Regression and Time Series for Actuaries

Edward W. Frees

Contents

4

Regression Using Categorial Independent Variables

4.1 Introduction

Categorical variables provide a numerical label for observations that fall in distinct groups, or categories. In Section 4.5, we considered *indicator* variables; these indicate the presence, or absence, of an attribute. As an example, to label a voter's political party affiliation, we might construct a variable to be one if the voter is Democrat, and zero otherwise. Categorical variables provide an extension of the idea of indicator variables. For example, we could get more information about a voter's political party affiliation by investigating the categorical variable constructed so that the variable is one if a voter is Democrat, two if Republican, three if Libertarian, four if Socialist, five if a member of Ross Perot's new United We Stand party, and six otherwise.

For categorical variables, there may or may not be an ordering of the groups. As an example of a categorization that displays ordering, we might consider the age of an apartment building as new, intermediate and old. Any continuous variable, such as age, can be grouped into distinct categories and be treated as a categorical variable. As an example of a categorization that does not display ordering, in the section, we will discuss the type of the car. Another interesting example is the political party affiliation example cited above. For some studies, one might argue that there is an ordering of political philosophies, for example, from strongly conservative to strongly liberal. For other studies, it may be difficult to make this argument. In our treatment, we do not make use of the ordering of categories within a factor. *Factor* is another term used for a categorical, independent variable.

Historically, factors were used primarily to represent grouped continuous variables. By categorizing a continuous variable, the precision of measuring the variable is less of an issue than if an exact measurement was used. Models using only factors became so widely used in experimental research that a separate literature has been developed called the study of *ANOVA models* (for analysis of variance). Leading researchers in statistics, such as R. A. Fisher, realized that ANOVA models could be written as a special case of regression models. However, because of the lack of readily available computing prior to 1960, regression analysis was not a widely used tool and this connection was not appreciated by many researchers.

An important theme of this chapter is that traditional ANOVA models can be expressed using the regression model. A consequence of this theme is that the inferences, introduced in Chapter 3 and further described in Chapter 6, are also available in the ANOVA model set-up. Still, it is also useful to present ANOVA models separately for at least two reasons. First, in the ANOVA context, the formulas for parameter estimates and partitioning the variability can be done directly using only averages and sums of squares. In particular, no matrix manipulations are required. Second, even though we are able to write ANOVA models in the regression set-up, the details can be cumbersome. Interpretations of the estimation formulas and the model coefficients are more intuitive in the ANOVA set-up.

4.2 Regression Using Indicator Variables

The most direct way of handling categorical variables in regression is through the use of indicator variables. A categorical variable with c levels can be represented using c indicator variables, one for each category. In regression analysis with an intercept term, we use c-1 of these indicator variables. The remaining one enters implicitly through the intercept term. Consider the following example.

Illustration 4.1: 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 LN_PRICE, the natural logarithm of the car price, as the response variable of interest. We use H/P, the car's horsepower, as a continuous explanatory variable. Presumably, consumers are willing to pay more for more powerful cars, and H/P is a standard industry measure of a car's power. We also consider CLASS_CD, the car class, where there are $c = 5$ different types of cars. The variable CLASS_CD 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. We know all the possible outcomes of the categorical variable CLASS_CD, so it need not be examined in isolation of the other variables.

TABLE 4.1 Summary Statistics of Each Continuous Variable

	Number	Mean	Median	Standard deviation	Minimum	Maximum
LN_PRICE	173	9.80	9.70	0.60	8.81	11.612
H/P	173	147	134	60	55	400

The next step is to display the distribution of the continuous variables. So that we can use it later in a more complex setting, we now introduce a graphical method for displaying a variable's distribution called the *box plot.* For examining larger data sets, analysts have found this to be a useful graphical form. Figure 4.1 illustrates the box plot for the logarithm of car price. Here the box captures the middle 50% of the data and the so-called "whiskers" capture the middle 80%. By using the box and whiskers to capture the majority of the data, the viewer can get a quick sense of the distribution and important summary statistics without examining individuals observations. (For comparison with a normal curve, you will also see the box and whiskers defined in relation to percentiles from a standard normal curve.)

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

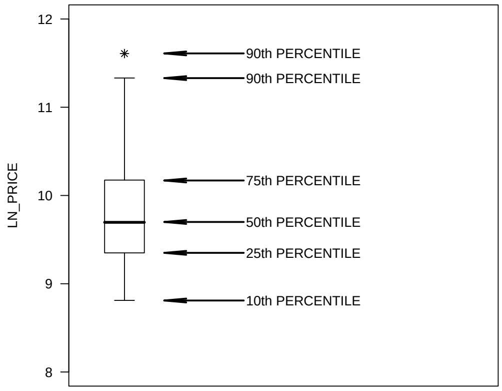

Fig. 4.1. Box plot of car price in logarithmic units. *Source: Motor Trend's 1993 New Car Buyer's Guide*

TABLE 4.2 Summary Statistics of Logarithmic Price By Car Type

	CLASS_CD	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.

