Chapter 11

Analysis of Variance and Regression

"I've wasted time enough," said Lestrade rising. "I believe in hard work and not in sitting by the fire spinning fine theories."

Inspector Lestrade
The Adventure of the Noble Bachelor

11.1 Introduction

Up until now, we have modeled a random variable with a pdf or pmf that depended on parameters to be estimated. In many situations, some of which follow, a random variable can be modeled not only with unknown parameters but also with known (and sometimes controllable) covariates. This chapter describes the methodologies of analysis of variance (ANOVA) and regression analysis. They are based on an underlying assumption of a linear relationship and form a large core of the statistical methods that are used in practice.

The analysis of variance (commonly referred to as the ANOVA) is one of the most widely used statistical techniques. A basic idea of the ANOVA, that of partitioning variation, is a fundamental idea of experimental statistics. The ANOVA belies its name in that it is not concerned with analyzing variances but rather with analyzing *variation in means*.

We will study a common type of ANOVA, the oneway ANOVA. For a thorough treatment of the different facets of ANOVA designs, there is the classic text by Cochran and Cox (1957) or the more modern, but still somewhat classic, treatments by Dean and Voss (1999) and Kuehl (2000). The text by Neter, Wasserman, and Whitmore (1993) provides a guide to overall strategies in experimental statistics.

The technique of regression, in particular linear regression, probably wins the prize as the most popular statistical tool. There are all forms of regression: linear, nonlinear, simple, multiple, parametric, nonparametric, etc. In this chapter we will look at the simplest case, linear regression with one predictor variable. (This is usually called *simple* linear regression, as opposed to *multiple* linear regression, which deals with many predictor variables.)

A major purpose of regression is to explore the dependence of one variable on others. In simple linear regression, the mean of a random variable, Y, is modeled as a function of another observable variable, x, by the relationship $\mathrm{E}Y = \alpha + \beta x$. In general, the function that gives $\mathrm{E}Y$ as a function of x is called the *population regression function*.

Good overall references for regression models are Christensen (1996) and Draper and Smith (1998). A more theoretical treatment is given in Stuart, Ord, and Arnold (1999, Chapter 27).

11.2 Oneway Analysis of Variance

In its simplest form, the ANOVA is a method of estimating the means of several populations, populations often assumed to be normally distributed. The heart of the ANOVA, however, lies in the topic of statistical design. How can we get the most information on the most populations with the fewest observations? The ANOVA design question is not our major concern, however; we will be concerned with inference, that is, with estimation and testing, in the ANOVA.

Classic ANOVA had testing as its main goal—in particular, testing what is known as "the ANOVA null hypothesis." But more recently, especially in the light of greater computing power, experimenters have realized that testing one hypothesis (a somewhat ludicrous one at that, as we shall see) does not make for good experimental inference. Thus, although we will derive the test of the ANOVA null, it is far from the most important part of an analysis of variance. More important is estimation, both point and interval. In particular, inference based on *contrasts* (to be defined) is of major importance.

In the oneway analysis of variance (also known as the oneway classification) we assume that data, Y_{ij}, are observed according to a model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \tag{11.2.1}$$

where the θ_i are unknown parameters and the ϵ_{ij} are error random variables.

Example 11.2.1 (Oneway ANOVA) Schematically, the data, y_{ij}, from a oneway ANOVA will look like this:

Treatments				
1	2	3	...	k
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	y_{k3}
		y_{3n_3}		$\vdots$
y_{1n_1}				
	y_{2n_2}			y_{kn_k}

Note that we do not assume that there are equal numbers of observations in each treatment group.

As an example, consider the following experiment performed to assess the relative effects of three toxins and a control on the liver of a certain species of trout. The data are the amounts of deterioration (in standard units) of the liver in each sacrificed fish.

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
	31	24	
	34		

Without loss of generality we can assume that $E\epsilon_{ij} = 0$, since if not, we can rescale the ϵ_{ij} and absorb the leftover mean into θ_i. Thus it follows that

$$EY_{ij} = \theta_i, \quad j = 1, \ldots, n_i,$$

so the θ_is are the means of the Y_{ij}s. The θ_is are usually referred to as *treatment means*, since the index often corresponds to different treatments or to *levels* of a particular treatment, such as dosage levels of a particular drug.

There is an alternative model to (11.2.1), sometimes called the *overparameterized model*, which can be written as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \tag{11.2.2}$$

where, again, $E\epsilon_{ij} = 0$. It follows from this model that

$$EY_{ij} = \mu + \tau_i.$$

In this formulation we think of μ as a grand mean, that is, the common mean level of the treatments. The parameters τ_i then denote the unique effect due to treatment i, the deviation from the mean level that is caused by the treatment. However, we cannot estimate both τ_i and μ separately, because there are problems with *identifiability*.

Definition 11.2.2 A parameter θ for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is *identifiable* if distinct values of θ correspond to distinct pdfs or pmfs. That is, if $\theta \neq \theta'$, then $f(x|\theta)$ is not the same function of x as $f(x|\theta')$.

Identifiability is a property of the model, not of an estimator or estimation procedure. However, if the model is not identifiable, then there is difficulty in doing inference. For example, if $f(x|\theta) = f(x|\theta')$, then observations from both distributions look exactly the same and we would have no way of knowing whether the true value of the parameter was θ or θ'. In particular, both θ and θ' would give the likelihood function the same value.

Realize that problems with identifiability can usually be solved by redefining the model. One reason that we have not encountered identifiability problems before is that our models have not only made intuitive sense but also were identifiable (for example, modeling a normal population in terms of its mean and variance). Here, however, we have a model, (11.2.2), that makes intuitive sense but is not identifiable. In Chapter 12 we will see a parameterization of the bivariate normal distribution that models a situation well but is not identifiable.

In the parameterization of (11.2.2), there are $k+1$ parameters, $(\mu, \tau_1, \ldots, \tau_k)$, but only k means, $EY_{ij}, i = 1, \ldots, k$. Without any further restriction on the parameters, more than one set of values for $(\mu, \tau_1, \ldots, \tau_k)$ will lead to the same distribution. It is common in this model to add the restriction that $\sum_{i=1}^{k} \tau_i = 0$, which effectively reduces the number of parameters to k and makes the model identifiable. The restriction also has the effect of giving the τ_is an interpretation as deviations from an overall mean level. (See Exercise 11.5.)

For the oneway ANOVA the model (11.2.1), the *cell means model*, which has a more straightforward interpretation, is the one that we prefer to use. In more complicated ANOVAs, however, there is sometimes an interpretive advantage in model (11.2.2).

11.2.1 Model and Distribution Assumptions

Under model (11.2.1), a minimum assumption that is needed before any estimation can be done is that $E\epsilon_{ij} = 0$ and $\operatorname{Var} \epsilon_{ij} < \infty$ for all i, j. Under these assumptions, we can do some estimation of the θ_is (as in Exercise 7.41). However, to do any confidence interval estimation or testing, we need distributional assumptions. Here are the classic ANOVA assumptions.

Oneway ANOVA assumptions

Random variables Y_{ij} are observed according to the model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,$$

where

(i) $E\epsilon_{ij} = 0, \operatorname{Var} \epsilon_{ij} = \sigma_i^2 < \infty$, for all i, j. $\operatorname{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ for all i, i', j, and j' unless $i = i'$ and $j = j'$.

(ii) The ϵ_{ij} are independent and normally distributed (normal errors).

(iii) $\sigma_i^2 = \sigma^2$ for all i (equality of variance, also known as *homoscedasticity*).

Without assumption (ii) we could do only point estimation and possibly look for estimators that minimize variance within a class, but we could not do interval estimation or testing. If we assume some distribution other than normal, intervals and tests can be quite difficult (but still possible) to derive. Of course, with reasonable sample sizes and populations that are not too asymmetric, we have the Central Limit Theorem (CLT) to rely on.

The equality of variance assumption is also quite important. Interestingly, its importance is linked to the normality assumption. In general, if it is suspected that the data badly violate the ANOVA assumptions, a first course of attack is usually to try to transform the data nonlinearly. This is done as an attempt to more closely satisfy the ANOVA assumptions, a generally easier alternative than finding another model for the untransformed data. A number of common transformations can be found in Snedecor and Cochran (1989); also see Exercises 11.1 and 11.2. (Other research on transformations has been concerned with the Box–Cox family of power transformations. See Exercise 11.3.)

The classic paper of Box (1954) shows that the robustness of the ANOVA to the assumption of normality depends on how equal the variances are (equal being better). The problem of estimating means when variances are unequal, known as the Behrens–Fisher problem, has a rich statistical history which can be traced back to Fisher (1935, 1939). A full account of the Behrens–Fisher problem can be found in Stuart, Ord, and Arnold (1999).

For the remainder of this chapter we will do what is done in most of the experimental situations and we will assume that the three classic assumptions hold. If the data are such that transformations and the CLT are needed, we assume that such measures have been taken.

11.2.2 The Classic ANOVA Hypothesis

The classic ANOVA test is a test of the null hypothesis

$$H_0: \quad \theta_1 = \theta_2 = \cdots = \theta_k,$$

a hypothesis that, in many cases, is silly, uninteresting, and not true. An experimenter would not usually believe that the different treatments have *exactly* the same mean. More reasonably, an experiment is done to find out which treatments are better (for example, have a higher mean), and the real interest in the ANOVA is not in testing but in estimation. (There are some specialized situations where there is interest in the ANOVA null in its own right.) Most situations are like the following.

Example 11.2.3 (The ANOVA hypothesis) The ANOVA evolved as a method of analyzing agricultural experiments. For example, in a study of the effect of various fertilizers on the zinc content of spinach plants (y_{ij}), five treatments are investigated. Each treatment consists of a mixture of fertilizer material (magnesium, potassium, and zinc) and the data look like the layout of Example 11.2.1. The five treatments, in pounds per acre, are

Treatment	Magnesium	Potassium	Zinc
1	0	0	0
2	0	200	0
3	50	200	0
4	200	200	0
5	0	200	15

The classic ANOVA null hypothesis is really of no interest since the experimenter is sure that the different fertilizer mixtures have some different effects. The interest is in quantifying these effects. ||

We will spend some time with the ANOVA null but mostly use it as a means to an end. Recall the connection between testing and interval estimation established in Chapter 9. By using this connection, we can derive confidence regions by deriving, then inverting, appropriate tests (an easier route here).

The alternative to the ANOVA null is simply that the means are not all equal; that is, we test

$$(11.2.3) \qquad H_0: \quad \theta_1 = \theta_2 = \cdots = \theta_k \qquad \text{versus} \qquad H_1: \quad \theta_i \neq \theta_j, \text{ for some } i, j.$$

Equivalently, we can specify H_1 as H_1: not H_0. Realize that if H_0 is rejected, we can conclude only that there is *some* difference in the θ_is, but we can make no inference as to where this difference might be. (Note that if H_1 is accepted, we are *not* saying that all of the θ_is are different, merely that at least two are.)

One problem with the ANOVA hypotheses, a problem shared by many multivariate hypotheses, is that the interpretation of the hypotheses is not easy. What would be more useful, rather than concluding just that some θ_is are different, is a statistical description of the θ_is. Such a description can be obtained by breaking down the ANOVA hypotheses into smaller, more easily describable pieces.

We have already encountered methods for breaking down complicated hypotheses into smaller, more easily understood pieces—the union–intersection and intersection–union methods of Chapter 8. For the ANOVA, the union–intersection method is best suited, as the ANOVA null is the intersection of more easily understood univariate hypotheses, hypotheses expressed in terms of *contrasts*. Furthermore, in the cases we will consider, the resulting tests based on the union–intersection method are identical to LRTs (see Exercise 11.13). Hence, they enjoy all the properties of likelihood tests.

