \quad \text{and} \quad \hat{\tau}_j = \hat{\mu}_j - \hat{\mu} = \bar{y}_j - \bar{y}.$$

4.4 Combining Categorical and Continuous Explanatory Variables

There are several ways to combine categorical and continuous explanatory variables. We initially present the case of only one categorical and one continuous variable. We then briefly present the general case, called the *general linear model*. When combining categorical and continuous variable models, we use the terminology *factor* for the categorical variable and *covariate* for the continuous variable.

Combining a Factor and Covariate

Let us begin with the simplest models that use a factor and a covariate. In Section 4.3, we introduced the one factor model:

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j, \ j = 1, \ldots, c.$$

In Chapter 2, we introduced basic linear regression in terms of one continuous variable, or covariate, using:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}.$$

To summarize different approaches for combining these variables, Table 4.7 describes several models that could be used to represent combinations of a factor and covariate.

Table 4.7. *Several Models that Represent Combinations of One Factor and One Covariate*

Model Description	Notation
One factor ANOVA (no covariate model)	$y_{ij} = \mu_j + \varepsilon_{ij}$
Regression with constant intercept and slope (no factor model)	$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$
Regression with variable intercept and constant slope (analysis of covariance model)	$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$
Regression with constant intercept and variable slope	$y_{ij} = \beta_0 + \beta_{1j} x_{ij} + \varepsilon_{ij}$
Regression with variable intercept and slope	$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$

We can interpret the regression with variable intercept and constant slope to be an additive model, because we are adding the factor effect, β_{0j}, to the covariate effect, $\beta_1 x_{ij}$. Note that one could also use the notation, μ_j, in lieu of $\beta_{0,j}$ to suggest the presence of a factor effect. This is also know as an *analysis of covariance (ANCOVA) model*. The regression with variable intercept and slope can be thought of as an *interaction model*. Here, both the intercept, β_{0j}, and slope, $\beta_{1,j}$, may vary by level of the factor. In this sense, we interpret the factor and covariate to be "interacting." The model with constant intercept and variable slope is typically not used in practice; it is included here for completeness. With this model, the factor and covariate interact only through the variable slope. Figures 4.5, 4.6 and 4.7 illustrate the expected responses of these models.

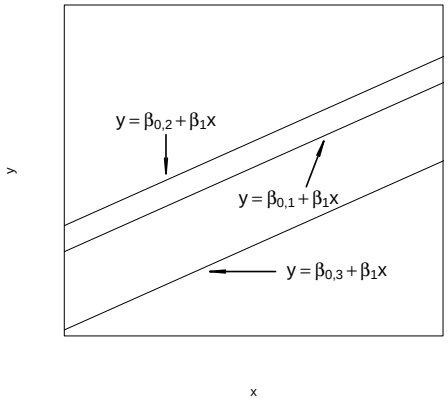

Fig. 4.5. Plot of the expected response versus the covariate for the regression model with variable intercept and constant slope.

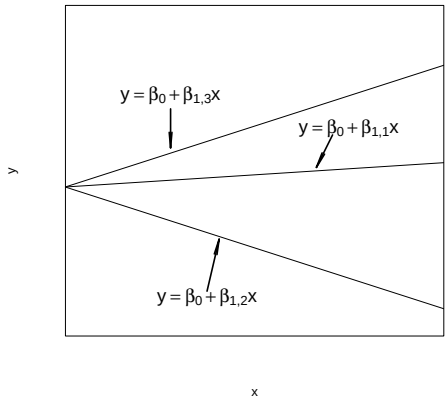

Fig. 4.6. Plot of the expected response versus the covariate for the regression model with constant intercept and variable slope.

For each model presented in Table 4.7, parameter estimates can be calculated using the method of least squares. As usual, this means writing the expected response, E

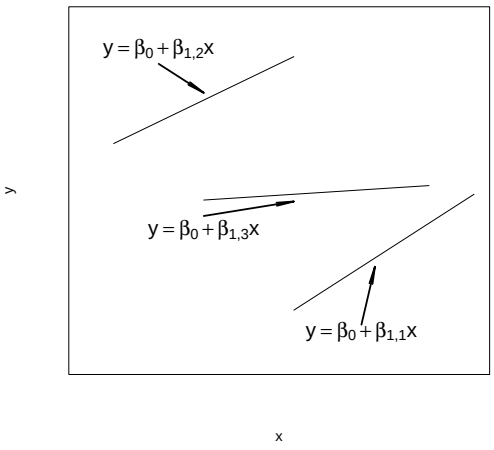

Fig. 4.7. Plot of the expected response versus the covariate for the regression model with variable intercept and variable slope.

y_{ij}, as a function of known variables and unknown parameters. Then, for candidate estimates of the parameters, an error sum of squares can be calculated and minimized over all candidate estimates. For the regression model with variable intercept and constant slope, the least squares estimates can be expressed compactly as:

$$b_1 = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}$$

and $b_{0,j} = \bar{y}_j - b_1\bar{x}_j$. Similarly, the least squares estimates for the regression model with variable intercept and slope can be expressed as:

$$b_{1,j} = \frac{\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}$$

and $b_{0,j} = \bar{y}_j - b_1\bar{x}_j$. With these parameter estimates, fitted values may be calculated.

For each model, the error sum of squares is defined as the sum of squared deviations between the observation and the corresponding fitted values, that is,

$$\text{Error SS} = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \hat{y}_{ij})^2.$$

Fitted values are defined to be the expected response with the unknown parameters replaced by their least squares estimates. For example, for the regression model with variable intercept and constant slope the fitted values are $\hat{y}_{ij} = b_{0,j} + b_1 x_{ij}$.

Example. Hospital Charges - Continued. To illustrate, we now consider the Hospital Charges case study introduced in Section 4.3. To streamline the presentation, we now consider only costs associated with three diagnostic related groups (DRGs), DRG #209, DRG #391 and DRG #430.

The covariate, x, is the natural logarithm of the number of discharges. In ideal

settings, hospitals with more patients enjoy lower costs due to economies of scale. In non-ideal settings, hospitals may not have excess capacity and thus, hospitals with more patients have higher costs. One purpose of this analysis is to investigate the relationship between hospital costs and hospital utilization.

Recall that our measure of hospital charges is the logarithm of costs per discharge (y). The scatter plot in Figure 4.8 gives a preliminary idea of the relationship between y and x. We note that there appears to be a negative relationship between y and x.