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 LN_PRICE and H/P. The correlation coefficient turns out to be 87.2%.

Also in Figure 4.3 you will notice that a letter coding was used as plotting 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 LN_PRICE as the response, H/P as an explanatory variable and four indicator variables of the car class. Here, we define C to be an indicator variable for convertibles so that $C = 1$ if the car

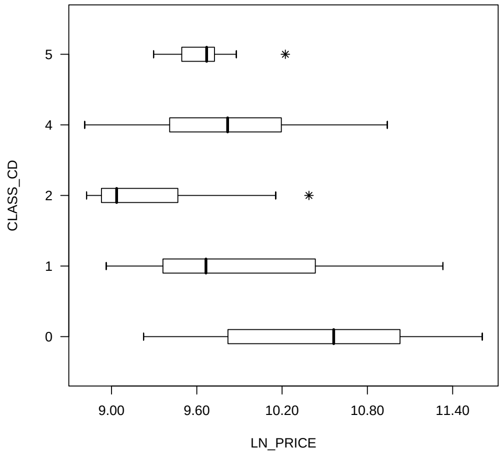

Fig. 4.2. Box plot of Logarithmic price by car type

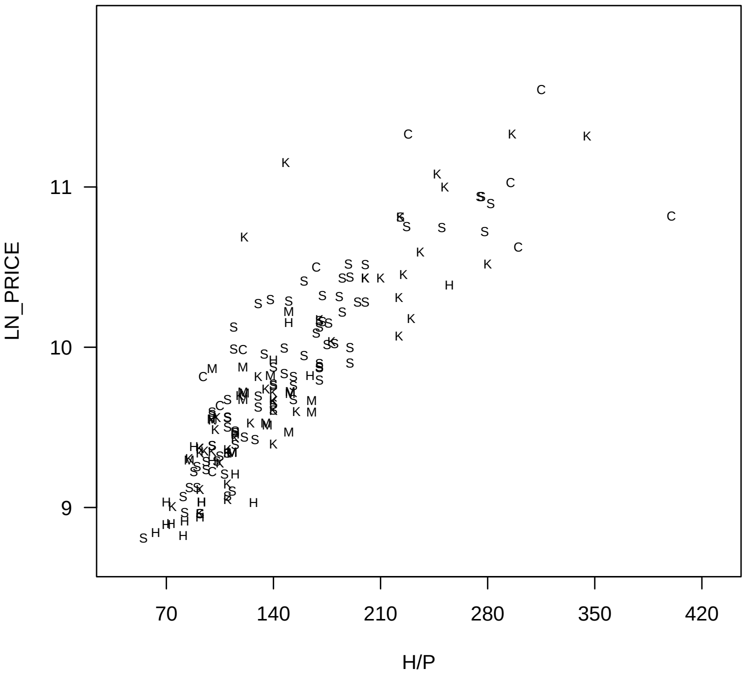

Fig. 4.3. Letter plot of logarithmic prices versus horsepower. Here, the letter codes are 'C' for convertible, 'K' for coupe, 'H' for hatchback, 'S' for sedan and 'M' for mini-van.

is a convertible and $C = 0$ for the other car types. Similarly, define the indicator 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 H/P, 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

H/P show that it is a significant explanatory variable because $t(b_{H/P}) = 21.2$ is very large.

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
H/P	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

DISPLAY 4.3 ANOVA table and coefficient estimates for Example 4.1

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

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

Thus, for example, for a convertible with H/P = 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 H/P = 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.

Similarly, we may interpret the regression coefficient of each indicator variable to be the estimated expected difference between the variable being indicated and the one being dropped. For a variable with c categories, we only use $c - 1$ indicator variables. The reasons will be discussed in greater detail in Section 4.2 and 6.3. Because no assumption is made regarding the ordering of the categories, it does not matter which variable is dropped with regard to the fit of the model. However, as we have seen, it does matter for the interpretation of the regression coefficients.