Definition 11.2.4 Let $\mathbf{t} = (t_1, \ldots, t_k)$ be a set of variables, either parameters or statistics, and let $\mathbf{a} = (a_1, \ldots, a_k)$ be known constants. The function

$$(11.2.4) \qquad \sum_{i=1}^{k} a_i t_i$$

is called a *linear combination* of the t_is. If, furthermore, $\sum a_i = 0$, it is called a *contrast*.

Contrasts are important because they can be used to compare treatment means. For example, if we have means $\theta_1, \ldots, \theta_k$ and constants $\mathbf{a} = (1, -1, 0, \ldots, 0)$, then

$$\sum_{i=1}^{k} a_i \theta_i = \theta_1 - \theta_2$$

is a contrast that compares θ_1 to θ_2. (See Exercise 11.10 for more about contrasts.)

The power of the union–intersection approach is increased understanding. The individual null hypotheses, of which the ANOVA null hypothesis is the intersection, are quite easy to visualize.

Theorem 11.2.5 *Let $\theta = (\theta_1, \ldots, \theta_k)$ be arbitrary parameters. Then*

$$\theta_1 = \theta_2 = \cdots = \theta_k \Leftrightarrow \sum_{i=1}^{k} a_i \theta_i = 0 \quad \textit{for all } \mathbf{a} \in \mathcal{A},$$

where $\mathcal{A}$ is the set of constants satisfying $\mathcal{A} = \{\mathbf{a} = (a_1, \ldots, a_k) : \sum a_i = 0\}$; that is, all contrasts must satisfy $\sum a_i \theta_i = 0$.

Proof: If $\theta_1 = \cdots = \theta_k = \theta$, then

$$\sum_{i=1}^{k} a_i\theta_i = \sum_{i=1}^{k} a_i\theta = \theta\sum_{i=1}^{k} a_i = 0, \quad \text{(because } \mathbf{a} \text{ satisfies } \textstyle\sum a_i = 0\text{)}$$

proving one implication ($\Rightarrow$). To prove the other implication, consider the set of $\mathbf{a}_i \in \mathcal{A}$ given by

$$\mathbf{a}_1 = (1,-1,0,\ldots,0), \quad \mathbf{a}_2 = (0,1,-1,0,\ldots,0), \quad \ldots, \quad \mathbf{a}_{k-1} = (0,\ldots,0,1,-1).$$

(The set $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1})$ *spans* the elements of $\mathcal{A}$. That is, any $\mathbf{a} \in \mathcal{A}$ can be written as a linear combination of $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1})$.) Forming contrasts with these $\mathbf{a}_i$s, we get that

$$\mathbf{a}_1 \Rightarrow \theta_1 = \theta_2, \quad \mathbf{a}_2 \Rightarrow \theta_2 = \theta_3, \quad \ldots, \quad \mathbf{a}_{k-1} \Rightarrow \theta_{k-1} = \theta_k,$$

which, taken together, imply that $\theta_1 = \cdots = \theta_k$, proving the theorem. □

It immediately follows from Theorem 11.2.5 that the ANOVA null can be expressed as a hypothesis about contrasts. That is, the null hypothesis is true if and only if the hypothesis

$$H_0\colon \quad \sum_{i=1}^{k} a_i\theta_i = 0 \quad \text{for all } (a_1,\ldots,a_k) \text{ such that } \sum_{i=1}^{k} a_i = 0$$

is true. Moreover, if H_0 is false, we now know that there must be at least one nonzero contrast. That is, the ANOVA alternative, H_1: not all θ_is equal, is equivalent to the alternative

$$H_1\colon \quad \sum_{i=1}^{k} a_i\theta_i \neq 0 \quad \text{for some } (a_1,\ldots,a_k) \text{ such that } \sum_{i=1}^{k} a_i = 0.$$

Thus, we have gained in that the use of contrasts leaves us with hypotheses that are a little easier to understand and perhaps are a little easier to interpret. The real gain, however, is that the use of contrasts now allows us to think and operate in a univariate manner.

11.2.3 Inferences Regarding Linear Combinations of Means

Linear combinations, in particular contrasts, play an extremely important role in the analysis of variance. Through understanding and analyzing the contrasts, we can make meaningful inferences about the θ_is. In the previous section we showed that the ANOVA null is really a statement about contrasts. In fact, most interesting inferences in an ANOVA can be expressed as contrasts or sets of contrasts. We start simply with inference about a single linear combination.

Working under the oneway ANOVA assumptions, we have that

$$Y_{ij} \sim \mathrm{n}(\theta_i, \sigma^2), \quad i = 1,\ldots,k, \quad j = 1,\ldots,n_i.$$

Therefore,

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim \mathrm{n}\big(\theta_i, \sigma^2/n_i\big), \quad i = 1, \ldots, k.$$

A note on notation: It is a common convention that if a subscript is replaced by a $\cdot$ (dot), it means that subscript has been summed over. Thus, $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$ and $Y_{\cdot j} = \sum_{i=1}^{k} Y_{ij}$. The addition of a "bar" indicates that a mean is taken, as in $\bar{Y}_{i\cdot}$ above. If both subscripts are summed over and the overall mean (called the *grand mean*) is calculated, we will break this rule to keep notation a little simpler and write $\bar{\bar{Y}} = (1/N)\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$, where $N = \sum_{i=1}^{k} n_i$.

For any constants $\mathbf{a} = (a_1, \ldots, a_k)$, $\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot}$ is also normal (see Exercise 11.8) with

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot}\right) = \sum_{i=1}^{k} a_i \theta_i \qquad \text{and} \qquad \mathrm{Var}\left(\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot}\right) = \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i},$$

and furthermore

$$\frac{\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{\sigma^2 \sum_{i=1}^{k} a_i^2/n_i}} \sim \mathrm{n}(0, 1).$$

Although this is nice, we are usually in the situation of wanting to make inferences about the θ_is without knowledge of σ. Therefore, we want to replace σ with an estimate. In each population, if we denote the sample variance by S_i^2, that is,

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2, \quad i = 1, \ldots, k,$$

then S_i^2 is an estimate of σ^2 and $(n_i - 1)S_i^2/\sigma^2 \sim \chi^2_{n_i - 1}$. Furthermore, under the ANOVA assumptions, since each S_i^2 estimates the same σ^2, we can improve the estimators by combining them. We thus use the pooled estimator of σ^2, S_p^2, given by

$$S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k} (n_i - 1)S_i^2 = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2. \tag{11.2.5}$$

Note that $N - k = \sum(n_i - 1)$. Since the S_i^2s are independent, Lemma 5.3.2 shows that $(N-k)S_p^2/\sigma^2 \sim \chi^2_{N-k}$. Also, S_p^2 is independent of each $\bar{Y}_{i\cdot}$ (see Exercise 11.6) and thus

$$\frac{\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^{k} a_i^2/n_i}} \sim t_{N-k}, \tag{11.2.6}$$

Student's t with $N - k$ degrees of freedom.

To test

$$H_0: \sum_{i=1}^{k} a_i\theta_i = 0 \qquad \text{versus} \qquad H_1: \sum_{i=1}^{k} a_i\theta_i \neq 0$$

at level α, we would reject H_0 if

$$\left| \frac{\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot}}{\sqrt{S_p^2 \sum_{i=1}^{k} a_i^2/n_i}} \right| > t_{N-k,\alpha/2}. \tag{11.2.7}$$

(Exercise 11.9 shows some other tests involving linear combinations.) Furthermore, (11.2.6) defines a pivot that can be inverted to give an interval estimator of $\sum a_i\theta_i$. With probability $1-\alpha$,

$$\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - t_{N-k,\alpha/2}\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \leq \sum_{i=1}^{k} a_i\theta_i$$

$$\leq \sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} + t_{N-k,\alpha/2}\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}. \tag{11.2.8}$$

Example 11.2.6 (ANOVA contrasts) Special values of $\mathbf{a}$ will give particular tests or confidence intervals. For example, to compare treatments 1 and 2, take $\mathbf{a} = (1, -1, 0, \ldots, 0)$. Then, using (11.2.6), to test $H_0: \theta_1 = \theta_2$ versus $H_1: \theta_1 \neq \theta_2$, we would reject H_0 if

$$\left| \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| > t_{N-k,\alpha/2}.$$

Note, the difference between this test and the two-sample t test (see Exercise 8.41) is that here information from treatments $3, \ldots, k$, as well as treatments 1 and 2, is used to estimate σ^2.

Alternatively, to compare treatment 1 to the average of treatments 2 and 3 (for example, treatment 1 might be a control, 2 and 3 might be experimental treatments, and we are looking for some overall effect), we would take $\mathbf{a} = (1, -\frac{1}{2}, -\frac{1}{2}, 0, \ldots, 0)$ and reject $H_0: \theta_1 = \frac{1}{2}(\theta_2 + \theta_3)$ if

$$\left| \frac{\bar{Y}_{1\cdot} - \frac{1}{2}\bar{Y}_{2\cdot} - \frac{1}{2}\bar{Y}_{3\cdot}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{4n_2} + \frac{1}{4n_3}\right)}} \right| > t_{N-k,\alpha/2}.$$

Using either (11.2.6) or (11.2.8), we have a way of testing or estimating any linear combination in the ANOVA. By judiciously choosing our linear combination we can

learn much about the treatment means. For example, if we look at the contrasts $\theta_1 - \theta_2, \theta_2 - \theta_3$, and $\theta_1 - \theta_3$, we can learn something about the ordering of the θ_is. (Of course, we have to be careful of the overall α level when doing a number of tests or intervals, but we can use the Bonferroni Inequality. See Example 11.2.9.)

We also must use some care in drawing formal conclusions from combinations of contrasts. Consider the hypotheses

$$H_0: \quad \theta_1 = \frac{1}{2}(\theta_2 + \theta_3) \qquad \text{versus} \qquad H_1: \quad \theta_1 < \frac{1}{2}(\theta_2 + \theta_3)$$

and

$$H_0: \quad \theta_2 = \theta_3 \qquad \text{versus} \qquad H_1: \quad \theta_2 < \theta_3.$$

If we reject both null hypotheses, we can conclude that θ_3 is greater than both θ_1 and θ_2, although we can draw no formal conclusion about the ordering of θ_2 and θ_1 from these two tests. (See Exercise 11.10.) ‖

Now we will use these univariate results about linear combinations and the relationship between the ANOVA null hypothesis and contrasts given in Theorem 11.2.5 to derive a test of the ANOVA null hypothesis.

11.2.4 The ANOVA F Test

In the previous section we saw how to deal with single linear combinations and, in particular, contrasts in the ANOVA. Also, in Section 11.2, we saw that the ANOVA null hypothesis is equivalent to a hypothesis about contrasts. In this section we will use this equivalence, together with the union–intersection methodology of Chapter 8, to derive a test of the ANOVA hypothesis.