The negative relationship between y and x suggested by Figure 4.8 is misleading and is induced by an *omitted variable*, the category of the cost (DRG). To see the joint effect of the categorical variable DRG and the continuous variable x, in Figure 4.9 is a plot of y versus x where the plotting symbols are codes for the level of the categorical variable. From this plot, we see that the level of cost varies by level of the factor DRG. Moreover, for each level of DRG, the slope between y and x is either zero or positive. The slopes are not negative, as suggested by Figure 4.8.

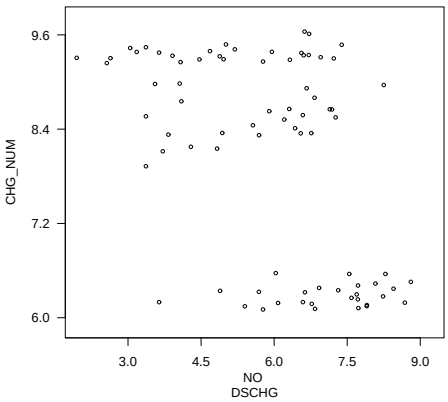

Fig. 4.8. Plot of natural logarithm of cost per discharge versus natural logarithm of the number of discharges.

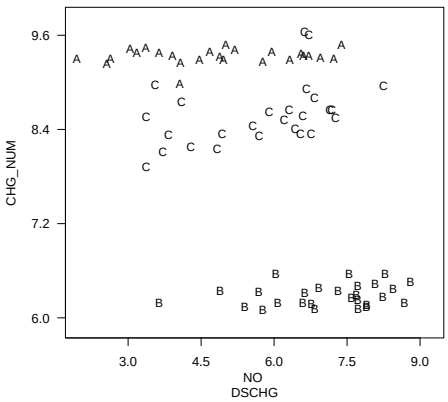

Fig. 4.9. Letter plot of natural logarithm of cost per discharge versus natural logarithm of the number of discharges by DRG. Here, A is for DRG #209, B is for DRG #391 and C is for DRG #430.

Each of the five models defined in Table 4.7 was fit to this subset of the Hospital case study. The summary statistics are in Table 4.8. For this data set, there are $n = 79$ observations and $c = 3$ levels of the DRG factor. For each model, the model degrees of freedom is the number of model parameters minus one. The error degrees of freedom is the number of observations minus the number of model parameters.

Using binary variables, each of the models in Table 4.7 can be written in a regression format. As we have seen in Section 4.2, when a model can be written as a subset of another, larger model, we have formal testing procedures available to decide which model is more appropriate. To illustrate this testing procedure with our DRG example, from Table 4.8 and the associated plots, it seems clear that the DRG factor is important. Further, a t-test, not presented here, shows that the covariate x

Table 4.8. *Degree of Freedom and Error Sum of Squares of Several Models to Represent One Factor and One Covariate for the DRG Example*

Model Description	Model degrees of freedom	Error degrees of freedom	Error Sum of Squares	R^2 (%)	Error Mean Square
One factor ANOVA	2	76	9.396	93.3	0.124
Regression with constant intercept and slope	1	77	115.059	18.2	1.222
Regression with variable intercept and constant slope	3	75	7.482	94.7	0.100
Regression with constant intercept and variable slope	3	75	14.048	90.0	0.187
Regression with variable intercept and slope	5	73	5.458	96.1	0.075

is important. Thus, let's compare the full model E $y_{ij} = \beta_{0,j} + \beta_{1,j}x$ to the reduced model E $y_{ij} = \beta_{0,j} + \beta_1 x$. In other words, is there a different slope for each DRG?

Using the notation from Section 4.2, we call the variable intercept and slope the full model. Under the null hypothesis, $H_0 : \beta_{1,1} = \beta_{1,2} = \beta_{1,3}$, we get the variable intercept, constant slope model. Thus, using the F-ratio in equation (4.2), we have

$$F\text{-ratio} = \frac{(Error\ SS)_{reduced} - (Error\ SS)_{full}}{ps^2_{full}} = \frac{7.482 - 5.458}{2(0.075)} = 13.535.$$

The 95th percentile from the F-distribution with $df_1 = p = 2$ and $df_2 = (df)_{full} = 73$ is approximately 3.13. Thus, this test leads us to reject the null hypothesis and declare the alternative, the regression model with variable intercept and variable slope, to be valid.

General Linear Model

In Section 4.1, we saw that we only use $c - 1$ indicator variables to represent a categorical variable with c levels. Similarly, in Section 4.3 we saw that the one factor ANOVA model could be expressed as a regression model with c indicator variables. However, if we had attempted to estimate the model in equation (4.8), the method of least squares would not have arrived at a unique set of regression coefficient estimates. The reason is that, in equation (4.8), each explanatory variable can be expressed as a linear combination of the others. For example, observe that $x_c = 1 - (x_1 + x_2 + \ldots + x_{c-1})$.

The fact that parameter estimates are not unique is a drawback, but not an overwhelming one. In fact, we now introduce the *general linear model*,

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon, \tag{4.10}$$

where $\{\varepsilon_i\}$ is a random sample from an unknown population with mean zero. We follow standard terminology and view the linear regression model as a special case of the general linear model. To distinguish the two sets of models, we assume that the explanatory variables are not linear combinations of one another in the linear

regression model context. This restriction is not made in the general linear model case. To illustrate, the models in equations (4.8) and (4.13) are examples of general linear models that are not regression models.

In the linear regression model case, the assumption that the explanatory variables are not linear combinations of one another means that we can compute unique estimates of the regression coefficients using the method of least squares. In the general linear model case, the parameter estimates need not be unique. However, an important feature of the general linear model is that the resulting fitted values turn out to be unique, using the method of least squares.

Specifically, suppose that we are considering the model in equation (4.15) and, using the method of least squares, our regression coefficient estimates are $b_0^o, b_1^o, \ldots, b_k^o$. This set of regression coefficients estimates minimizes our error sum of squares, but there are other sets of coefficients that also minimize the error sum of squares. The fitted values are computed as $\hat{y}_i = b_0^o + b_1^o x_{i1} + \ldots + b_k^o x_{ik}$. It can be shown that the resulting fitted values are unique, in the sense that any set of coefficients that minimize the error sum of squares produce the same fitted values.

Thus, for a set of data and a specified general linear model, fitted values are unique. Because residuals are computed as observed responses minus fitted values, we have that the residuals are unique. Because residuals are unique, we have the error sums of squares are unique. Thus, it seems reasonable, and is true, that we can use the general test of hypotheses described in Section 4.2 to decide whether collections of explanatory variables are important.