To illustrate, the regression model was re-run with H/P as a continuous explanatory variable and C, K, S and M as indicator explanatory variables. The analysis of variance table is the same as given in Display 4.3. The coefficient estimates are given in Table 4.4. Unlike Display 4.3, we see that almost all of the t-ratios of the indicator 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 indicator 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 Display 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 4.3 Coefficient Estimates of a Regression of Car Price on Horsepower, and Indicators of Convertible, Coupe, Sedan and Mini-van

Explanatory variable	Coefficient	Standard error	t-ratio
Constant	8.37338	0.07950	104.3
H/P	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 indicator variables are significantly different from the hatchbacks, the omitted binary variable

4.3 One Factor ANOVA Model

To illustrate a one factor ANOVA model, 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, $\ldots$, 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:

$$\begin{array}{lcccc}
\text{Data for DRG 1} & y_{11} & y_{21} & \cdots & y_{n1,1} \\
\text{Data for DRG 2} & y_{12} & y_{22} & \cdots & y_{n2,1} \\
. & . & . & \cdots & . \\
\text{Data for DRG } c & y_{1c} & y_{2c} & \cdots & y_{nc,c}
\end{array}$$

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 $_j$th 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 (see Section 6.6 for more discussion on this point.)

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

When introducing the concept of random error in Section 2.8, we decomposed the response as

$$response(y) = deterministic\ component(\mu) + random\ error(e)$$

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

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

This is short-hand notation for $n_1 + n_2 + \ldots + n_c = n$ equations, one for each observation. The random errors $\{e_{ij}\}$ 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

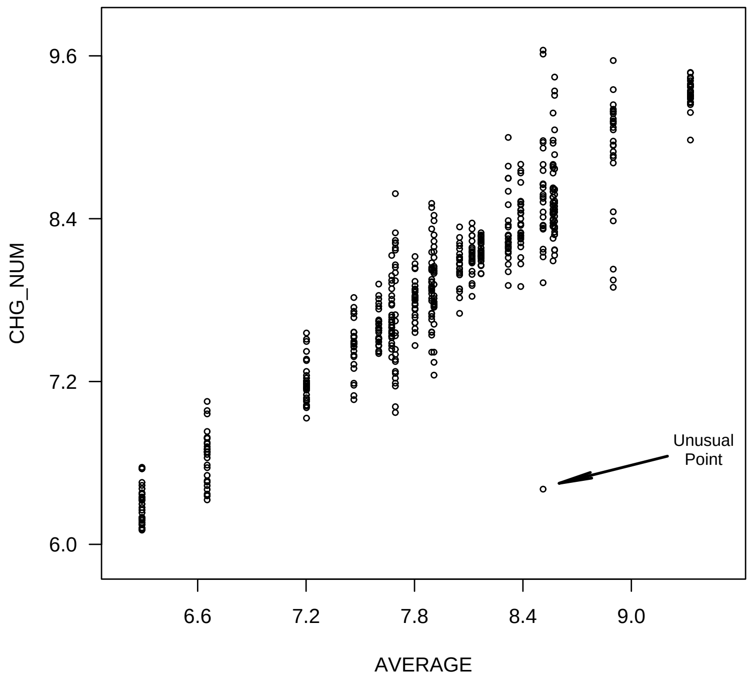

Fig. 4.4. Scatter plot of responses versus average response over diagnosis-related group (DRG). *Source: Wisconsin Department of Health and Human Services.*

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 squares*, introduced in Section 3.1. That is, let be some 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 $\hat{\mu}^*$ that minimizes this sum of squares is $\bar{y}_j$. Thus, yj 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 4.4 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 $s_y^2 = (Total\ SS)/(n-1)$.

Illustration 4.2: Machine Run Times

Before continuing, we consider a small, hypothetical, data set to illustrate the computational issues. Suppose that we have measured the time it takes for three types of machines to run a given benchmark test. The run times are presented in Table 4.4.

TABLE 4.5 Hypothetical Run Times for Three Machines

Machine	Run times	Average run time
1	$14, 12, 10, 12$	$\bar{y}_1 = 12$
2	$9, 16, 15, 12$	$\bar{y}_2 = 13$
3	$8, 10, 7, 7$	$\bar{y}_3 = 8$

Recall that the notation y_{ij} means the ith run from the jth machine. For example, $y_{21} = 12$and $y_{23} = 10$. The average run times are computed as $\bar{y}_j = (\sum_{i=1}^{4} y_{ij})/4$.

Note that Machine 1, with average run time $\bar{y}_3 = 8$, is the fastest. The issue is, based on the data, can we be confident that there is a real difference in machines or could this difference be due to sampling variability, that is, chance?

To this end, we first construct the ANOVA table. You should check that

$$\bar{y} = \frac{\sum_{j=1}^{3}\sum_{i=1}^{4} y_{ij}}{12} = 11 \quad \text{and} \quad \text{Total SS } = \sum_{j=1}^{3}\sum_{i=1}^{4} \frac{(y_{ij}-\bar{y})^2}{(12-1)} = 100.$$

Further, we have

$$\text{Factor SS} = \sum_{j=1}^{3} n_j(\bar{y}_j - \bar{y})^2 = 4(12-11)^2 + 4(13-11)^2 + 4(8-11)^2 = 56$$

.

Table 4.6 summarizes these calculations.