From Theorem 11.2.5, the ANOVA hypothesis test can be written

$$H_0: \quad \sum_{i=1}^{k} a_i\theta_i = 0 \text{ for all } \mathbf{a} \in \mathcal{A} \qquad \text{versus} \qquad H_1: \quad \sum_{i=1}^{k} a_i\theta_i \neq 0 \text{ for some } \mathbf{a} \in \mathcal{A},$$

where $\mathcal{A} = \{\mathbf{a} = (a_1, \ldots, a_k) : \sum_{i=1}^{k} a_i = 0\}$. To see this more clearly as a union–intersection test, define, for each $\mathbf{a}$, the set

$$\Theta_{\mathbf{a}} = \{\theta = (\theta_1, \ldots, \theta_k) : \sum_{i=1}^{k} a_i\theta_i = 0\}.$$

Then we have

$$\theta \in \{\theta : \theta_1 = \theta_2 = \cdots = \theta_k\} \Leftrightarrow \theta \in \Theta_{\mathbf{a}} \quad \text{for all } \mathbf{a} \in \mathcal{A} \Leftrightarrow \theta \in \bigcap_{\mathbf{a}\in\mathcal{A}} \Theta_{\mathbf{a}},$$

showing that the ANOVA null can be written as an intersection.

Now, recalling the union–intersection methodology from Section 8.2.3, we would reject $H_0: \theta \in \cap_{\mathbf{a}\in\mathcal{A}}\Theta_{\mathbf{a}}$ (and, hence, the ANOVA null) if we can reject

$$H_{0_{\mathbf{a}}}: \quad \theta \in \Theta_{\mathbf{a}} \qquad \text{versus} \qquad H_{1_{\mathbf{a}}}: \quad \theta \notin \Theta_{\mathbf{a}}$$

for any $\mathbf{a}$. We test $H_{0_\mathbf{a}}$ with the t statistic of (11.2.6),

$$T_\mathbf{a} = \left| \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^k a_i^2/n_i}} \right|. \tag{11.2.9}$$

We then reject $H_{0_\mathbf{a}}$ if $T_\mathbf{a} > k$ for some constant k. From the union–intersection methodology, it follows that if we could reject for any $\mathbf{a}$, we could reject for the $\mathbf{a}$ that maximizes $T_\mathbf{a}$. Thus, the union–intersection test of the ANOVA null is to reject H_0 if $\sup_\mathbf{a} T_\mathbf{a} > k$, where k is chosen so that $P_{H_0}(\sup_\mathbf{a} T_\mathbf{a} > k) = \alpha$.

Calculation of $\sup_\mathbf{a} T_\mathbf{a}$ is not straightforward, although with a little care it is not difficult. The calculation is that of a constrained maximum, similar to problems previously encountered (see, for example, Exercise 7.41, where a constrained minimum is calculated). We will attack the problem in a manner similar to what we have done previously and use the Cauchy–Schwarz Inequality. (Alternatively, a method such as Lagrange multipliers could be used, but then we would have to use second-order conditions to verify that a maximum has been found.)

Most of the technical maximization arguments will be given in the following lemma and the lemma will then be applied to obtain the supremum of $T_\mathbf{a}$. The lemma is just a statement about constrained maxima of quadratic functions. The proof of the lemma may be skipped by the fainthearted.

Lemma 11.2.7 *Let $(v_1, \ldots, v_k)$ be constants and let $(c_1, \ldots, c_k)$ be positive constants. Then, for $\mathcal{A} = \{\mathbf{a} = (a_1, \ldots, a_k) : \sum a_i = 0\}$,*

$$\max_{\mathbf{a} \in \mathcal{A}} \left\{ \frac{\left(\sum_{i=1}^k a_i v_i\right)^2}{\sum_{i=1}^k a_i^2/c_i} \right\} = \sum_{i=1}^k c_i (v_i - \bar{v}_c)^2, \tag{11.2.10}$$

where $\bar{v}_c = \sum c_i v_i / \sum c_i$. The maximum is attained at any $\mathbf{a}$ of the form $a_i = K c_i (v_i - \bar{v}_c)$, where K is a nonzero constant.

Proof: Define $\mathcal{B} = \{\mathbf{b} = (b_1, \ldots, b_k) : \sum b_i = 0 \text{ and } \sum b_i^2/c_i = 1\}$. For any $\mathbf{a} \in \mathcal{A}$, define $\mathbf{b} = (b_1, \ldots, b_k)$ by

$$b_i = \frac{a_i}{\sqrt{\sum_{i=1}^k a_i^2/c_i}}$$

and note that $\mathbf{b} \in \mathcal{B}$. For any $\mathbf{a} \in \mathcal{A}$,

$$\frac{\left(\sum_{i=1}^k a_i v_i\right)^2}{\sum_{i=1}^k a_i^2/c_i} = \left(\sum_{i=1}^k b_i v_i\right)^2.$$

We will find an upper bound on $(\sum b_i v_i)^2$ for $\mathbf{b} \in \mathcal{B}$, and then we will show that the maximizing $\mathbf{a}$ given in the lemma achieves the upper bound.

Since we are dealing with the sum of products, the Cauchy–Schwarz Inequality (see Section 4.7) is a natural thing to try, but we have to be careful to build in the

constraints involving the c_is. We can do this in the following way. Define $C = \sum c_i$ and write

$$\frac{1}{C^2}\left(\sum_{i=1}^{k} b_i v_i\right)^2 = \left\{\sum_{i=1}^{k}\left(\frac{b_i}{c_i}\right)(v_i)\left(\frac{c_i}{C}\right)\right\}^2.$$

This is the square of a *covariance* for a probability measure defined by the ratios c_i/C. Formally, if we define random variables B and V by

$$P\left(B = \frac{b_i}{c_i}, V = v_i\right) = \frac{c_i}{C}, \quad i = 1, \ldots, k,$$

then $\mathrm{E}B = \sum(b_i/c_i)(c_i/C) = \sum b_i/C = 0$. Thus,

$$\begin{aligned}\left\{\sum_{i=1}^{k}\left(\frac{b_i}{c_i}\right)(v_i)\left(\frac{c_i}{C}\right)\right\}^2 &= (\mathrm{E}BV)^2 \\ &= (\mathrm{Cov}(B,V))^2 && (\mathrm{E}B = 0) \\ &\le (\mathrm{Var}\ B)(\mathrm{Var}\ V) && (\text{Cauchy–Schwarz Inequality}) \\ &= \left(\sum_{i=1}^{k}\left(\frac{b_i}{c_i}\right)^2\left(\frac{c_i}{C}\right)\right)\left(\sum_{i=1}^{k}(v_i - \bar{v}_c)^2\left(\frac{c_i}{C}\right)\right). && \left(\bar{v}_c = \frac{\sum c_i v_i}{\sum c_i}\right)\end{aligned}$$

Using the fact that $\sum b_i^2/c_i = 1$ and canceling common terms, we obtain

$$\left(\sum_{i=1}^{k} b_i v_i\right)^2 \le \sum_{i=1}^{k} c_i (v_i - \bar{v}_c)^2 \quad \text{for any } \mathbf{b} \in \mathcal{B}. \tag{11.2.11}$$

Finally, we see that if $a_i = Kc_i(v_i - \bar{v}_c)$ for any nonzero constant K, then $\mathbf{a} \in \mathcal{A}$ and

$$b_i = \frac{Kc_i(v_i - \bar{v}_c)}{\sqrt{\sum_{i=1}^{k}(Kc_i(v_i - \bar{v}_c))^2/c_i}} = \frac{c_i(v_i - \bar{v}_c)}{\sqrt{\sum_{i=1}^{k} c_i(v_i - \bar{v}_c)^2}}.$$

Since $\sum c_i(v_i - \bar{v}_c) = 0$,

$$\begin{aligned}\sum_{i=1}^{k} b_i v_i &= \frac{\sum_{i=1}^{k} c_i(v_i - \bar{v}_c)v_i}{\sqrt{\sum_{i=1}^{k} c_i(v_i - \bar{v}_c)^2}} \\ &= \frac{\sum_{i=1}^{k} c_i(v_i - \bar{v}_c)^2}{\sqrt{\sum_{i=1}^{k} c_i(v_i - \bar{v}_c)^2}} = \sqrt{\sum_{i=1}^{k} c_i(v_i - \bar{v}_c)^2},\end{aligned}$$

and the inequality in (11.2.11) is an equality. Thus, the upper bound is attained and the function is maximized at such an $\mathbf{a}$. □

Returning to $T_{\mathbf{a}}$ of (11.2.9), we see that maximizing $T_{\mathbf{a}}$ is equivalent to maximizing $T_{\mathbf{a}}^2$. We have

$$T_{\mathbf{a}}^2 = \frac{\left(\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i\right)^2}{S_p^2 \sum_{i=1}^k a_i^2/n_i} = \frac{\left(\sum_{i=1}^k a_i \bar{U}_i\right)^2}{S_p^2 \sum_{i=1}^k a_i^2/n_i}. \qquad (\bar{U}_i = \bar{Y}_{i\cdot} - \theta_i)$$

Noting that S_p^2 has no effect on the maximization, we can apply Lemma 11.2.7 to the above expression to get the following theorem.

Theorem 11.2.8 *For $T_{\mathbf{a}}$ defined in expression (11.2.9),*

$$\sup_{\mathbf{a}:\sum a_i=0} T_{\mathbf{a}}^2 = \frac{\sum_{i=1}^k n_i \left((\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta})\right)^2}{S_p^2}, \tag{11.2.12}$$

where $\bar{\bar{Y}} = \sum n_i \bar{Y}_{i\cdot} / \sum n_i$ and $\bar{\theta} = \sum n_i \theta_i / \sum n_i$. Furthermore, under the ANOVA assumptions,

$$\sup_{\mathbf{a}:\sum a_i=0} T_{\mathbf{a}}^2 \sim (k-1) F_{k-1,N-k}, \tag{11.2.13}$$

that is, $\sup_{\mathbf{a}:\Sigma a_i=0} T_{\mathbf{a}}^2/(k-1)$ has an F distribution with $k-1$ and $N-k$ degrees of freedom. (Recall that $N = \sum n_i$.)

Proof: To prove (11.2.12), use Lemma 11.2.7 and identify v_i with $\bar{U}_i$ and c_i with n_i. The result is immediate.

To prove (11.2.13), we must show that the numerator and denominator of (11.2.12) are independent chi squared random variables, each divided by its degrees of freedom. From the ANOVA assumptions two things follow. The numerator and denominator are independent and $S_p^2 \sim \sigma^2 \chi_{N-k}^2/(N-k)$. A little work must be done to show that

$$\frac{1}{\sigma^2} \sum_{i=1}^k n_i \left((\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta})\right)^2 \sim \chi_{k-1}^2.$$

This can be done, however, and is left as an exercise. (See Exercise 11.7.) □

If $H_0: \theta_1 = \theta_2 = \cdots = \theta_k$ is true, $\theta_i = \bar{\theta}$ for all $i = 1, \ldots, k$ and the $\theta_i - \bar{\theta}$ terms drop out of (11.2.12). Thus, for an α level test of the ANOVA hypotheses

$$H_0: \quad \theta_1 = \theta_2 = \cdots = \theta_k \qquad \text{versus} \qquad H_1: \quad \theta_i \neq \theta_j \text{ for some } i, j,$$

we reject H_0 if

$$\frac{\sum_{i=1}^k n_i \left((\bar{Y}_{i\cdot} - \bar{\bar{Y}})\right)^2}{S_p^2} > (k-1) F_{k-1,N-k,\alpha}. \tag{11.2.14}$$

This rejection region is usually written as

$$\text{reject } H_0 \text{ if } F = \frac{\sum_{i=1}^{k} n_i \left((\bar{Y}_{i\cdot} - \bar{\bar{Y}})\right)^2 / (k-1)}{S_p^2} > F_{k-1,N-k,\alpha},$$

and the test statistic F is called the *ANOVA F statistic.*

11.2.5 Simultaneous Estimation of Contrasts

We have already seen how to estimate and test a single contrast in the ANOVA; the t statistic and interval are given in (11.2.6) and (11.2.8). However, in the ANOVA we are often in the position of wanting to make more than one inference and we know that the simultaneous inference from many α level tests is not necessarily at level α. In the context of the ANOVA this problem has already been mentioned.