To summarize, for general linear models, parameter estimates are not unique and thus not meaningful. An important part of regression models is the interpretation of regression coefficients. This interpretation is not necessarily available in the general linear model context. However, for general linear models, we may still discuss the important of an individual variable or collection of variables through partial F-tests. Further, fitted values, and the corresponding exercise of prediction, works in the general linear model context. The advantage of the general linear model context is that we need not worry about the type of restrictions to impose on the parameters. Although not the subject of this text, this advantage is particularly important in complicated experimental designs used in the life sciences. Searle (1987) is one reference for these designs and for further details of the general linear model. The reader will find that general linear model estimation routines are widely available in statistical software packages available on the market today.

4.5 Two Factor ANOVA Model

Suppose that we now wish to consider two categorical explanatory variables, or factors. We first establish some notation in the context of the following example.

Example: Machine Run Times. In this example, the response of interest is the length of time it takes a machine to complete a certain benchmark test. The explanatory variables considered are the type of machine and the type of person operating the machine. For convenience, think of the operator as either experienced or inexperienced (a rookie). We refer to the operator as Factor 1 and the type of

machine as Factor 2, although these designations are interchangeable. Table 4.9 presents the sample data.

Table 4.9. *Hypothetical Run Times of Three Machines by Two Operators*

	Operator (Factor 1)		
Machine (Factor 2)	Rookie	Experienced	Machine Averages
1	14, 12	10, 12	$\bar{y}_{.1.} = 12$
2	15, 16	9, 12	$\bar{y}_{.2.} = 13$
3	8, 10	7, 7	$\bar{y}_{.3.} = 8$
Operator Average	$\bar{y}_{1..} = 12.5$	$\bar{y}_{2..} = 9.5$	$\bar{y}_{3..} = 11$

Extending the notation introduced in Section 4.3, we use y_{ijk} to denote the kth observation of the ith operator of the jth machine. We suppose that there are $K = 2$ observations for each combination of the $I = 2$ types of operators and $c = 3$ types of machines. Thus, there are $n = IcK = 2(3)2 = 12$ observations in total.

As in Section 4.3, an important issue that can be addressed with this data is whether the (population) mean run times differ among machine types. To this end, we define the average for each machine, $j = 1, 2, 3$, using $\bar{y}_{\cdot j \cdot} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ijk}$. Here, the notation $\{\cdot j \cdot\}$ in the subscript means sum over $i = 1, \ldots, I$, leave j fixed, and sum over $k = 1, \ldots, K$. One goal of this section is to explain part of the unknown variability in terms of the type of operator. To this end, we can define the average for each operator, $i = 1, 2$, using $\bar{y}_{i\cdot\cdot} = \frac{1}{cK} \sum_{j=1}^{c} \sum_{k=1}^{K} y_{ijk}$. It is also convenient for subsequent analyses to define the average over each combination of operator and machine $\bar{y}_{ij\cdot} = \frac{1}{K} \sum_{k=1}^{K} y_{ijk}$.

Table 4.9 shows that there may be a difference between the two types of operators as well as the three types of machines. Are the differences in sample mean run times due to sampling variability or are they due to differences in population mean run times? How does accounting for the type of operator help understand the performance of different machine types? To respond to these and related questions, we now introduce two models of variability.

Model Assumptions and Analysis - Additive Model

If we wish to put two categorical, or attribute, variables together in one model, there are two basic approaches. These are called *additive* and *interaction* models, respectively. To illustrate these two approaches, we begin with the simpler additive model.

Using the one factor formulation in equation (4.8), we interpret the parameter μ to be the population mean and τ_j to be the difference due to the jth level of Factor 2. Similarly, we introduce the parameter β_i to be interpreted as the difference due to the ith level of Factor 1. With these parameters, the two factor additive model is

$$y_{ijk} = \mu + \beta_i + \tau_j + \varepsilon_{ijk} \tag{4.11}$$

where $i = 1, \ldots, I, j = 1, \ldots, c$, and $k = 1, \ldots, K$. The errors, $\{\varepsilon_{ijk}\}$, are assumed

to be random, independent draws from a common population with mean zero and variance σ^2.

As with the one factor model, we again need to impose certain restrictions on the factor differences. So that all levels of each factor are treated in the same fashion, we require

$$\beta_1 + \beta_2 + \ldots + \beta_I = 0 \quad \text{and} \quad \tau_1 + \tau_2 + \ldots + \tau_c = 0. \tag{4.12}$$

Note that in this section we do not use the number of observations in our restrictions as we did in Section 4.3. This is because of the fact that in this section the data are assumed to be *balanced*. That is, for each combination of levels of Factors 1 and 2, we assume that there are an equal number, K, of observations available. This assumption is made primarily in order to simplify the presentation. It is possible to present the formulas where the number of observations may vary by combinations of levels (see, for example, Searle, 1987). Instead, in this chapter, we handle unbalanced data a special case of the general linear model, introduced in Section 4.5.

The least squares parameter estimates are determined by minimizing the sum of squares

$$SS(\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_I, \hat{\tau}_1, \ldots, \hat{\tau}_c) = \sum_{i=1}^{I}\sum_{j=1}^{c}\sum_{k=1}^{K}(y_{ijk} - (\hat{\mu} + \hat{\beta}_i + \hat{\tau}_j))^2.$$

Here, $\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_I, \hat{\tau}_1, \ldots, \hat{\tau}_c$ are candidate estimates of $\mu, \beta_1, \ldots, \beta_I, \tau_1, \ldots, \tau_c$. Minimizing this sum of squares subject to the restrictions in equation (4.12), the least squares estimates are

$$\hat{\mu} = \bar{y}_{\cdots}, \quad \hat{\beta}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}, \quad \text{and } \hat{\tau}_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots} \ . \tag{4.13}$$

Thus, the variability still unaccounted for, after the introduction of the parameters μ, β and τ, is summarized by

$$\begin{aligned} \text{Error SS} &= SS(\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_I, \hat{\tau}_1, \ldots, \hat{\tau}_c) = \sum_{i=1}^{I}\sum_{j=1}^{c}\sum_{k=1}^{K}(y_{ijk} - (\hat{\mu} + \hat{\beta}_i + \hat{\tau}_j))^2. \\ &= \sum_{i=1}^{I}\sum_{j=1}^{c}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2 \end{aligned} \tag{4.14}$$