TABLE 4.6 ANOVA Table for Hypothetical Run Times for Three Machines

Source	Sum of Squares	*df*	Mean Square
Machine	56	2	28.00
Error	44	9	4.89
Total	100	11	

From this table, we can compute $R^2 = 56\%$ and $s = (Error\ MS)^{1/2} = (4.89)^{1/2} = 2.21$.

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.

In our machine example in Illustration 4.2, we have realized average run times of $\bar{y}_1 = 12$, $\bar{y}_2 = 13$ and $\bar{y}_3 = 8$. Is there a real difference? We hypothesize H$_0$: $\mu_1 = \mu_2 = \mu_3$, no difference among the true run times. From Table 4.6, we calculate F-ratio = (Factor MS)/(Error MS) = (28.00)/(4.89) = 4.726. Is this large? Looking to the F-table, at the 5% level of significance with $df_1 = c-1 = 2$ and $df_2 = n-c = 9$, we have F-value = 4.256. Because the F-ratio exceeds the F-value, we reject the null hypothesis and declare that there does seem to be a real difference among the run times. The data does not provide us with an indication as to the cause of this difference, only that it is unlikely that the difference can be ascribed to mere sampling variability.

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 4.7 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 = (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 = 10th 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

As described in Section 4.1, an important feature of the 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. The regression model relies on more general, yet more cumbersome, estimation methodologies. However, 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, justifications of the tests and intervals estimates need only be done in the regression case and need not be repeated in the ANOVA context. Further, the remedies for model inadequacy that we will present in Chapter 5 and the additional inference techniques in Chapter 6 will be available for both the regression and ANOVA models.

To this end, for a categorical variable with c levels, define c indicator variables, x_1, $x_2, \ldots, x_c$. Here, x_1 is a one if the observation falls in the first level and is zero otherwise. Similarly, x_2 is an indicator variable for an observation falling in the second level, and so on. Thus, x_j indicates whether or not an observation falls in the jth level.

With these variables, we can rewrite our one factor ANOVA model

$$y = \mu_j + e$$

as

$$y = \mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_c x_c + e. \tag{4.1}$$

The regression model in equation (4.1) 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 μ_j=τj+μ in equation (4.1), we have

$$y = \mu + \tau_1 x_1 + \tau_2 x_2 + \ldots + \tau_c x_c + e. \tag{4.2}$$

Thus, we have re-written the model into what appears to be our usual regression format, as in equation (4.2).

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 (4.2) 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 (4.2) is

$$y_{ij} = \mu + \tau_j + e_{ij}. \tag{4.3}$$

Comparing equations (4.1) and (4.2), we see that the number of parameters has increased by one. That is, in equation (4.1), there are c parameters, $\mu_1, \ldots, \mu_c$, even though in equation (4.2) there are $c+1$ parameters, μ and $\tau_1, \ldots, \tau_c$. The model in equation (4.2) is said to be *overparameterized.* To make these two expressions equivalent, we now present two ways of *restricting* the movement of the parameters in (4.2).

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

$$y = \mu + \tau_1 x_1 + \tau_2 x_2 + \ldots + \tau_{c-1} x_{c-1} + e, \tag{4.4}$$

dropping x_c. With this formulation, it is easy to fit the model in equation (4.4) 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 (4.1) and (4.2), 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

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

To illustrate, consider the hypothetical machine run time data described in Illustration 4.1. The estimate of the mean levels are $\hat{\mu}_1 = \bar{y}_1 = 12$,

$\hat{\mu}_2 = \bar{y}_2 = 13$ and $\hat{\mu}_3 = \bar{y}_3 = 8$. To apply the first restriction, we could drop, for instance, the third category. The resulting regression equation would be E $y = \mu + \tau_1 x_1 + \tau_2 x_2 = \mu_3 + (\mu_1 - \mu_3) x_1 + (\mu_2 - \mu_3) x_2$. The corresponding fitted equation would be

$$\hat{y} = 8 + 4x_1 + 5x_2.$$

If you have a regression package available, input the data in Table 4.5 and verify that this is the correct fitted equation.

The second restriction yields $\mu \equiv (1/n)\sum_{j=1}^{c} n_j \mu_{_j} = (1/c)\sum_{j=1}^{c} \mu_j$, because there are an equal number of observations in each cell. Here, c is 3, so $\mu = (\mu_1+\mu_2+\mu_3)/3$. The estimate is $\hat{\mu} = \bar{y} = 11$. The estimated differences are

$$\hat{\tau}_1 = \bar{y}_1 - \bar{y} = 12 - 11 = 1, \quad \hat{\tau}_2 = \bar{y}_2 - \bar{y} = 13 - 11 = 2 \quad \textit{and}$$

$$\hat{\tau}_3 = \bar{y}_3 - \bar{y} = 8 - 11 = -3.$$

The model remains the same regardless of our choice of restriction on parameters. Because the model is the same, all predictions and other inferences are the same. To illustrate, suppose that we are interested in getting a point estimate for the first machine. For an observation at the first level, we have $x_1 = 1$ and $x_2 = 0$. Thus, using the estimated regression equation from our first restriction, we have $\hat{y} = 8 + 4(1) + 5(0) = 12$. Using the model express with the second restriction, our fitted value is $\hat{\mu} + \hat{\tau}_1 = 11 + 1 = 12$. The two methods produce equivalent fitted values, as anticipated.