Example 11.2.9 (Pairwise differences) Many times there is interest in pairwise differences of means. Thus, if an ANOVA has means $\theta_1, \ldots, \theta_k$, there may be interest in interval estimates of $\theta_1 - \theta_2$, $\theta_2 - \theta_3$, $\theta_3 - \theta_4$, etc. With the Bonferroni Inequality, we can build a simultaneous inference statement. Define

$$C_{ij} = \left\{ \theta_i - \theta_j : \theta_i - \theta_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_{N-k,\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \right\}.$$

Then $P(C_{ij}) = 1-\alpha$ for *each* C_{ij}, but, for example, $P(C_{12} \text{ and } C_{23}) < 1-\alpha$. However, this last inference is the kind that we want to make in the ANOVA.

Recall the Bonferroni Inequality, given in expression (1.2.10), which states that for any sets $A_1, \ldots, A_n$,

$$P\left(\bigcap_{i=1}^{n} A_i \right) \geq \sum_{i=1}^{n} P(A_i) - (n-1).$$

In this case we want to bound $P(\cap_{i,j} C_{ij})$, the probability that all of the pairwise intervals cover their respective differences.

If we want to make a simultaneous $1-\alpha$ statement about the coverage of m confidence sets, then, from the Bonferroni Inequality, we can construct each confidence set to be of level γ, where γ satisfies

$$1 - \alpha = \sum_{i=1}^{m} \gamma - (m-1),$$

or, equivalently,

$$\gamma = 1 - \frac{\alpha}{m}.$$

A slight generalization is also possible in that it is not necessary to require each individual inference at the same level. We can construct each confidence set to be of

level γ_i, where γ_i satisfies

$$1 - \alpha = \sum_{i=1}^{m} \gamma_i - (m - 1).$$

In an ANOVA with k treatments, simultaneous inference on all $k(k-1)/2$ pairwise differences can be made with confidence $1 - \alpha$ if each t interval has confidence $1 - 2\alpha/[k(k-1)]$. ||

An alternative and quite elegant approach to simultaneous inference is given by Scheffé (1959). Scheffé's procedure, sometimes called the *S method*, allows for simultaneous confidence intervals (or tests) on *all* contrasts. (Exercise 11.14 shows that Scheffé's method can also be used to set up simultaneous intervals for any linear combination, not just for contrasts.) The procedure allows us to set a confidence coefficient that will be valid for *all contrast intervals simultaneously*, not just a specified group. The Scheffé procedure would be preferred if a large number of contrasts are to be examined. If the number of contrasts is small, the Bonferroni bound will almost certainly be smaller. (See the Miscellanea section for a discussion of other types of multiple comparison procedures.)

The proof that the Scheffé procedure has simultaneous $1 - \alpha$ coverage on all contrasts follows easily from the union–intersection nature of the ANOVA test.

Theorem 11.2.10 *Under the ANOVA assumptions, if* $\mathbf{M} = \sqrt{(k-1)F_{k-1,N-k,\alpha}}$, *then the probability is* $1 - \alpha$ *that*

$$\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - \mathbf{M}\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \;\le\; \sum_{i=1}^{k} a_i \theta_i \;\le\; \sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} + \mathbf{M}\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}$$

simultaneously for all $\mathbf{a} \in \mathcal{A} = \{\mathbf{a} = (a_1, \ldots, a_k) : \sum a_i = 0\}$.

Proof: The simultaneous probability statement requires $\mathbf{M}$ to satisfy

$$P\left(\left|\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - \sum_{i=1}^{k} a_i \theta_i\right| \le \mathbf{M}\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \text{ for all } \mathbf{a} \in \mathcal{A}\right) = 1 - \alpha$$

or, equivalently,

$$P(T_{\mathbf{a}}^2 \le \mathbf{M}^2 \text{ for all } \mathbf{a} \in \mathcal{A}) = 1 - \alpha,$$

where $T_{\mathbf{a}}$ is defined in (11.2.9). However, since

$$P(T_{\mathbf{a}}^2 \le \mathbf{M}^2 \text{ for all } \mathbf{a} \in \mathcal{A}) = P\left(\sup_{\mathbf{a}:\sum a_i = 0} T_{\mathbf{a}}^2 \le \mathbf{M}^2\right),$$

Theorem 11.2.8 shows that choosing $\mathbf{M}^2 = (k-1)F_{k-1,N-k,\alpha}$ satisfies the probability requirement. □

One of the real strengths of the Scheffé procedure is that it allows legitimate "data snooping." That is, in classic statistics it is taboo to test hypotheses that have been suggested by the data, since this can bias the results and, hence, invalidate the inference. (We normally would not test $H_0: \theta_1 = \theta_2$ just because we noticed that $\bar{Y}_1.$ was different from $\bar{Y}_2.$. See Exercise 11.18.) However, with Scheffé's procedure such a strategy is legitimate. The intervals or tests are valid for *all* contrasts. Whether they have been suggested by the data makes no difference. They already have been taken care of by the Scheffé procedure.

Of course, we must pay for all of the inferential power offered by the Scheffé procedure. The payment is in the form of the lengths of the intervals. In order to guarantee the simultaneous confidence level, the intervals may be quite long. For example, it can be shown (see Exercise 11.15) that if we compare the t and F distributions, for any ν, α, and k, the cutoff points satisfy

$$t_{\nu,\alpha/2} \leq \sqrt{(k-1)F_{k-1,\nu,\alpha}},$$

and so the Scheffé intervals are always wider, sometimes much wider, than the single-contrast intervals (another argument in favor of the doctrine that nothing substitutes for careful planning and preparation in experimentation). The interval length phenomenon carries over to testing. It also follows from the above inequality that Scheffé tests are less powerful than t tests.

11.2.6 Partitioning Sums of Squares

The ANOVA provides a useful way of thinking about the way in which different treatments affect a measured variable—the idea of allocating variation to different sources. The basic idea of allocating variation can be summarized in the following identity.

Theorem 11.2.11 *For any numbers $y_{ij}, i = 1, \ldots, k$, and $j = 1, \ldots, n_i$,*

$$\text{(11.2.15)} \qquad \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2,$$

where $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_j y_{ij}$ and $\bar{\bar{y}} = \sum_i n_i \bar{y}_{i\cdot} / \sum_i n_i$.

Proof: The proof is quite simple and relies only on the fact that, when we are dealing with means, the cross-term often disappears. Write

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{\bar{y}})\right)^2,$$

expand the right-hand side, and regroup terms. (See Exercise 11.21.) □

The sums in (11.2.15) are called *sums of squares* and are thought of as measuring variation in the data ascribable to different sources. (They are sometimes called *corrected sums of squares*, where the word *corrected* refers to the fact that a mean has

been subtracted.) In particular, the terms in the oneway ANOVA model,

$$Y_{ij} = \theta_i + \epsilon_{ij},$$

are in one-to-one correspondence with the terms in (11.2.15). Equation (11.2.15) shows how to allocate variation to the treatments (variation *between* treatments) and to random error (variation *within* treatments). The left-hand side of (11.2.15) measures variation without regard to categorization by treatments, while the two terms on the right-hand side measure variation due only to treatments and variation due only to random error, respectively. The fact that these sources of variation satisfy the above identity shows that the variation in the data, measured by sums of squares, is additive in the same way as the ANOVA model.

One reason it is easier to deal with sums of squares is that, under normality, corrected sums of squares are chi squared random variables and we have already seen that independent chi squareds can be added to get new chi squareds.

Under the ANOVA assumptions, in particular if $Y_{ij} \sim \mathrm{n}(\theta_i, \sigma^2)$, it is easy to show that

$$\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi^2_{N-k}, \tag{11.2.16}$$

because for each $i = 1, \ldots, k$, $\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi^2_{n_i-1}$, all independent, and, for independent chi squared random variables, $\sum_{i=1}^{k}\chi^2_{n_i-1} \sim \chi^2_{N-k}$. Furthermore, if $\theta_i = \theta_j$ for every i, j, then

$$\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 \sim \chi^2_{k-1} \quad \text{and} \quad \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\bar{Y}})^2 \sim \chi^2_{N-1}. \tag{11.2.17}$$

Thus, under $H_0\colon \theta_1 = \cdots = \theta_k$, the sum of squares partitioning of (11.2.15) is a partitioning of chi squared random variables. When scaled, the left-hand side is distributed as a χ^2_{N-1}, and the right-hand side is the sum of two independent random variables distributed, respectively, as χ^2_{k-1} and χ^2_{N-k}. Note that the χ^2 partitioning is true only if the terms on the right-hand side of (11.2.15) are independent, which follows in this case from the normality in the ANOVA assumptions. The partitioning of χ^2s does hold in a slightly more general context, and a characterization of this is sometimes referred to as Cochran's Theorem. (See Searle 1971 and also the Miscellanea section.)

In general, it is possible to partition a sum of squares into sums of squares of uncorrelated contrasts, each with 1 degree of freedom. If the sum of squares has ν degrees of freedom and is χ^2_ν, it is possible to partition it into ν independent terms, each of which is χ^2_1.

The quantity $(\sum a_i\bar{Y}_{i\cdot})^2/(\sum a_i^2/n_i)$ is called the *contrast sum of squares* for a treatment contrast $\sum a_i\bar{Y}_{i\cdot}$. In a oneway ANOVA it is always possible to find sets of constants $\mathbf{a}^{(l)} = (a_1^{(l)}, \ldots, a_k^{(l)})$, $l = 1, \ldots, k-1$, to satisfy

$$\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 = \frac{\sum_{i=1}^{k} a_i^{(1)}\bar{Y}_{i\cdot}^2}{\sum_{i=1}^{k}(a_i^{(1)})^2/n_i} + \frac{\sum_{i=1}^{k} a_i^{(2)}\bar{Y}_{i\cdot}^2}{\sum_{i=1}^{k}(a_i^{(2)})^2/n_i} + \cdots + \frac{\sum_{i=1}^{k} a_i^{(k-1)}\bar{Y}_{i\cdot}^2}{\sum_{i=1}^{k}(a_i^{(k-1)})^2/n_i}$$

Table 11.2.1. *ANOVA table for oneway classification*

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Between treatment groups	$k-1$	SSB = $\sum n_i(\bar{y}_i - \bar{\bar{y}})^2$	MSB = SSB/$(k-1)$	$F = \frac{\text{MSB}}{\text{MSW}}$
Within treatment groups	$N-k$	SSW = $\sum\sum(y_{ij} - \bar{y}_{i\cdot})^2$	MSW = SSW/$(N-k)$	
Total	$N-1$	SST = $\sum\sum(y_{ij} - \bar{\bar{y}})^2$		

and

$$\sum_{i=1}^{k} \frac{a_i^{(l)} a_i^{(l')}}{n_i} = 0 \quad \text{for all } l \neq l'. \tag{11.2.18}$$

Thus, the individual contrast sums of squares are all uncorrelated and hence independent under normality (Lemma 5.3.3). When suitably normalized, the left-hand side of (11.2.18) is distributed as a χ^2_{k-1} and the right-hand side is $k-1$ χ^2_1s. (Such contrasts are called *orthogonal contrasts*. See Exercises 11.10 and 11.11.)

It is common to summarize the results of an ANOVA F test in a standard form, called an ANOVA table, shown in Table 11.2.1. The table also gives a number of useful, intermediate statistics. The headings should be self-explanatory.