To account for each source of the variability, consider the decomposition

$$\underset{(1)}{y_{ijk} - \bar{y}_{\cdots}} = \underset{(2)}{(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})} + \underset{(3)}{(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})} + \underset{(4)}{(y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})}. \tag{4.15}$$

Interpret this equation as (1) the total deviation equals (2) the deviation explained by Factor 1 plus (3) the deviation explained by Factor 2 plus (4) the unexplained deviation. Squaring each side of equation (4.15) and summing over all observations yields

$$\text{Total SS} = \text{Factor 1 SS} + \text{Factor 2 SS} + \text{Error SS}.$$

Here, Error SS is defined in equation (4.14) and, with equation (4.13),

$$\begin{aligned} \text{Total SS} &= \sum_{i=1}^{I}\sum_{j=1}^{c}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{...})^2, \qquad (4.16a) \\ \text{Factor 1 SS} &= cK\sum_{i=1}^{I}(\bar{y}_{j\cdot\cdot} - \bar{y}_{...})^2 = cK\sum_{i=1}^{I}\hat{\beta}_i^2 \\ \text{Factor 2 SS} &= IK\sum_{j=1}^{c}(\bar{y}_{\cdot j\cdot} - \bar{y}_{...})^2 = IK\sum_{j=1}^{c}\hat{\tau}_j^2. \end{aligned}$$

The variability decomposition is summarized in the following analysis of variance (ANOVA) table.

ANOVA Table for Two Factor Additive Model

Source	Sum of Squares	*df*	Mean Square
Factor 1	Factor 1 SS	$I-1$	Factor 1 MS
Factor 2	Factor 2 SS	$c-1$	Factor 2 MS
Error	Error SS	$n-(I+c-1)$	Error MS
Total	Total SS	$n-1$	

Again, the mean squares (MS) column is defined by the sum of squares (SS) column divided by the degrees of freedom (*df*) column. Thus, Factor 1 MS $\equiv$ (Factor 1 SS)$/(I-1)$, Factor 2 MS $\equiv$ (Factor 2 SS)$/(c-1)$ and Error MS $\equiv$ (Error SS)$/(n-(I+c-1))$. To understand the degrees of freedom column for the errors, first note that there are $1+I+c$ parameters, one for μ, I for β and c for τ. However, there are two restrictions on $\{\beta_i\}$ and $\{\tau_j\}$, resulting in $I+c-1$ free parameters. Thus, the error degrees of freedom follows the same rule as all regression models, the number of observations, n, minus the number of (free) parameters, $I+c-1$.

To illustrate, Table 4.10 presents results for the data in Table 4.9.

Table 4.10. *ANOVA Table for Two Factor Additive Model of Hypothetical Run Times*

Source	Sum of Squares	*df*	Mean Square
Operator (Factor 1)	27	1	27
Machine (Factor 2)	56	2	28
Error	17	8	2.12
Total	100	11	

As before, tests of hypotheses allows us to test formally for differences among levels of each factor. For example, the notation H_0: $\beta_1 = \ldots = \beta_I = 0$ stands for the null hypothesis: all Factor 1 level mean differences are equal to zero. In other words, this

is the hypothesis that there is no difference among levels of Factor 1. The alternative hypothesis is that at least some of the means differ from one another. For this test, we examine the F-ratio=(Factor 1 MS)/(Error MS). The null hypothesis is rejected in favor of the alternative if $F-ratio > F-value$. Here, the F-value is a percentile from the F-distribution with $df_1 = I-1$ and $df_2 = n-(I+c-1)$ degrees of freedom. The percentile is one - significance level. In our machine example, with $df_1 = 1$ and $df_2 = 8$, at the 5% significance level, we have F-value=5.318 from a table of the F-distribution. From Table 4.10, we have F-ratio = 27/2.12 = 12.74. Because 12.74 = F-ratio > F-value = 5.318, we reject the null hypothesis and conclude that there is a real difference between types of operators. This result reinforces our examination of the data in Table 4.9.

The test of hypothesis for differences among levels of Factor 2 is similar. To summarize, consider the Table 4.11.

Table 4.11. *Test of Hypothesis of Differences Among Levels for Two Factor Additive Model*

Factor	Null hypothesis	Alternative hypothesis	Test statistic (F-ratio)	Degree of Freedom to use with the F-value
1	H_0: $\beta_1 = \ldots = \beta_I = 0$	H_a: At least one $\beta \neq 0$	(Factor 1 MS)/ (Error MS)	$df_1 = I-1$, $df_2 = n-(I+c-1)$
2	H_0: $\tau_1 = \ldots = \tau_c = 0$	H_a: At least one $\tau \neq 0$	(Factor 2 MS)/ (Error MS)	$df_1 = c-1$, $df_2 = n-(I+c-1)$

For example, to test differences among types of machines, we hypothesize H_0: $\tau_1 = \tau_2 = \tau_3 = 0$. To perform the test, we first calculate F-ratio = (Factor 2 MS)/(Error MS) = 28/2.12 = 13.21. From the F-table with $df_1 = 2$ and $df_2 = 8$, at the 5% significance level, we have F-value = 4.459. Because 13.21 > 4.459, we reject the null hypothesis that there is no difference among machines.

Model Assumptions and Analysis - Interaction Model

For the two factor additive model, we assumed that we could simply add together the impact of each variable, together with a population mean, to form the expected response. However, it may be that reality is better represented by examining more complicated interactions between the two factors. For example, in our hypothetical machine example, it may be that experienced operators run certain types of machines much faster than inexperienced operators even though, for other types of machines, experienced operators post only marginally faster run times.

To accommodate potential interactions, we use the model

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}. \tag{4.17}$$

Here, μ_{ij} represents the mean response for the ith level of Factor 1 and the jth level of Factor 2. As with equation (4.7), we would like to rewrite this model into interpretable components. To this end, define

$$\text{(a) } \mu = \frac{1}{Ic}\sum_{i=1}^{I}\sum_{j=1}^{c}\mu_{ij}, \quad \text{(b) } \beta_i = (\frac{1}{c}\sum_{j=1}^{c}\mu_{ij}) - \mu,$$

$$\text{(c) } \tau_j = (\frac{1}{I}\sum_{i=1}^{I}\mu_{ij}) - \mu, \quad \text{(d) } (\beta\tau)_{ij} = \mu_{ij} - \beta_i - \tau_j - \mu.$$

As with the additive model, μ represents the overall mean, β_i represents Factor 1 differences and τ_j represents Factor 2 differences. We use the term $(\beta\tau)_{ij}$ to represent the interaction between the two factors.