4.4 Two Factor ANOVA Model

Suppose that we now wish to consider two independent categorical variables, or factors. To be specific, we again consider the hypothetical machine run time example introduced in Illustration 4.1. In this example, the response of interest is the length of time it takes a machine to complete a certain benchmark test. The explanatory variable considered above was the type of machine. Suppose that we have now additional information available on another factor, 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 4.8 presents the data for the run times.

TABLE 4.8 Hypothetical Run Times of Three Machines by Two Operators

	Operator (Factor 1)		
Machine (Factor 2)	Rookie	Experiened	Machine averages
1	14, 12	10, 12	$\bar{y}_{\cdot 1 \cdot} = 12$
2	15, 16	9, 12	$\bar{y}_{\cdot 2 \cdot} = 13$
3	8, 10	7, 7	$\bar{y}_{\cdot 3 \cdot} = 8$

Operator Avearge	$\bar{y}_{1 \cdot\cdot} = 12.5$	$\bar{y}_{2 \cdot\cdot} = 9.5$	$\bar{y}_{3 \cdot\cdot} = 11$

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 $k = 1, \ldots, K$. Extending the example in Section 4.3, 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

$$\bar{y}_{ij \cdot} = \frac{1}{K} \sum_{k=1}^{K} y_{ijk}.$$

Based on an examination of the data in Table 4.8, it appears 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 (4.3), 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

$$y_{ijk} = \mu + \beta_i + \tau_j + e_{ijk} \qquad (4.5)$$

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

$$\beta_1 + \beta_2 + \ldots \ + \beta I = 0 and \tau 1 + \tau 2 + \ldots \ + \tau c = 0. \qquad (4.6)$$

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 (4.6), the least squares estimates are

$$\hat{\mu} = \bar{y}_{\cdots}, \quad \hat{\beta}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}, \quad \text{and } \hat{\tau}_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots} \ . \tag{4.7}$$

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

$$\begin{aligned} \text{Error SS} &= SS(\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_I, \hat{\tau}_1, \ldots, \hat{\tau}_c) = \sum_{i=1}^{I} \sum_{j=1}^{c} \sum_{k=1}^{K} (y_{ijk} - (\hat{\mu} + \hat{\beta}_i + \hat{\tau}_j))^2 \quad (4.8) \\ &= \sum_{i=1}^{I} \sum_{j=1}^{c} \sum_{k=1}^{K} (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2 \end{aligned}$$

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

$$\begin{array}{cccc} y_{ijk} - \bar{y}_{\cdots} = & (\bar{y}_{\cdot i\cdot} - \bar{y}_{\cdots}) + & (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) + & (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}). \\ (1) & (2) & (3) & (4) \end{array} \tag{4.9}$$

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 (4.9) 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 (4.8) and, with equation (4.7),

$$\begin{aligned} \text{Total SS} &= \sum_{i=1}^{I}\sum_{j=1}^{c}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{...})^2, \qquad (4.10a) \\ \text{Factor 1 SS} &= cK\sum_{i=1}^{I}(\bar{y}_{j..} - \bar{y}_{...})^2 = cK\sum_{i=1}^{I}\hat{\beta}_i^2 \\ \text{Factor 2 SS} &= IK\sum_{j=1}^{c}(\bar{y}_{.j.} - \bar{y}_{...})^2 = IK\sum_{j=1}^{c}\hat{\tau}_j^2. \end{aligned}$$

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 ≡ (Factor 1 SS)/(I-1), Factor 2 MS ≡ (Factor 2 SS)/(c-1) and Error MS ≡ (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 4.9 presents results for the data in Table 4.8.

TABLE 4.9 ANOVA Table for Two Factor Additive Model of Hypothetical Run Times

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

As before, tests of hy

potheses 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 the F-table. From Table 4.9, we have F-ratio $= 27/2.12 = 12.74$. Because $12.74 = F\text{-ratio} > F\text{-value} = 5.318$, 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 4.8.

The test of hypothesis for differences among levels of Factor 2 is similar. To summarize, consider the Table 4.10.

TABLE 4.10 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.