Example 11.2.12 (Continuation of Example 11.2.1) The ANOVA table for the fish toxin data is

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Treatments	3	995.90	331.97	26.09
Within	15	190.83	12.72	
Total	18	1,186.73		

The F statistic of 26.09 is highly significant, showing that there is strong evidence the toxins produce different effects. ‖

It follows from equation (11.2.15) that the sum of squares column "adds"—that is, SSB + SSW = SST. Similarly, the degrees of freedom column adds. The mean square column, however, does not, as these are means rather than sums.

The ANOVA table contains no new statistics; it merely gives an orderly form for calculation and presentation. The F statistic is exactly the same as derived before and, moreover, MSW is the usual pooled, unbiased estimator of σ^2, S_p^2 of (11.2.5) (see Exercise 11.22).

11.3 Simple Linear Regression

In the analysis of variance we looked at how one factor (variable) influenced the means of a response variable. We now turn to simple linear regression, where we try to better understand the functional dependence of one variable on another. In particular, in simple linear regression we have a relationship of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i, \tag{11.3.1}$$

where Y_i is a random variable and x_i is another observable variable. The quantities α and β, the *intercept* and *slope* of the regression, are assumed to be fixed and unknown parameters and ϵ_i is, necessarily, a random variable. It is also common to suppose that $\mathrm{E}\epsilon_i = 0$ (otherwise we could just rescale the excess into α), so that, from (11.3.1), we have

$$\mathrm{E}Y_i = \alpha + \beta x_i. \tag{11.3.2}$$

In general, the function that gives $\mathrm{E}Y$ as a function of x is called the *population regression function*. Equation (11.3.2) defines the population regression function for simple linear regression.

One main purpose of regression is to predict Y_i from knowledge of x_i using a relationship like (11.3.2). In common usage this is often interpreted as saying that Y_i *depends* on x_i. It is common to refer to Y_i as the *dependent* variable and to refer to x_i as the *independent* variable. This terminology is confusing, however, since this use of the word *independent* is different from our previous usage. (The x_is are not necessarily random variables, so they cannot be statistically "independent" according to our usual meaning.) We will not use this confusing terminology but will use alternative, more descriptive terminology, referring to Y_i as the *response* variable and to x_i as the *predictor* variable.

Actually, to keep straight the fact that our inferences about the relationship between Y_i and x_i assume knowledge of x_i, we could write (11.3.2) as

$$\mathrm{E}(Y_i \mid x_i) = \alpha + \beta x_i. \tag{11.3.3}$$

We will tend to use (11.3.3) to reinforce the conditional aspect of any inferences.

Recall that in Chapter 4 we encountered the word *regression* in connection with conditional expectations (see Exercise 4.13). There, the regression of Y on X was defined as $\mathrm{E}(Y|x)$, the conditional expectation of Y given $X = x$. More generally, the word *regression* is used in statistics to signify a relationship between variables. When we refer to *regression that is linear*, we can mean that the conditional expectation of Y given $X = x$ is a linear function of x. Note that, in equation (11.3.3), it does not matter whether x_i is fixed and known or it is a realization of the observable random

variable X_i. In either case, equation (11.3.3) has the same interpretation. This will not be the case in Section 11.3.4, however, when we will be concerned with inference using the joint distribution of X_i and Y_i.

The term *linear regression* refers to a specification that is *linear in the parameters*. Thus, the specifications $\mathrm{E}(Y_i|x_i) = \alpha+\beta x_i^2$ and $\mathrm{E}(\log Y_i|x_i) = \alpha+\beta(1/x_i)$ both specify linear regressions. The first specifies a linear relationship between Y_i and x_i^2, and the second between $\log Y_i$ and $1/x_i$. In contrast, the specification $\mathrm{E}(Y_i|x_i) = \alpha + \beta^2 x_i$ does not specify a linear regression.

The term *regression* has an interesting history, dating back to the work of Sir Francis Galton in the 1800s. (See Freedman *et al.* 1991 for more details or Stigler 1986 for an in-depth historical treatment.) Galton investigated the relationship between heights of fathers and heights of sons. He found, not surprisingly, that tall fathers tend to have tall sons and short fathers tend to have short sons. However, he also found that very tall fathers tend to have shorter sons and very short fathers tend to have taller sons. (Think about it—it makes sense.) Galton called this phenomenon *regression toward the mean* (employing the usual meaning of *regression*, "to go back"), and from this usage we get the present use of the word *regression*.

Example 11.3.1 (Predicting grape crops) A more modern use of regression is to predict crop yields of grapes. In July, the grape vines produce clusters of berries, and a count of these clusters can be used to predict the final crop yield at harvest time. Typical data are like the following, which give the cluster counts and yields (tons/acre) for a number of years.

Year	Yield (Y)	Cluster count (x)
1971	5.6	116.37
1973	3.2	82.77
1974	4.5	110.68
1975	4.2	97.50
1976	5.2	115.88
1977	2.7	80.19
1978	4.8	125.24
1979	4.9	116.15
1980	4.7	117.36
1981	4.1	93.31
1982	4.4	107.46
1983	5.4	122.30

The data from 1972 are missing because the crop was destroyed by a hurricane. A plot of these data would show that there is a strong linear relationship. ‖

When we write an equation like (11.3.3) we are implicitly making the assumption that the regression of Y on X *is* linear. That is, the conditional expectation of Y, given that $X = x$, is a linear function of x. This assumption may not be justified, because there may be no underlying theory to support a linear relationship. However, since a linear relationship is so convenient to work with, we might want to assume

that the regression of Y on X can be adequately approximated by a linear function. Thus, we really do not expect (11.3.3) to hold, but instead we hope that

$$\mathrm{E}(Y_i|x_i) \approx \alpha + \beta x_i \tag{11.3.4}$$

is a reasonable approximation. If we start from the (rather strong) assumption that the pair (X_i, Y_i) has a bivariate normal distribution, it immediately follows that the regression of Y on X is linear. In this case, the conditional expectation $\mathrm{E}(Y|x)$ is linear in the parameters (see Definition 4.5.10 and the subsequent discussion).

There is one final distinction to be made. When we do a regression analysis, that is, when we investigate the relationship between a predictor and a response variable, there are two steps to the analysis. The first step is a totally data-oriented one, in which we attempt only to summarize the observed data. (This step is always done, since we almost always calculate sample means and variances or some other summary statistic. However, this part of the analysis now tends to get more complicated.) It is important to keep in mind that this "data fitting" step is not a matter of statistical inference. Since we are interested only in the data at hand, we do not have to make any assumptions about parameters.

The second step in the regression analysis is the statistical one, in which we attempt to infer conclusions about the relationship in the population, that is, about the population regression function. To do this, we need to make assumptions about the population. In particular, if we want to make inferences about the slope and intercept of a population linear relationship, we need to assume that there are parameters that correspond to these quantities.

In a simple linear regression problem, we observe data consisting of n pairs of observations, $(x_1, y_1), \ldots, (x_n, y_n)$. In this section, we will consider a number of different models for these data. The different models will entail different assumptions about whether x or y or both are observed values of random variables $\boldsymbol{X}$ or Y.

In each model we will be interested in investigating a linear relationship between x and y. The n data points will not fall exactly on a straight line, but we will be interested in *summarizing* the sample information by *fitting a line* to the observed data points. We will find that many different approaches lead us to the same line.

Based on the data $(x_1, y_1), \ldots, (x_n, y_n)$, define the following quantities. The *sample means* are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{11.3.5}$$

The *sums of squares* are

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{11.3.6}$$

and the *sum of cross-products* is

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{11.3.7}$$

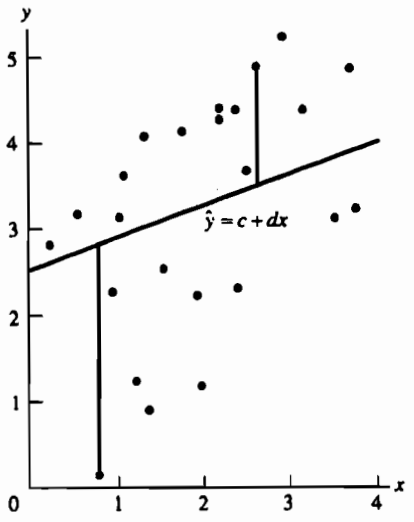

Figure 11.3.1. *Data from Table 11.3.1: Vertical distances that are measured by RSS*

Then the most common estimates of α and β in (11.3.4), which we will subsequently justify under various models, are denoted by a and b, respectively, and are given by

$$b = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}. \tag{11.3.8}$$

11.3.1 Least Squares: A Mathematical Solution

Our first derivation of estimates for α and β makes no statistical assumptions about the observations (x_i, y_i). Simply consider $(x_1, y_1), \ldots, (x_n, y_n)$ as n pairs of numbers plotted in a scatterplot as in Figure 11.3.1. (The 24 data points pictured in Figure 11.3.1 are listed in Table 11.3.1.) Think of drawing through this cloud of points a straight line that comes "as close as possible" to all the points.

Table 11.3.1. *Data pictured in Figure 11.3.1*

x	y	x	y	x	y	x	y
3.74	3.22	0.20	2.81	1.22	1.23	1.76	4.12
3.66	4.87	2.50	3.71	1.00	3.13	0.51	3.16
0.78	0.12	3.50	3.11	1.29	4.05	2.17	4.40
2.40	2.31	1.35	0.90	0.95	2.28	1.99	1.18
2.18	4.25	2.36	4.39	1.05	3.60	1.53	2.54
1.93	2.24	3.13	4.36	2.92	5.39	2.60	4.89
$\bar{x} = 1.95$	$\bar{y} = 3.18$	$S_{xx} = 22.82$		$S_{yy} = 43.62$		$S_{xy} = 15.48$	

For any line $y = c + dx$, the *residual sum of squares* (RSS) is defined to be

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (c + dx_i))^2.$$

The RSS measures the *vertical* distance from each data point to the line $c + dx$ and then sums the squares of these distances. (Two such distances are shown in Figure 11.3.1.) The *least squares estimates* of α and β are defined to be those values a and b such that the line $a + bx$ minimizes RSS. That is, the least squares estimates, a and b, satisfy

$$\min_{c,d} \sum_{i=1}^{n} (y_i - (c + dx_i))^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2.$$

This function of two variables, c and d, can be minimized in the following way. For any fixed value of d, the value of c that gives the minimum value can be found by writing

$$\sum_{i=1}^{n} (y_i - (c + dx_i))^2 = \sum_{i=1}^{n} ((y_i - dx_i) - c)^2.$$

From Theorem 5.2.4, the minimizing value of c is

$$c = \frac{1}{n} \sum_{i=1}^{n} (y_i - dx_i) = \bar{y} - d\bar{x}. \tag{11.3.9}$$

Thus, for a given value of d, the minimum value of RSS is

$$\sum_{i=1}^{n} ((y_i - dx_i) - (\bar{y} - d\bar{x}))^2 = \sum_{i=1}^{n} ((y_i - \bar{y}) - d(x_i - \bar{x}))^2 = S_{yy} - 2dS_{xy} + d^2 S_{xx}.$$

The value of d that gives the overall minimum value of RSS is obtained by setting the derivative of this quadratic function of d equal to 0. The minimizing value is

$$d = \frac{S_{xy}}{S_{xx}}. \tag{11.3.10}$$

This value is, indeed, a minimum since the coefficient of d^2 is positive. Thus, by (11.3.9) and (11.3.10), a and b from (11.3.8) are the values of c and d that minimize the residual sum of squares.