By substituting the expression for $(\beta\tau)_{ij}$ into (4.17), we get

$$y_{ijk} = \mu + \beta_i + \tau_j + (\beta\tau)_{ij} + \varepsilon_{ijk}. \tag{4.18}$$

When comparing equations (4.17) and (4.18), we see that there are Ic linear parameters in equation (4.17) even though there are $1 + I + c + Ic$ parameters in equation (4.18). As before, certain restrictions need to be imposed on the parameters in equation (4.18) so that these models are equivalent. The restrictions adopted here are:

$$\sum_{i=1}^{I}\beta_i = 0, \ \sum_{j=1}^{c}\tau_j = 0, \ \sum_{i=1}^{I}(\beta\tau)_{ij} = 0\,, \quad \text{for each } j,$$

$$\text{and } \sum_{j=1}^{c}(\beta\tau)_{ij} = 0, \quad \text{for each } i.$$

These restrictions impose $I + c + 1$ constraints, so that there are Ic free parameters in each expression.

Parameter estimation and partitioning the variability of the interaction model parallel the development of the additive model. Thus, only a brief outline is presented here. The least squares estimates of μ, β_i and τ_j are the same as presented in equation (4.13). The least squares estimate of $(\beta\tau)_{ij}$ turns out to be $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}$. Partitioning the variability yields:

$$\text{Total SS} = \text{Factor 1 SS} + \text{Factor 2 SS} + \text{Interaction SS} + \text{Error SS}.$$

Here, Total SS, Factor 1 SS and Factor 2 SS are defined in equation (4.16a) and

$$\text{Interaction SS} = K\sum_{i=1}^{I}\sum_{j=1}^{c}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

$$\text{and} \quad \text{Error SS} = \sum_{i=1}^{I}\sum_{j=1}^{c}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij\cdot})^2.$$

These results can be summarized in the following analysis of variance table.

ANOVA Table for Two Factor Interaction Model

Source	Sum of Squares	*df*	Mean Square
Factor 1	Factor 1 SS	$I-1$	Factor 1 MS
Factor 2	Factor 2 SS	$c-1$	Factor 2 MS
Interaction	Interaction SS	$(I-1)(c-1)$	Interaction MS
Error	Error SS	$n-Ic$	Error MS
Total	Total SS	$n-1$	

We remark that the degrees of freedom for the unexplained variability, the Error Sum of Squares, is the number of observations, n, minus the number of free parameters, Ic.

From the error degrees of freedom, we see that it is necessary to have more than one observation for each combination of the two factors. That is, K must be greater than one. If K equals one, then the number of observations, $n = IcK$, equals the number of parameters. In this case, the data fits the model perfectly, there is no error, and there are no degrees of freedom available for the error sum of squares. This is not the case in the additive model where we may have $K = 1$. This is because the error degrees of freedom, $n - (I + c + 1) = IcK - (I + c + 1)$, can be greater than zero even if $K = 1$.

To test whether or not the interaction terms are important, we hypothesize H_0: all $(\beta\tau)_{ij}$'s $= 0$ versus the alternative hypothesis H_a: at least one $(\beta\tau)_{ij} \neq 0$. This null hypothesis is rejected in favor of the alternative if F-ratio = (Interaction MS)/(Error MS) $>$ F-value, where the F-value is a (1 - significance level) percentile from the F-distribution with $df_1 = (I-1)(c-1)$ and $df_2 = n - Ic$ degrees of freedom. To illustrate this test, consider the machine run data presented in Table 4.9. Table 4.12 presents the analysis of variance for the two factor interaction model fit of this example.

Table 4.12. *ANOVA Table for Two Factor Interaction Model of Hypothetical Run Times*

Source	Sum of Squares	*df*	Mean Square
Operator (Factor 1)	27	1	27
Machine (Factor 2)	56	2	28
Interaction	6	2	3
Error	11	6	1.83
Total	100	11	

To test for the presence of significant interaction terms, we compute F-ratio $= 3/1.83 = 1.64$. From an F-table with $df_1 = 2$ and $df_2 = 6$ degrees of freedom, at the 5% level of significance we have F-value $= 5.143$. Thus, we cannot reject the null hypothesis that the interaction effects are significantly different from zero.

It is also possible to test the hypothesis of no differences among factor levels using the interaction model. One would simply use the procedures outlined in Table 4.11 for the additive model but using the interaction model error mean squares and degrees of freedom. However, the interpretation of this decision-making procedure

is not clear. Under the interaction model, the terms $(\beta\tau)_{ij}$ represent the interaction, or joint effect, of the ith level of Factor 1 and the jth level of Factor 2. With terms of this type present, it is difficult to interpret the decision that either Factor 1 or Factor 2 is not important.

Deciding whether or not a factor is important may be the main goal of the data analysis. One way to address this is to test first whether or not the interaction terms are important. If not, as in the machine example above, the analyst can then represent the data using the additive model where the importance of a factor can be tested. It is important to note that with this procedure, we are fitting two models to the data and that the usual caveats apply.

Link with Regression

In this subsection, we show how to connect the two factor ANOVA models to a regression model using binary variables. To this end, for the I levels of Factor 1, define $x_{1,i}$ to be a one if the observation falls in the ith level of Factor one and is zero otherwise. Similarly, define $x_{2,j}$ to be an indicator of the jth level of Factor 2. With this notation, we can re-express the two factor additive model in equation (4.11) as

$$y = \mu + \sum_{i=1}^{I} \beta_i x_{1,i} + \sum_{j=1}^{c} \tau_j x_{2,j} + \varepsilon. \tag{4.19}$$

For example, for an observation falling in the third level of Factor 1 and the fourth level of Factor 2, we have $x_{1,3} = 1$, $x_{2,4} = 1$ and all other x's $= 0$. Thus, equation (4.19) reduces to $y_{23,k} = \mu + \beta_3 + \tau_4 + \varepsilon_{34,k}$, as in equation (4.18).