Recall that this was the same decision made in Section 4.3 when we examined only one factor. The advantage of introducing a second explanatory factor is that we have significantly reduced our estimate of the variability, s^2. Thus, we can be more confident in the decisions that we make.

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

$$y_{ijk} = \mu_{ij} + e_{ijk}. \tag{4.11}$$

Here, μ_{ij} represents the mean response for the ith level of Factor 1 and the jth level of Factor 2. As with equation (4.3), 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 (4.11), we get

$$y_{ijk} = \mu + \beta_i + \tau_j + (\beta\tau)_{ij} + e_{ijk}. \tag{4.12}$$

When comparing equations (4.11) and (4.12), we see that there are Ic linear parameters in equation (4.11) even though there are $1 + I + c + Ic$ parameters in equation (4.12). As before, certain restrictions need to be imposed on the parameters in equation (4.12) 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 (4.6). The least squares estimate of $(\beta\tau)ij$ turns out to be $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}$. 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 (4.10) and

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

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

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 Source, is the number of observations, n, minus the number of free parameter, 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 4.8. Table 4.11 presents the analysis of variance for the two factor interaction model fit of this example.

TABLE 4.11 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 the 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 4.10 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 indicator variables. To this end, for the I levels of Factor 1, define $x_{1,1}$ to be a one if the observation falls in the first level of Factor one and is zero otherwise. Similarly, $x_{1,2}$ is an indicator variable for an observation fall in the second level of Factor 1, and so on up to $x_{1,I}$, an indicator for the Ith level of Factor 1. Thus, we define $x_{1,i}$ to be an indicator of the ith level of Factor 1. 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 (4.5) as

$$y = \mu + \sum_{i=1}^{I} \beta_i x_{1,i} + \sum_{j=1}^{c} \tau_j x_{2,j} + e. \tag{4.13}$$

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 (4.13) reduces to $y_{23,k} = \mu + \beta_3 + \tau_4 + e_{34,k}$, as in equation (4.5).

As with equation (4.5), 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

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

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 (4.12) as

$$y = \mu + \sum_{i=1}^{I} \beta_i x1, i + \sum_{j=1}^{c} \tau_j x_{2,j} + \sum_{i=1}^{I} \sum_{j=1}^{c} (\beta\tau)_{ij} x_{1,i} x_{2,j} + e. \tag{4.14}$$

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

$$y = \mu + \sum_{i=1}^{I-1} \beta_i x1, i + \sum_{j=1}^{c-1} \tau_j x_{2,j} + \sum_{i=1}^{I-1} \sum_{j=1}^{c-1} (\beta\tau)_{ij} x_{1,i} x_{2,j} + e. \tag{4.15}$$

Again, equation (4.14) must be estimated using restricted parameters. Equation (4.15) provides an equivalent formulation without the need to restrict the parameters. To illustrate equation (4.15), consider our machine example with $I = 2$ and $c = 3$. In this case, equation (4.15) 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.

4.5 Regression using Categorical and Continuous Variables

In Section 4.4, we introduced two ways of combining two categorical variables, additive models and interaction models. In Section 4.5, several ways of combining continuous variables were presented. In that section, we also discussed ways of modelling combinations of indicator and continuous variables. In this section, we extend that discussion by presenting ways of combining categorical and continuous 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 One Categorical and One Continuous Variable

Combining categorical and continuous variables models begins with separate models for each variable. In Section 4.3 is a discussion of the one factor model:

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

In Chapter 2 is a discussion of the one continuous variable, or covariate, model:

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

Table 4.12 describes several models that could be used to represent combinations of a factor and covariate.

TABLE 4.12 Several Models that Represent Combinations of One Factor and One Covariate

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

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. 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 through the variable slope but there is no main effect, as represented by the constant intercept. Figures 4.5 through 4.7 illustrate the expected responses of these models.

For each model presented in Table 4.12, parameter estimates can be calculated using the method of least squares. As usual, this means writing the expected response, E 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. It turns out, for the regression model with variable intercept and constant slope, that the least squares estimates can be expressed compactly as:

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

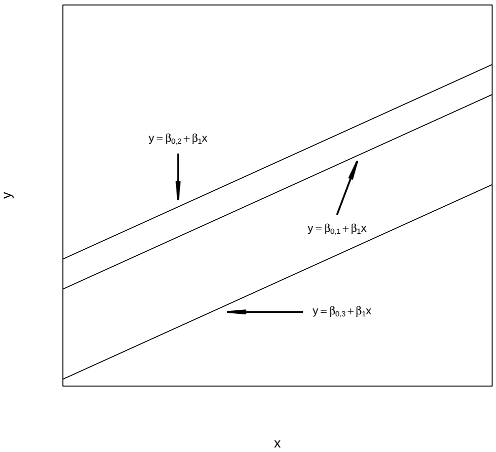

Fig. 4.5. Plot of the expected response versus the covariate for the regression model with variable intercept and constant slope.