The RSS is only one of many reasonable ways of measuring the distance from the line $c + dx$ to the data points. For example, rather than using vertical distances we could use horizontal distances. This is equivalent to graphing the y variable on the horizontal axis and the x variable on the vertical axis and using vertical distances as we did above. Using the above results (interchanging the roles of x and y), we find the least squares line is $\hat{x} = a' + b'y$, where

$$b' = \frac{S_{xy}}{S_{yy}} \quad \text{and} \quad a' = \bar{x} - b'\bar{y}.$$

Reexpressing the line so that y is a function of x, we obtain $\hat{y} = -(a'/b') + (1/b')x$.

Usually the line obtained by considering horizontal distances is different from the line obtained by considering vertical distances. From the values in Table 11.3.1, the *regression of y on x* (vertical distances) is $\hat{y} = 1.86 + .68x$. The *regression of x on y* (horizontal distances) is $\hat{y} = -2.31 + 2.82x$. In Figure 12.2.2, these two lines are shown (along with a third line discussed in Section 12.2). If these two lines were the same, then the slopes would be the same and $b/(1/b')$ would equal 1. But, in fact, $b/(1/b') \leq 1$ with equality only in special cases. Note that

$$\frac{b}{1/b'} = bb' = \frac{(S_{xy})^2}{S_{xx}S_{yy}}.$$

Using the version of Hölder's Inequality in (4.7.9) with $p = q = 2, a_i = x_i - \bar{x}$, and $b_i = y_i - \bar{y}$, we see that $(S_{xy})^2 \leq S_{xx}S_{yy}$ and, hence, the ratio is less than 1.

If x is the predictor variable, y is the response variable, and we think of predicting y from x, then the vertical distance measured in RSS is reasonable. It measures the distance from y_i to the predicted value of $y_i, \hat{y}_i = c + dx_i$. But if we do not make this distinction between x and y, then it is unsettling that another reasonable criterion, horizontal distance, gives a different line.

The least squares method should be considered only as a method of "fitting a line" to a set of data, not as a method of statistical inference. We have no basis for constructing confidence intervals or testing hypotheses because, in this section, we have not used any statistical model for the data. When we think of a and b in the context of this section, it might be better to call them least squares *solutions* rather than least squares *estimates* because they are the solutions of the mathematical problem of minimizing the RSS rather than estimates derived from a statistical model. But, as we shall see, these least squares solutions have optimality properties in certain statistical models.

11.3.2 Best Linear Unbiased Estimators: A Statistical Solution

In this section we show that the estimates a and b from (11.3.8) are optimal in the class of linear unbiased estimates under a fairly general statistical model. The model is described as follows. Assume that the values $x_1, \ldots, x_n$ are known, fixed values. (Think of them as values the experimenter has chosen and set in a laboratory experiment.) The values $y_1, \ldots, y_n$ are observed values of uncorrelated random variables $Y_1, \ldots, Y_n$. The linear relationship assumed between the xs and the ys is

$$\mathrm{E}Y_i = \alpha + \beta x_i, \quad i = 1, \ldots, n, \tag{11.3.11}$$

where we also assume that

$$\mathrm{Var}\, Y_i = \sigma^2. \tag{11.3.12}$$

There is no subscript in σ^2 because we are assuming that all the Y_is have the same (unknown) variance. These assumptions about the first two moments of the Y_is are the only assumptions we need to make to proceed with the derivation in this subsection. For example, we do not need to specify a probability distribution for the $Y_1, \ldots, Y_n$.

The model in (11.3.11) and (11.3.12) can also be expressed in this way. We assume that

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{11.3.13}$$

where $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated random variables with

$$\mathrm{E}\,\epsilon_i = 0 \quad \text{and} \quad \mathrm{Var}\,\epsilon_i = \sigma^2. \tag{11.3.14}$$

The $\epsilon_1, \ldots, \epsilon_n$ are called the *random errors*. Since Y_i depends only on ϵ_i and the ϵ_is are uncorrelated, the Y_is are uncorrelated. Also, from (11.3.13) and (11.3.14), the expressions for $\mathrm{E}Y_i$ and $\mathrm{Var}\,Y_i$ in (11.3.11) and (11.3.12) are easily verified.

To derive estimators for the parameters α and β, we restrict attention to the class of *linear estimators*. An estimator is a linear estimator if it is of the form

$$\sum_{i=1}^{n} d_i Y_i, \tag{11.3.15}$$

where $d_1, \ldots, d_n$ are known, fixed constants. (Exercise 7.39 concerns linear estimators of a population mean.) Among the class of linear estimators, we further restrict attention to unbiased estimators. This restricts the values of $d_1, \ldots, d_n$ that can be used.

An unbiased estimator of the slope β must satisfy

$$\mathrm{E}\sum_{i=1}^{n} d_i Y_i = \beta,$$

regardless of the true value of the parameters α and β. This implies that

$$\beta = \mathrm{E}\sum_{i=1}^{n} d_i Y_i = \sum_{i=1}^{n} d_i \mathrm{E}Y_i = \sum_{i=1}^{n} d_i(\alpha + \beta x_i)$$
$$= \alpha\left(\sum_{i=1}^{n} d_i\right) + \beta\left(\sum_{i=1}^{n} d_i x_i\right).$$

This equality is true for *all* α and β if and only if

$$\sum_{i=1}^{n} d_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} d_i x_i = 1. \tag{11.3.16}$$

Thus, $d_1, \ldots, d_n$ must satisfy (11.3.16) in order for the estimator to be an unbiased estimator of β.

In Chapter 7 we called an unbiased estimator "best" if it had the smallest variance among all unbiased estimators. Similarly, an estimator is the *best linear unbiased estimator* (*BLUE*) if it is the linear unbiased estimator with the smallest variance. We will now show that the choice of $d_i = (x_i - \bar{x})/S_{xx}$ that defines the estimator $b = S_{xY}/S_{xx}$ is the best choice in that it results in the linear unbiased estimator of β

with the smallest variance. (The d_is must be known, fixed constants but the x_is are known, fixed constants, so this choice of d_is is legitimate.)

A note on notation: The notation S_{xY} stresses the fact that S_{xY} is a random variable that is a function of the random variables $Y_1, \ldots, Y_n$. S_{xY} also depends on the nonrandom quantities $x_1, \ldots, x_n$.

Because $Y_1, \ldots, Y_n$ are uncorrelated with equal variance σ^2, the variance of *any* linear estimator is given by

$$\operatorname{Var} \sum_{i=1}^{n} d_i Y_i = \sum_{i=1}^{n} d_i^2 \operatorname{Var} Y_i = \sum_{i=1}^{n} d_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} d_i^2.$$

The BLUE of β is, therefore, defined by constants $d_1, \ldots, d_n$ that satisfy (11.3.16) and have the minimum value of $\sum_{i=1}^{n} d_i^2$. (The presence of σ^2 has no effect on the minimization over linear estimators since it appears as a multiple of the variance of every linear estimator.)

The minimizing values of the constants $d_1, \ldots, d_n$ can now be found by using Lemma 11.2.7. To apply the lemma to our minimization problem, make the following correspondences, where the left-hand sides are notation from Lemma 11.2.7 and the right-hand sides are our current notation. Let

$$k = n, \quad v_i = x_i, \quad c_i = 1, \quad \text{and} \quad a_i = d_i,$$

which implies $\bar{v}_c = \bar{x}$. If d_i is of the form

$$d_i = K c_i (v_i - \bar{v}_c) = K(x_i - \bar{x}), \quad i = 1, \ldots, n, \tag{11.3.17}$$

then, by Lemma 11.2.7, $d_1, \ldots, d_n$ maximize

$$\frac{\left(\sum_{i=1}^{n} d_i x_i\right)^2}{\sum_{i=1}^{n} d_i^2} \tag{11.3.18}$$

among all $d_1, \ldots, d_n$ that satisfy $\sum d_i = 0$. Furthermore, since

$$\{(d_1, \ldots, d_n) : \sum d_i = 0, \sum d_i x_i = 1\} \subset \{(d_1, \ldots, d_n) : \sum d_i = 0\},$$

if d_is of the form (11.3.17) also satisfy (11.3.16), they certainly maximize (11.3.18) among all $d_1, \ldots, d_n$ that satisfy (11.3.16). (Since the set over which the maximum is taken is smaller, the maximum cannot be larger.) Now, using (11.3.17), we have

$$\sum_{i=1}^{n} d_i x_i = \sum_{i=1}^{n} K(x_i - \bar{x}) x_i = K S_{xx}.$$

The second constraint in (11.3.16) is satisfied if $K = \frac{1}{S_{xx}}$. Therefore, with $d_1, \ldots, d_n$ defined by

$$d_i = \frac{(x_i - \bar{x})}{S_{xx}}, \quad i = 1, \ldots, n, \tag{11.3.19}$$

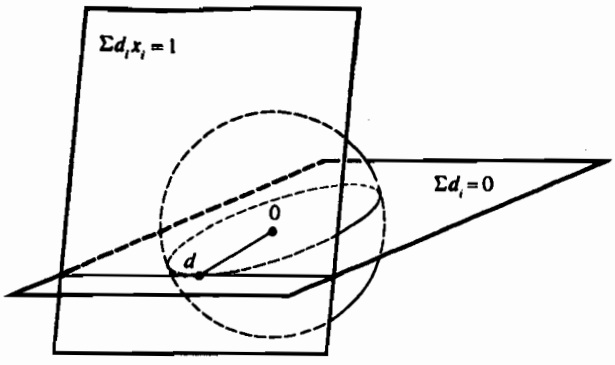

Figure 11.3.2. *Geometric description of the BLUE*

both constraints of (11.3.16) are satisfied and this set of d_is produces the maximum. Finally, note that for all $d_1, \ldots, d_n$ that satisfy (11.3.16),

$$\frac{\left(\sum_{i=1}^n d_i x_i\right)^2}{\sum_{i=1}^n d_i^2} = \frac{1}{\sum_{i=1}^n d_i^2}.$$

Thus, for $d_1, \ldots, d_n$ that satisfy (11.3.16), maximization of (11.3.18) is equivalent to minimization of $\sum d_i^2$. Hence, we can conclude that the d_is defined in (11.3.19) give the minimum value of $\sum d_i^2$ among all d_is that satisfy (11.3.16), and the linear unbiased estimator defined by these d_is, namely,

$$b = \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i = \frac{S_{xy}}{S_{xx}},$$

is the BLUE of β.

A geometric description of this construction of the BLUE of β is given in Figure 11.3.2, where we take $n = 3$. The figure shows three-dimensional space with coordinates d_1, d_2, and d_3. The two planes represent the vectors (d_1, d_2, d_3) that satisfy the two linear constraints in (11.3.16), and the line where the two planes intersect consists of the vectors (d_1, d_2, d_3) that satisfy both equalities. For any point on the line, $\sum_{i=1}^n d_i^2$ is the square of the distance from the point to the origin $\mathbf{0}$. The vector (d_1, d_2, d_3) that defines the BLUE is the point on the line that is closest to $\mathbf{0}$. The sphere in the figure is the smallest sphere that intersects the line, and the point of intersection is the point (d_1, d_2, d_3) that defines the BLUE of β. This, we have shown, is the point with $d_i = (x_i - \bar{x})/S_{xx}$.