As with equation (4.18), certain restriction must be applied to the parameters. In the ANOVA models, the restriction is that the sum over levels of the parameters is zero. For regression routines, it is more straightforward to drop an explanatory indicator variable from each factor. Dropping the last variable of each factor yields

$$y \equiv \mu + \sum_{i=1}^{I-1} \beta_i x_{1,i} + \sum_{j=1}^{c-1} \tau_j x_{2,j} + \varepsilon.$$

Here, we interpret μ to be the mean response for the Ith level of Factor 1 and the cth level of Factor 2. The parameter β_i is interpreted to be the difference in mean responses between the ith and the Ith levels of Factor 1. Similarly, the parameter τ_j is interpreted to be the difference in mean responses between the jth and the cth levels of Factor 2. Thus, for the model fit, it does not matter which variables are dropped from the equation. However, it does matter when interpreting the parameters, and their resulting estimates.

The case of the two factor interaction model is similar. We can rewrite equation (4.19) as

$$y = \mu + \sum_{i=1}^{I} \beta_i x1, i + \sum_{j=1}^{c} \tau_j x_{2,j} + \sum_{i=1}^{I} \sum_{j=1}^{c} (\beta\tau)_{ij} x_{1,i} x_{2,j} + \varepsilon. \tag{4.20}$$

Dropping one indicator variable from each factor yields the analogous regression model

$$y = \mu + \sum_{i=1}^{I-1} \beta_i x_{1,i} + \sum_{j=1}^{c-1} \tau_j x_{2,j} + \sum_{i=1}^{I-1} \sum_{j=1}^{c-1} (\beta\tau)_{ij} x_{1,i} x_{2,j} + \varepsilon. \tag{4.21}$$

Again, equation (4.20) must be estimated using restricted parameters. Equation (4.21) provides an equivalent formulation without the need to restrict the parameters. To illustrate equation (4.21), consider our machine example with $I = 2$ and $c = 3$. In this case, equation (4.21) reduces to:

$$y = \mu + \beta_1 x_{1,1} + \tau_1 x_{2,1} + \tau_2 x_{2,2} + (\beta\tau)_{11} x_{1,1} x_{2,1} + (\beta\tau)_{12} x_{1,1} x_{2,2} + e.$$

This is a multiple linear regression model with five independent variables.

4.6 Technical Supplement - Matrix Expressions

4.6.1 Expressing Models with Categorical Variables in Matrix Form

In Chapter 4, we explored the analysis for models of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is a matrix of explanatory variables such that $\mathbf{X}'\mathbf{X}$ is invertible. In this section, we show how to use this model to form two models with categorical variables. In the next section, we will consider models where $\mathbf{X}'\mathbf{X}$ need not be invertible.

One Categorical Variable Model. Consider the model with one categorical variable introduced in Section 4.3, $y_j = \mu_j + \varepsilon_j$. In this model, there are c levels of the categorical variable. As in equation (4.7), this model can be written as

$$y = \mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_c x_c + \varepsilon.$$

where x_j is an indicator variable that the observation falls in the jth level. Using matrix notation, equation (4.7) can be expressed as

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n_1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{1,c} \\ \cdot \\ \cdot \\ \cdot \\ y_{n_c,c} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \cdot \\ \cdot \\ \cdot \\ \mu_c \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{n_1,1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{1,c} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{n_c,c} \end{bmatrix} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.22}$$

To make the notation more compact, we write $\mathbf{0}$ and $\mathbf{1}$ for a column of zeros and ones, respectively. With this convention, another way to express equation (4.22) is

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_1 & \cdots & \mathbf{0}_1 \\ \mathbf{0}_2 & \mathbf{1}_2 & \cdots & \mathbf{0}_2 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_c & \mathbf{0}_c & \cdots & \mathbf{1}_c \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_c \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.23}$$

Here, $\mathbf{0}_1$ and $\mathbf{1}_1$ stand for vector columns of length n_1 of zeros and ones, respectively, and similarly for $\mathbf{0}_2, \mathbf{1}_2, \ldots, \mathbf{0}_c, \mathbf{1}_c$.

Equation (4.23) allows us to apply the machinery developed for the regression model to the model with one categorical variable. As an intermediate calculation, we have

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \left(\begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_2 & \cdots & \mathbf{0}_c \\ \mathbf{0}_1 & \mathbf{1}_2 & \cdots & \mathbf{0}_c \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_1 & \mathbf{0}_2 & \cdots & \mathbf{1}_c \end{bmatrix}' \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_1 & \cdots & \mathbf{0}_1 \\ \mathbf{0}_2 & \mathbf{1}_2 & \cdots & \mathbf{0}_2 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_c & \mathbf{0}_c & \cdots & \mathbf{1}_c \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & n_c \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \frac{1}{n_c} \end{bmatrix}. \end{aligned} \tag{4.24}$$

Thus, the parameter estimates are

$$\mathbf{b} = \begin{bmatrix} \hat{\mu}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\mu}_c \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \frac{1}{n_c} \end{bmatrix} \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_2 & \cdots & \mathbf{0}_c \\ \mathbf{0}_1 & \mathbf{1}_2 & \cdots & \mathbf{0}_c \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_1 & \mathbf{0}_2 & \cdots & \mathbf{1}_c \end{bmatrix}' \begin{bmatrix} y_{1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n_1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{1,c} \\ \cdot \\ \cdot \\ \cdot \\ y_{n_c,c} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \frac{1}{n_c} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n_1} y_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1}^{n_c} y_{ic} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \cdot \\ \cdot \\ \cdot \\ \bar{y}_c \end{bmatrix} \tag{4.25}$$

Of course, the fact that $\bar{y}_j$ is the least squares estimate of μ_j could have been obtained directly from equation (4.7). However, by rewriting the model in matrix regression notation, we can appeal to linear regression model results and need not prove properties of models with categorical variables from first principles. That is, because this model is in regression format, we immediately have all the properties of the regression model.