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,

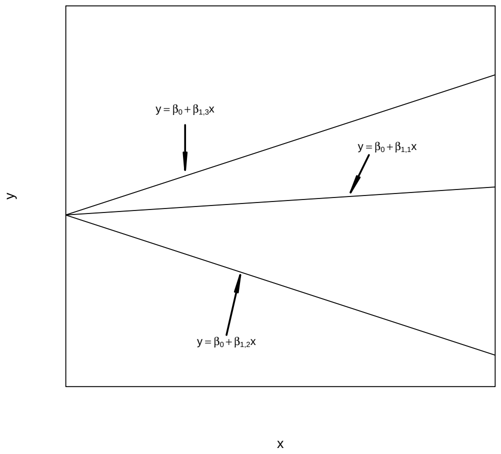

Fig. 4.6. Plot of the expected response versus the covariate for the regression model with constant intercept and variable slope.

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

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 $\hat{y}_{ij} = b_{0,j} + b_1 x_{ij}$.

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

The covariate, x, that we consider 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) and our measure of hospital utilization is the logarithm of the number of discharges (x). The scatter plot in Figure 4.8 gives a preliminary idea of the relationship between y and x. Here, we see the unusual point in the lower left hand part of the plot, corresponding to an observation with a small number of patients discharged. We also note what 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 log discharges x, in Figure 4.9 is a scatterplot 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 the scatterplot in Figure 4.8.

The misleading results produced by omitting categorical variables is sometimes referred to as a problem of *aggregation of data.* The idea here is that we should be analyzing the data at the DRG level as in Figure 4.9. When we omit this factor and consider all levels of DRG simultaneously as in Figure 4.8, we look at the less complex, more "aggregate" levels of the data. As we have seen in this example, this may produce misleading results.

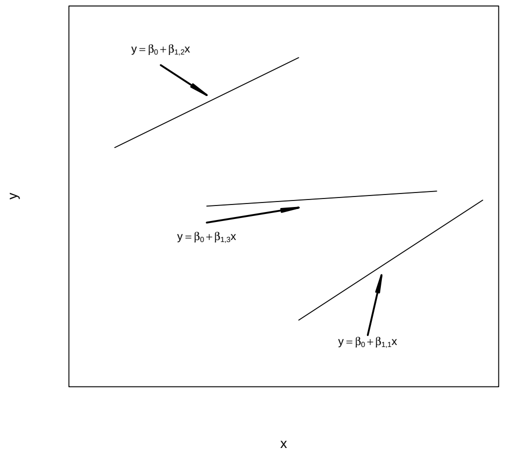
$y = \beta_0 + \beta_{1,2}x$
$y = \beta_0 + \beta_{1,3}x$
$y = \beta_0 + \beta_{1,1}x$
y
x

Fig. 4.7. Scatter plot of natural logarithm of cost per discharge versus natural logarithm of the number of discharges.

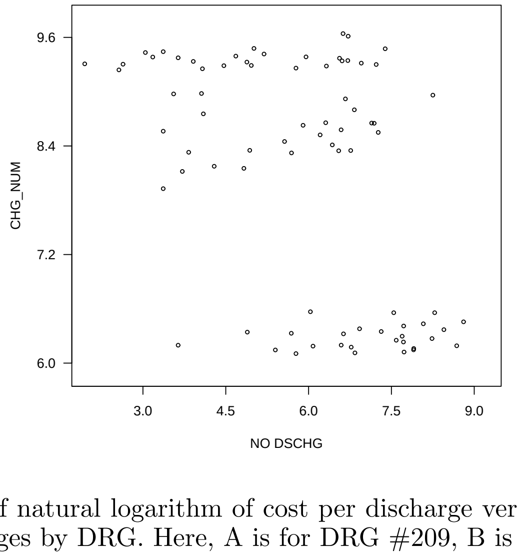

Fig. 4.8. Letter plot of natural logarithm of cost per discharge versus natural logarithm of the number of discharges by DRG. Here, A is for DRG #209, B is for DRG #391 and C is for DRG #430.

Each of the five models defined in Table 4.12 was fit to this subset of the Hospital case study. The summary statistics are in Table 4.13. 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 and the error degrees of freedom is the number of observations minus the number of model parameters. The error sum of squares is defined above. The coefficient of determination is one minus the ratio of the error sum of squares to the total of squares and the error mean squares is the error sum of squares divided by the error degrees of freedom.