The variance of b is

$$\text{Var } b = \sigma^2 \sum_{i=1}^n d_i^2 = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{11.3.20}$$

Since $x_1, \ldots, x_n$ are values chosen by the experimenter, they can be chosen to make S_{xx} large and the variance of the estimator small. That is, the experimenter can *design*

the experiment to make the estimator more precise. Suppose that all the $x_1, \ldots, x_n$ must be chosen in an interval $[e, f]$. Then, if n is even, the choice of $x_1, \ldots, x_n$ that makes S_{xx} as large as possible is to take half of the x_is equal to e and half equal to f (see Exercise 11.26). This would be the best design in that it would give the most precise estimate of the slope β if the experimenter were certain that the model described by (11.3.11) and (11.3.12) was correct. In practice, however, this design is seldom used because an experimenter is hardly ever certain of the model. This *two-point design* gives information about the value of $\mathrm{E}(Y|x)$ at only two values, $x = e$ and $x = f$. If the population regression function $\mathrm{E}(Y|x)$, which gives the mean of Y as a function of x, is nonlinear, it could never be detected from data obtained using the "optimal" two-point design.

We have shown that b is the BLUE of β. A similar analysis will show that a is the BLUE of the intercept α. The constants $d_1, \ldots, d_n$ that define a linear estimator of α must satisfy

$$\sum_{i=1}^{n} d_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} d_i x_i = 0. \tag{11.3.21}$$

The details of this derivation are left as Exercise 11.27. The fact that least squares estimators are BLUEs holds in other linear models also. This general result is called the *Gauss–Markov Theorem* (see Christensen 1996; Lehmann and Casella 1998, Section 3.4, or the more general treatment in Harville 1981).

11.3.3 Models and Distribution Assumptions

In this section, we will introduce two more models for paired data $(x_1, y_1), \ldots, (x_n, y_n)$ that are called simple linear regression models.

To obtain the least squares estimates in Section 11.3.1, we used no statistical model. We simply solved a mathematical minimization problem. Thus, we could not derive any statistical properties about the estimators obtained by this method because there were no probability models to work with. There are not really any parameters for which we could construct hypothesis tests or confidence intervals.

In Section 11.3.2 we made some statistical assumptions about the data. Specifically, we made assumptions about the first two moments, the mean, variance, and covariance of the data. These are all statistical assumptions related to probability models for the data, and we derived statistical properties for the estimators. The properties of unbiasedness and minimum variance, which we proved for the estimators a and b of the parameters α and β, are statistical properties.

To obtain these properties we did not have to specify a complete probability model for the data, only assumptions about the first two moments. We were able to obtain a general optimality property under these minimal assumptions, but the optimality was only in a restricted class of estimators—linear unbiased estimators. We were not able to derive exact tests and confidence intervals under this model because the model does not specify enough about the probability distribution of the data. We now present two statistical models that completely specify the probabilistic structure of the data.

Conditional normal model

The *conditional normal model* is the most common simple linear regression model and the most straightforward to analyze. The observed data are the n pairs, $(x_1, y_1), \ldots, (x_n, y_n)$. The values of the predictor variable, $x_1, \ldots, x_n$, are considered to be known, fixed constants. As in Section 11.3.2, think of them as being chosen and set by the experimenter. The values of the response variable, $y_1, \ldots, y_n$, are observed values of random variables, $Y_1, \ldots, Y_n$. The random variables $Y_1, \ldots, Y_n$ are assumed to be independent. Furthermore, the distribution of the Y_is is normal, specifically,

$$Y_i \sim \mathrm{n}(\alpha + \beta x_i, \sigma^2), \quad i = 1, \ldots, n. \tag{11.3.22}$$

Thus the population regression function is a linear function of x, that is, $\mathrm{E}(Y|x) = \alpha + \beta x$, and all the Y_is have the same variance, σ^2. The conditional normal model can be expressed similar to (11.3.13) and (11.3.14), namely,

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{11.3.23}$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathrm{n}(0, \sigma^2)$ random variables.

The conditional normal model is a special case of the model considered in Section 11.3.2. The population regression function, $\mathrm{E}(Y|x) = \alpha + \beta x$, and the variance, $\mathrm{Var}\, Y = \sigma^2$, are as in that model. The uncorrelatedness of $Y_1, \ldots, Y_n$ (or, equivalently, $\epsilon_1, \ldots, \epsilon_n$) has been strengthened to independence. And, of course, rather than just the first two moments of the distribution of $Y_1, \ldots, Y_n$, the exact form of the probability distribution is now specified.

The joint pdf of $Y_1, \ldots, Y_n$ is the product of the marginal pdfs because of the independence. It is given by

$$\begin{aligned} f(\mathbf{y}|\alpha, \beta, \sigma^2) &= f(y_1, \ldots, y_n|\alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^{n} f(y_i|\alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(y_i - (\alpha + \beta x_i))^2/(2\sigma^2)\right] \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\left(\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right)/(2\sigma^2)\right]. \end{aligned} \tag{11.3.24}$$

It is this joint probability distribution that will be used to develop the statistical procedures in Sections 11.3.4 and 11.3.5. For example, the expression in (11.3.24) will be used to find MLEs of α, β, and σ^2.

Bivariate normal model

In all the previous models we have discussed, the values of the predictor variable, $x_1, \ldots, x_n$, have been fixed, known constants. But sometimes these values are actually observed values of random variables, $X_1, \ldots, X_n$. In Galton's example in Section 11.3,

$x_1, \ldots, x_n$ were observed heights of fathers. But the experimenter certainly did not choose these heights before collecting the data. Thus it is necessary to consider models in which the predictor variable, as well as the response variable, is random. One such model that is fairly simple is the *bivariate normal model.* A more complex model is discussed in Section 12.2.

In the bivariate normal model the data $(x_1, y_1), \ldots, (x_n, y_n)$ are observed values of the bivariate random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$. The random vectors are independent and the joint distribution of (X_i, Y_i) is assumed to be bivariate normal. Specifically, it is assumed that

$$(X_i, Y_i) \sim \text{bivariate normal}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho).$$

The joint pdf and various properties of a bivariate normal distribution are given in Definition 4.5.10 and the subsequent discussion. The joint pdf of all the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is the product of these bivariate pdfs.

In a simple linear regression analysis, we are still thinking of x as the predictor variable and y as the response variable. That is, we are most interested in predicting the value of y having observed the value of x. This naturally leads to basing inference on the conditional distribution of Y given $X = x$. For a bivariate normal model, the conditional distribution of Y given $X = x$ is normal. The population regression function is now a true conditional expectation, as the notation suggests, and is

$$\text{(11.3.25)} \qquad \mathrm{E}(Y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \left[\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right] + \left[\rho\frac{\sigma_Y}{\sigma_X}\right]x.$$

The bivariate normal model *implies* that the population regression is a linear function of x. We need not assume this as in the previous models. Here $\mathrm{E}(Y|x) = \alpha + \beta x$, where $\beta = \rho\frac{\sigma_Y}{\sigma_X}$ and $\alpha = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$. Also, as in the conditional normal model, the conditional variance of the response variable Y does not depend on x,

$$\text{(11.3.26)} \qquad \mathrm{Var}\,(Y|x) = \sigma_Y^2(1 - \rho^2).$$

For the bivariate normal model, the linear regression analysis is almost always carried out using the conditional distribution of $(Y_1, \ldots, Y_n)$ given $X_1 = x_1, \ldots, X_n = x_n$, rather than the unconditional distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. But then we are in the same situation as the conditional normal model described above. The fact that $x_1, \ldots, x_n$ are observed values of random variables is immaterial if we condition on these values and, in general, in simple linear regression we do not use the fact of bivariate normality except to define the conditional distribution. (Indeed, for the most part, the marginal distribution of X is of no consequence whatsoever. In linear regression it is the conditional distribution that matters.) Inference based on point estimators, intervals, or tests is the same for the two models. See Brown (1990b) for an alternative view.

11.3.4 Estimation and Testing with Normal Errors

In this and the next subsections we develop inference procedures under the conditional normal model, the regression model defined by (11.3.22) or (11.3.23).

First, we find the maximum likelihood estimates of the three parameters, α, β, and σ^2. Using the joint pdf in (11.3.24), we see that the log likelihood function is

$$\log L(\alpha, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}{2\sigma^2}.$$

For any fixed value of σ^2, $\log L$ is maximized as a function of α and β by those values, $\hat{\alpha}$ and $\hat{\beta}$, that minimize

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

But this function is just the RSS from Section 11.3.1! There we found that the minimizing values are

$$\hat{\beta} = b = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\alpha} = a = \bar{y} - b\bar{x} = \bar{y} - \hat{\beta}\bar{x}.$$

Thus, the least squares estimators of α and β are also the MLEs of α and β. The values $\hat{\alpha}$ and $\hat{\beta}$ are the maximizing values for any fixed value of σ^2. Now, substituting in the log likelihood, to find the MLE of σ^2 we need to maximize

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{2\sigma^2}.$$

This maximization is similar to finding the MLE of σ^2 in ordinary normal sampling (see Example 7.2.11), and we leave the details to Exercise 11.28. The MLE of σ^2, under the conditional normal model, is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2,$$

the RSS, evaluated at the least squares line, divided by the sample size. Henceforth, when we refer to RSS we mean the RSS evaluated at the least squares line.

In Section 11.3.2, we showed that $\hat{\alpha}$ and $\hat{\beta}$ were linear unbiased estimators of α and β. However, $\hat{\sigma}^2$ is not an unbiased estimator of σ^2. For the calculation of $\mathrm{E}\hat{\sigma}^2$ and in many subsequent calculations, the following lemma will be useful.

Lemma 11.3.2 *Let $Y_1, \ldots, Y_n$ be uncorrelated random variables with $\mathrm{Var}Y_i = \sigma^2$ for all $i = 1, \ldots, n$. Let $c_1, \ldots, c_n$ and $d_1, \ldots, d_n$ be two sets of constants. Then*

$$\mathrm{Cov}\left(\sum_{i=1}^n c_i Y_i, \sum_{i=1}^n d_i Y_i\right) = \left(\sum_{i=1}^n c_i d_i\right)\sigma^2.$$

Proof: This type of result has been encountered before. It is similar to Lemma 5.3.3 and Exercise 11.11. However, here we do not need either normality or independence of $Y_1, \ldots, Y_n$. □

We next find the bias in σ^2. From (11.3.23) we have

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

We define the *residuals from the regression* to be

(11.3.27) $$\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i,$$

and thus

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n}\text{RSS}.$$

It can be calculated (see Exercise 11.29) that

$$\text{E}\hat{\epsilon}_i = 0,$$

and a lengthy calculation (also in Exercise 11.29) gives

(11.3.28)

$$\text{Var}\,\hat{\epsilon}_i = \text{E}\hat{\epsilon}_i^2 = \left(\frac{n-2}{n} + \frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i\bar{x}\right)\right)\sigma^2.$$

Thus,

$$\begin{aligned}
\text{E}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \text{E}\hat{\epsilon}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n} + \frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i\bar{x}\right)\right]\sigma^2 \\
&= \left[\frac{n-2}{n} + \frac{1}{nS_{xx}}\left\{\sum_{j=1}^{n} x_j^2 + \sum_{i=1}^{n} x_i^2 - 2S_{xx} - 2\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\}\right]\sigma^2 \\
&\qquad\qquad \left(\textstyle\sum x_i\bar{x} = \frac{1}{n}(\sum x_i)^2\right) \\
&= \left(\frac{n-2}{n} + 0\right)\sigma^2 \qquad\qquad \left(\textstyle\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = S_{xx}\right) \\
&= \frac{n-2}{n}\sigma^2.
\end{aligned}$$

The MLE $\hat{\sigma}^2$ is a biased estimator of σ^2. The more commonly used estimator of σ^2, which is unbiased, is

(11.3.29) $$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

To develop estimation and testing procedures, based on these estimators, we need to know their sampling distributions. These are summarized in the following theorem.