To illustrate, from equation (4.25), the vector of fitted values is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_2 & \cdots & \mathbf{0}_c \\ \mathbf{0}_1 & \mathbf{1}_2 & \cdots & \mathbf{0}_c \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_1 & \mathbf{0}_2 & \cdots & \mathbf{1}_c \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \bar{y}_c \end{bmatrix} = \begin{bmatrix} \mathbf{1}_1\bar{y}_1 \\ \mathbf{1}_2\bar{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{1}_c\bar{y}_c \end{bmatrix}.$$

This establishes $\hat{y}_{ij} = \bar{y}_j$. Now, we have

$$\text{Error SS} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2,$$

and $s^2 = \text{Error MS} = \text{(Error SS)} \,/\, (n-c)$. This yields the one Factor ANOVA table that appears in Section 4.3. As another example, we have that the standard error of $\hat{\mu}_j$ is

$$se(\hat{\mu}_j) = s\,\sqrt{j\text{th } \textit{diagonal element} \text{ of } (\mathbf{X}'\mathbf{X})^{-1}} = s/\sqrt{n_j}.$$

One Categorical and One Continuous Variable Model. As another illustration, we consider the variable intercept and constant slope model in Table **??**. This can be expressed as a regression model using binary variables as

$$y_{ij} = \beta_{01} z_{i1} + \beta_{02} z_{i2} + ... + \beta_{0c} z_{ic} + \beta_1 x_{ij} + \varepsilon_{ij}.$$

Here, z_{ij} is an indicator variable that the observation falls in the jth level. This can be expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_1 & \cdots & \mathbf{0}_1 & \mathbf{x}_1 \\ \mathbf{0}_2 & \mathbf{1}_2 & \cdots & \mathbf{0}_2 & \mathbf{x}_2 \\ . & . & \cdots & . & . \\ . & . & \cdots & . & . \\ . & . & \cdots & . & . \\ \mathbf{0}_c & \mathbf{0}_c & \cdots & \mathbf{1}_c & \mathbf{x}_c \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ . \\ . \\ . \\ \beta_{0c} \\ \beta_1 \end{bmatrix}$$

As before, $\mathbf{0}_j$ and $\mathbf{1}_j$ stand for vector columns of length n_j of zeros and ones, respectively. Further, $\mathbf{x}_j = (x_{1j}, x_{2j}, ..., x_{n_j,j})'$ is the column of the continuous variable at the jth level. Now, straight-forward matrix algebra techniques provide the least squares estimates.

4.6.2 General Linear Model

Recall the general linear model from Section 4.5. That is, we use

$$y_i = x_{i0}\beta_0 + x_{i1}\beta_1 + ... + x_{ik}\beta_k + \varepsilon_i,$$

or, in matrix notation, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, the error terms $\{\varepsilon_i\}$ are assumed to be i.i.d. random variables with E $\varepsilon_i = 0$ and Var $\varepsilon_i = \sigma^2$. The explanatory variables $\{x_{i0}, x_{i1}, x_{i2}, ..., x_{ik}\}$ are assumed to be non-random.

In the general linear model, we do not require that $\mathbf{X}'\mathbf{X}$ be invertible. As we have seen in Chapter 4, an important reason for this generalization relates to handling categorical variables. That is, in order to use categorical variables, they are generally re-coded using binary variables. For this re-coding, generally some type of restrictions need to be made on the set of parameters associated with the indicator variables. However, it is not always clear what type of restrictions are the most intuitive. By expressing the model without requiring that $\mathbf{X}'\mathbf{X}$ be invertible, the restrictions can be imposed after the estimation is done, not before.

Normal Equations. Even when $\mathbf{X}'\mathbf{X}$ is not invertible, solutions to the normal equations still provide least squares estimates of $\boldsymbol{\beta}$. That is, the sum of squares is

$$SS(\mathbf{b}^*) = (\mathbf{y} - \mathbf{X}\mathbf{b}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b}^*),$$

where $\mathbf{b}^* = (b_0^*, b_1^*, \ldots, b_k^*)'$ is a vector of candidate estimates. Solutions of the normal equations are those vectors $\mathbf{b}^\circ$ that satisfy the normal equations

$$\mathbf{X}'\mathbf{X}\mathbf{b}^\circ = \mathbf{X}'\mathbf{y}. \tag{4.26}$$

We use the notation $^\circ$ to remind ourselves that $\mathbf{b}^\circ$ need not be unique. However, it is

a minimizer of the sum of squares. To see this, consider another candidate vector $\mathbf{b}^*$ and note that $SS(\mathbf{b}^*) = \mathbf{y}'\mathbf{y} - \mathbf{2b}^{*\prime}\mathbf{X}'\mathbf{y} + \mathbf{b}^{*\prime}\mathbf{X}'\mathbf{X}\mathbf{b}^*$. Then, using equation (4.26), we have

$$\begin{aligned} SS(\mathbf{b}^*) - SS(\mathbf{b}^\circ) &= -2\mathbf{b}^{*\prime}\mathbf{X}'\mathbf{y} + \mathbf{b}^{*\prime}\mathbf{X}'\mathbf{X}\mathbf{b}^* - (-2\mathbf{b}^{\circ\prime}\mathbf{X}\mathbf{y} + \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{X}\mathbf{b}^\circ) \\ &= -2\mathbf{b}^{*\prime}\mathbf{X}\mathbf{b}^\circ + \mathbf{b}^{*\prime}\mathbf{X}'\mathbf{X}\mathbf{b}^* + \mathbf{b}^{\circ\prime}\mathbf{X}'\mathbf{X}\mathbf{b}^\circ \\ &= (\mathbf{b}^*-\mathbf{b}^\circ)'\mathbf{X}'\mathbf{X}(\mathbf{b}^*-\mathbf{b}^\circ) = \mathbf{z}'\mathbf{z} \geq 0, \end{aligned}$$

where $\mathbf{z} = \mathbf{X}(\mathbf{b}^*-\mathbf{b}^\circ)$. Thus, any other candidate $\mathbf{b}^*$ yields a sum of squares at least as large as $SS(\mathbf{b}^\circ)$.

Unique Fitted Values. Despite the fact that there may be (infinitely) many solutions to the normal equations, the resulting fitted values, $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}^\circ$, are unique. To see this, suppose that $\mathbf{b}_1^\circ$ and $\mathbf{b}_2^\circ$ are two different solutions of equation (4.26). Let $\hat{\mathbf{y}}_1 = \mathbf{X}\mathbf{b}_1^\circ$ and $\hat{\mathbf{y}}_2 = \mathbf{X}\mathbf{b}_2^\circ$ denote the vectors of fitted values generated by these estimates. Then,

$$(\hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2)'(\hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2) = (\mathbf{b}_1^\circ-\mathbf{b}_2^\circ)'\mathbf{X}'\mathbf{X}(\mathbf{b}_1^\circ-\mathbf{b}_2^\circ) = 0$$

because $\mathbf{X}'\mathbf{X}(\mathbf{b}_1^\circ-\mathbf{b}_2^\circ) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} = \mathbf{0}$, from equation (4.26). Hence we have that $\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_2$ for any choice of $\mathbf{b}_1^\circ$ and $\mathbf{b}_2^\circ$, thus establishing the uniqueness of the fitted values.