TABLE 4.13 Degree of Freedom and Error Sum of Squares of Several Models to Represent One Factor and One Covariate for the DRG Example

	Model	Error		Coefficient of	
	degrees	degrees	Error Sum	determination	Error
Model Description	of freedom	of freedom	of Squares	(%)	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 constantslope	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

Using indicator variables, each of the models in Table 4.12 can be written in a regression format. Thus, we may interpret the summary statistics in Table 4.13 using the same principles that we introduced in the regression context, in Chapter 3. For example, when selecting the best model, we interpret the coefficient of determination, R^2, to be the proportion of variability explained by the model. Thus, we look for models with a high R^2. However, an algebraic fact shows that R^2 can always be increased by adding a variable to the models. The error mean square, Error MS, compensates for this. The Error MS $= s^2$ is our estimate of overall variability in the model that we would like to be as small as possible. Thus, we look for models with low Error MS.

As we have seen in Section 4.6, 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. That is, we can examine entire portions of the model to see if they should be included in our model specification. Recall that, when examining portions of a model, we call the larger model under consideration the *full* model. Thus, denote (Error SS)$_{full}$ and (df)$_{full}$ to be the error sum of squares and error degrees of freedom calculated using this model. The subset of this model under consideration is called the *reduced* model. Similarly, denote (Error SS)$_{reduced}$ and (df)$_{reduced}$ to be the error sum of squares and error degrees of freedom calculated using this model. Because the reduced model is a subset of the full model, we know that both the error sum of squares and the degrees of freedom are larger for the reduced model. We examine the test statistic

$$F\text{-ratio} = \frac{\frac{(\text{Error SS})_{reduced} - (\text{Error SS})_{full}}{(df)_{reduced} - (df)_{full}}}{\frac{(\text{Error SS})_{full}}{(df)_{full}}}$$

to see if the drop in the error sum of squares is significant. We reject the null hypothesis H_0: Reduced Model is valid in favor of the alternative hypothesis H_a: Full Model is valid if F-ratio ¿ F-value, where the F-value is a specified percentile from an F-distribution with $df1 = (df)_{reduced} - (df)_{full}$ and $df_2 = (df)_{full}$.

To illustrate this testing procedure with our DRG example, consider the summary statistics for several models presented in Table 4.13. From this table and the associated scatter plots, it seems clear that the DRG factor is important. Further, a t-test, not presented here, shows that the covariate x 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?

To this end, from Table 4.13, for the regression model with variable intercept and slope, we have $(\text{Error SS})_{full} = 5.458$ and $(\text{df})_{full} = 73$. For the regression model with variable intercept and constant slope, we have $(\text{Error SS})_{reduced} = 7.482$ and $(\text{df})_{reduced} = 75$. Thus, our test statistic is

$$F\text{-ratio} = \frac{\frac{7.482-5.458}{75-73}}{\frac{5.458}{73}} = 13.535$$

The 95th percentile from an F-distribution with $df_1 = (df)_{reduced} - (df)_{full} = 75 - 73 = 2$ 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. (Using the data sets available with the book, the reader may find it interesting to perform this test after omitting the outlying point noted above.)

General Linear Model

In Section 4.2, 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 (4.2), the method of least squares would not have arrived at a unique set of regression coefficient estimates. The reason is that, in equation (4.2), 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*,

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e, \tag{4.16}$$

where $\{e_i\}$ 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 (4.2) 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 4.6 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.

4.6 Summary

Chapter 3 introduced the multiple linear regression model, showed how to estimate the model parameters, and provided basic inference results. Chapter 4 extends this introduction by showing how to use categorical independent variables in the regression context. Section 4.2 was a direct extension, by showing how to use indicator variables to represent categorical variables in the regression context.

An important theme of this chapter is that traditional ANOVA models can be expressed using the regression model. To illustrate this theme, in Section 4.3 we considered the ANOVA models using only one factor. Many of the details concerning analysis of data, the assumptions of the model and some of the inferences that can be made using the model were presented. Further, the link between the ANOVA and regressions set-ups were constructed in detail. By showing that ANOVA models are a special case of the regression set-up, no new theory was needed to represent ANOVA models. In Section 4.4 we considered the ANOVA model using two factors. Here, the treatment was briefer than in Section 4.3, with only the important highlights underscored. The focus was mainly on the different ways that one can combine two categorical variables. We followed the same pattern as in Chapter 4: we first introduced additive models and then added an additional level of complexity, the interaction terms. In Section 4.5 we considered the general case, called the *general linear model.* The general linear model encompasses not only categorical variables, but also continuous variables and combinations of the two types of variables. In this section, we presented an important special case, combining one categorical and one continuous variable.

As in Chapter 3, in Chapter 4 only passing references are made to issues of model selection. We take up this important topic in Chapter 5.

Key Words, Phrases and Symbols - Chapter 4

After reading this chapter, you should be able to define each of the following important terms, phrases and symbols in your own words. If not, go to the page indicated and review the definition.

Insert Vocabulary here