Because the fitted values are unique, the residuals are also unique. Thus, the error sum of squares and estimates of variability (such as s^2) are also unique.

Generalized Inverses. A *generalized inverse* of a matrix $\mathbf{A}$ is a matrix $\mathbf{B}$ such that $\mathbf{ABA} = \mathbf{A}$. We use the notation $\mathbf{A}^-$ to denote the generalized inverse of $\mathbf{A}$. In the case that $\mathbf{A}$ is invertible, then $\mathbf{A}^-$ is unique and equals $\mathbf{A}^{-1}$. Although there are several definitions of generalized inverses, the above definition suffices for our purposes. See Searle (1987) for further discussion of alternative definitions of generalized inverses.

With this definition, it can be shown that a solution to the equation $\mathbf{Ab} = \mathbf{c}$ can be expressed as $\mathbf{b} = \mathbf{A}^-\mathbf{c}$. Thus, we can express a least squares estimate of $\boldsymbol{\beta}$ as $\mathbf{b}^\circ = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$. Statistical software packages can calculate versions of $(\mathbf{X}'\mathbf{X})^-$ and thus generate $\mathbf{b}^\circ$.

Estimable Functions. Above, we saw that each fitted value $\hat{y}_i$ is unique. Because fitted values are simply linear combinations of parameters estimates, it seems reasonable to ask what other linear combinations of parameter estimates are unique. To this end, we say that $\mathbf{C}\boldsymbol{\beta}$ is an *estimable function* of parameters if $\mathbf{C}\mathbf{b}^\circ$ does not depend (*is invariant*) to the choice of $\mathbf{b}^\circ$. Because fitted values are invariant to the choice of $\mathbf{b}^\circ$, we have that $\mathbf{X} = \mathbf{C}$ produces one type of estimable function. Interestingly, it turns out that all estimable functions are of the form $\mathbf{LXb}^\circ$, that is, $\mathbf{C} = \mathbf{LX}$. See Searle (1987, page 284) for a demonstration of this. Thus, all estimable function are linear combinations of fitted values, that is, $\mathbf{LXb}^\circ = \mathbf{L}\hat{\mathbf{y}}$.

Estimable functions are unbiased and a variance that does not depend on the choice of the generalized inverse. That is, it can be shown that E $\mathbf{C}\mathbf{b}^\circ = \mathbf{C}\boldsymbol{\beta}$ and Var $\mathbf{C}\mathbf{b}^\circ = \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^-\mathbf{C}'$ does not depend on the choice of $(\mathbf{X}'\mathbf{X})^-$.

Testable Hypotheses. As described in Section 4.2, if is often of interest to test H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{d}$ is a specified vector. This hypothesis is said to be *testable* if $\mathbf{C}\boldsymbol{\beta}$ is an estimable function, $\mathbf{C}$ is of full row rank, and the rank of $\mathbf{C}$ is less than the rank of $\mathbf{X}$. For consistency with the notation of Section 4.2, let p be the rank

of $\mathbf{C}$ and $k+1$ be the rank of $\mathbf{X}$. Recall that the rank of a matrix is the smaller of the number of linearly independent rows and linearly independent columns. When we say that $\mathbf{C}$ has full row rank, we mean that there are p rows in $\mathbf{C}$, so that the number of rows equals the rank.

General Linear Hypothesis. As in Section 4.2, the test statistic for examining H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ is

$$F - \text{ratio} = \frac{(\mathbf{C}\mathbf{b}^{\circ} - \mathbf{d})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{b}^{\circ} - \mathbf{d})}{ps^2_{full}}.$$

Note that the statistic F-ratio does not depend on the choice of $\mathbf{b}^{\circ}$ because $\mathbf{C}\mathbf{b}^{\circ}$ is invariant to $\mathbf{b}^{\circ}$. If H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ is a testable hypothesis and the errors ε_i are i.i.d. $N(0, \sigma^2)$, then the F-ratio has an F-distribution with $df_1 = p$ and $df_2 = n - (k+1)$.

One Categorical Variable Model. We now illustrate the general linear model by considering an over-parameterized version of the one factor model that appears in equation (4.8) using

$$y_{ij} = \mu + \tau_j + e_{ij} = \mu + \tau_1 x_{i1} + \tau_2 x_{i2} + ... + \tau_c x_{ic} + \varepsilon_{ij}.$$

At this point we do not impose additional restrictions in the parameters. As with equation (4.22), this can be written in matrix form as

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{1}_1 & \mathbf{0}_1 & \cdots & \mathbf{0}_1 \\ \mathbf{1}_2 & \mathbf{0}_2 & \mathbf{1_2} & \cdots & \mathbf{0}_2 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \mathbf{1}_c & \mathbf{0}_c & \mathbf{0}_c & \cdots & \mathbf{1}_c \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \cdot \\ \cdot \\ \cdot \\ \tau_c \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Thus, the $\mathbf{X}'\mathbf{X}$ matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & n_1 & n_2 & \cdots & n_c \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ n_c & 0 & 0 & \cdots & n_c \end{bmatrix}.$$

where $n = n_1 + n_2 + ... + n_c$. This matrix is not invertible. To see this, note that by adding the last c rows together yields the first row. Thus, the last c rows are an exact linear combination of the first row, meaning that the matrix is not full rank.

The (non-unique) least squares estimates can be expressed as

$$\mathbf{b}^{\circ} = \begin{bmatrix} \mu^{\circ} \\ \tau_1^{\circ} \\ \cdot \\ \cdot \\ \cdot \\ \tau_c^{\circ} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}.$$

Estimable functions are linear combinations of fitted values. Because fitted values

are $\hat{y}_{ij} = \bar{y}_j$, estimable functions can be expressed as $L = \sum_{j=1}^{c} a_i \bar{y}_j$ where $a_1, ..., a_c$ are constants. This linear combination of fitted values is an unbiased estimator of $\mathrm{E}\ L = \sum_{i=1}^{c} a_i(\mu + \tau_i)$.

Thus, for example, by choosing $a_1 = 1$, and the other $a_i = 0$, we see that $\mu + \tau_1$ is estimable. As another example, by choosing $a_1 = 1, a_2 = -1$, and the other $a_i = 0$, we see that $\tau_1 - \tau_2$ is estimable. It can be shown that μ is not an estimable parameter without further restrictions on $\tau_1, ..., \tau_c$.