Theorem 11.3.3 *Under the conditional normal regression model (11.3.22), the sampling distributions of the estimators $\hat{\alpha}$, $\hat{\beta}$, and S^2 are*

$$\hat{\alpha} \sim \mathrm{n}\left(\alpha, \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2\right), \qquad \hat{\beta} \sim \mathrm{n}\left(\beta, \frac{\sigma^2}{S_{xx}}\right),$$

with

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{-\sigma^2 \bar{x}}{S_{xx}}.$$

Furthermore, $(\hat{\alpha}, \hat{\beta})$ and S^2 are independent and

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Proof: We first show that $\hat{\alpha}$ and $\hat{\beta}$ have the indicated normal distributions. The estimators $\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of the independent normal random variables $Y_1, \ldots, Y_n$. Thus, by Corollary 4.6.10, they both have normal distributions. Specifically, in Section 11.3.2, we showed that $\hat{\beta} = \sum_{i=1}^{n} d_i Y_i$, where the d_i are given in (11.3.19), and we also showed that

$$\mathrm{E}\hat{\beta} = \beta \quad \text{and} \quad \mathrm{Var}\,\hat{\beta} = \frac{\sigma^2}{S_{xx}}.$$

The estimator $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$ can be expressed as $\hat{\alpha} = \sum_{i=1}^{n} c_i Y_i$, where

$$c_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}},$$

and thus it is straightforward to verify that

$$\begin{aligned} \mathrm{E}\hat{\alpha} = \sum_{i=1}^{n} c_i \mathrm{E}Y_i &= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)(\alpha + \beta x_i) = \alpha, \\ \mathrm{Var}\,\hat{\alpha} = \sigma^2 \sum_{i=1}^{n} c_i^2 &= \sigma^2\left[\frac{1}{nS_{xx}}\sum_{i=1}^{n} x_i^2\right], \end{aligned}$$

showing that $\hat{\alpha}$ and $\hat{\beta}$ have the specified distributions. Also, $\mathrm{Cov}(\hat{\alpha}, \hat{\beta})$ is easily calculated using Lemma 11.3.2. Details are left to Exercise 11.30.

We next show that $\hat{\alpha}$ and $\hat{\beta}$ are independent of S^2, a fact that will follow from Lemma 11.3.2 and Lemma 5.3.3. From the definition of $\hat{\epsilon}_i$ in (11.3.27), we can write

$$\hat{\epsilon}_i = \sum_{j=1}^{n} [\delta_{ij} - (c_j + d_j x_i)] Y_i, \tag{11.3.30}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}, \quad c_j = \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{S_{xx}}, \quad \text{and} \quad d_j = \frac{(x_j - \bar{x})}{S_{xx}}.$$

Since $\hat{\alpha} = \sum c_i Y_i$ and $\hat{\beta} = \sum d_i Y_i$, application of Lemma 11.3.2 together with some algebra will show that

$$\text{Cov}(\hat{\epsilon}_i, \hat{\alpha}) = \text{Cov}(\hat{\epsilon}_i, \hat{\beta}) = 0, \quad i = 1, \ldots, n.$$

Details are left to Exercise 11.31. Thus, it follows from Lemma 5.3.3 that, under normal sampling, $S^2 = \sum \hat{\epsilon}_i^2/(n-2)$ is independent of $\hat{\alpha}$ and $\hat{\beta}$.

To prove that $(n-2)S^2/\sigma^2 \sim \chi^2_{n-2}$, we write $(n-2)S^2$ as the sum of $n-2$ independent random variables, each of which has a χ^2_1 distribution. That is, we find constants $a_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, n-2$, that satisfy

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{j=1}^{n-2} \left(\sum_{i=1}^{n} a_{ij} Y_i \right)^2, \tag{11.3.31}$$

where

$$\sum_{i=1}^{n} a_{ij} = 0, \quad j = 1, \ldots, n-2, \qquad \text{and} \qquad \sum_{i=1}^{n} a_{ij} a_{ij'} = 0, \quad j \neq j'.$$

The details are somewhat involved because of the general nature of the x_is. We omit details. □

The RSS from the *linear* regression contains information about the worth of a polynomial fit of a higher order, over and above a linear fit. Since, in this model, we assume that the population regression is linear, the variation in this higher-order fit is just random variation. Robson (1959) gives a general recursion formula for finding coefficients for such higher-order polynomial fits, a formula that can be adapted to explicitly find the a_{ij}s of (11.3.31). Alternatively, Cochran's Theorem (see Miscellanea 11.5.1) can be used to establish that $\sum \hat{\epsilon}_i^2/\sigma^2 \sim \chi^2_{n-2}$.

Inferences regarding the two parameters α and β are usually based on the following two Student's t distributions. Their derivations follow immediately from the normal and χ^2 distributions and the independence in Theorem 11.3.3. We have

$$\frac{\hat{\alpha} - \alpha}{S\sqrt{(\sum_{i=1}^{n} x_i^2)/(nS_{xx})}} \sim t_{n-2} \tag{11.3.32}$$

and

$$\frac{\hat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t_{n-2}. \tag{11.3.33}$$

The joint distribution of these two t statistics is called a *bivariate Student's t distribution.* This distribution is derived in a manner analogous to the univariate case. We use the fact that the joint distribution of $\hat{\alpha}$ and $\hat{\beta}$ is bivariate normal and the same variance estimate S is used in both univariate t statistics. This joint distribution would be used if we wanted to do simultaneous inference regarding α and β. However, we shall deal only with the inferences regarding one parameter at a time.

Usually there is more interest in β than in α. The parameter α is the expected value of Y at $x = 0$, $\text{E}(Y|x = 0)$. Depending on the problem, this may or may not

be an interesting quantity. In particular, the value $x = 0$ may not be a reasonable value for the predictor variable. However, β is the rate of change of $\mathrm{E}(Y|x)$ as a function of x. That is, β is the amount that $\mathrm{E}(Y|x)$ changes if x is changed by one unit. Thus, this parameter relates to the entire range of x values and contains the information about whatever linear relationship exists between Y and x. (See Exercise 11.33.) Furthermore, the value $\beta = 0$ is of particular interest.

If $\beta = 0$, then $\mathrm{E}(Y|x) = \alpha + \beta x = \alpha$ and $Y \sim \mathrm{n}(\alpha, \sigma^2)$, which does not depend on x. In a well-thought-out experiment leading to a regression analysis we do not expect this to be the case, but we would be interested in knowing this if it were true.

The test that $\beta = 0$ is quite similar to the ANOVA test that all treatments are equal. In the ANOVA the null hypothesis states that the treatments are unrelated to the response *in any way*, while in linear regression the null hypothesis $\beta = 0$ states that the treatments (x) are unrelated to the response in a linear way.

To test

$$H_0: \quad \beta = 0 \qquad \text{versus} \qquad H_1: \quad \beta \neq 0 \tag{11.3.34}$$

using (11.3.33), we reject H_0 at level α if

$$\left| \frac{\hat{\beta} - 0}{S/\sqrt{S_{xx}}} \right| > t_{n-2,\alpha/2}$$

or, equivalently, if

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} > F_{1,n-2,\alpha}. \tag{11.3.35}$$

Recalling the formula for $\hat{\beta}$ and that $\mathrm{RSS} = \sum \hat{\epsilon}_i^2$, we have

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} = \frac{S_{xy}^2/S_{xx}}{\mathrm{RSS}/(n-2)} = \frac{\text{Regression sum of squares}}{\text{Residual sum of squares/df}}.$$

This last formula is summarized in the *regression ANOVA table*, which is like the ANOVA tables encountered in Section 11.2. For simple linear regression, the table, resulting in the test given in (11.3.35), is given in Table 11.3.2. Note that the table involves only a hypothesis about β. The parameter α and the estimate $\hat{\alpha}$ play the same role here as the grand mean did in Section 11.2. They merely serve to locate the overall level of the data and are "corrected" for in the sums of squares.

Example 11.3.4 (Continuation of Example 11.3.1) The regression ANOVA for the grape crop yield data follows.

ANOVA table for grape data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression	1	6.66	6.66	50.23
Residual	10	1.33	.133	
Total	11	7.99		

This shows a highly significant slope of the regression line. ‖

Table 11.3.2. *ANOVA table for simple linear regression*

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression (slope)	1	Reg. SS $= S_{xy}^2/S_{xx}$	MS(Reg) = Reg. SS	$F = \dfrac{\text{MS(Reg)}}{\text{MS(Resid)}}$
Residual	$n-2$	RSS $= \sum \hat{\epsilon}_i^2$	MS(Resid) = RSS$/(n-2)$	
Total	$n-1$	SST $= \sum (y_i - \bar{y})^2$		

We draw one final parallel with the analysis of variance. It may not be obvious from Table 11.3.2, but the partitioning of the sum of squares of the ANOVA has an analogue in regression. We have

Total sum of squares = Regression sum of squares + Residual sum of squares

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{11.3.36}$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. Notice the similarity of these sums of squares to those in ANOVA. The total sum of squares is, of course, the same. The RSS measures deviation of the fitted line from the observed values, and the regression sum of squares, analogous to the ANOVA treatment sum of squares, measures the deviation of predicted values ("treatment means") from the grand mean. Also, as in the ANOVA, the sum of squares identity is valid because of the disappearance of the cross-term (see Exercise 11.34). The total and residual sums of squares in (11.3.36) are clearly the same as in Table 11.3.2. But the regression sum of squares looks different. However, they are equal (see Exercise 11.34); that is,

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}.$$

The expression S_{xy}^2/S_{xx} is easier to use for computing and provides the link with the t test. But $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the more easily interpreted expression.

A statistic that is used to quantify how well the fitted line describes the data is the *coefficient of determination.* It is defined as the ratio of the regression sum of squares to the total sum of squares. It is usually referred to as r^2 and can be written in the various forms

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

The coefficient of determination measures the proportion of the total variation in $y_1, \ldots, y_n$ (measured by S_{yy}) that is explained by the fitted line (measured by the

regression sum of squares). From (11.3.36), $0 \le r^2 \le 1$. If $y_1, \ldots, y_n$ all fall exactly on the fitted line, then $y_i = \hat{y}_i$ for all i and $r^2 = 1$. If $y_1, \ldots, y_n$ are not close to the fitted line, then the residual sum of squares will be large and r^2 will be near 0. The coefficient of determination can also be (perhaps more straightforwardly) derived as the square of the sample correlation coefficient of the n pairs $(y_1, x_1), \ldots, (y_n, x_n)$ or of the n pairs $(y_1, \hat{y}_1), \ldots, (y_n, \hat{y}_n)$.

Expression (11.3.33) can be used to construct a $100(1-\alpha)\%$ confidence interval for β given by

$$\hat{\beta} - t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{xx}}} < \beta < \hat{\beta} + t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{xx}}}. \tag{11.3.37}$$

Also, a level α test of $H_0: \beta = \beta_0$ versus $H_1: \beta \neq \beta_0$ rejects H_0 if

$$\left|\frac{\hat{\beta} - \beta_0}{S/\sqrt{S_{xx}}}\right| > t_{n-2,\alpha/2}. \tag{11.3.38}$$

As mentioned above, it is common to test $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ to determine if there is some linear relationship between the predictor and response variables. However, the above test is more general, since any value of β_0 can be specified. The regression ANOVA, which is locked into a "recipe," can test only $H_0: \beta = 0$.

11.3.5 Estimation and Prediction at a Specified $x = x_0$

Associated with a specified value of the predictor variable, say $x = x_0$, there is a population of Y values. In fact, according to the conditional normal model, a random observation from this population is $Y \sim \mathrm{n}(\alpha + \beta x_0, \sigma^2)$. After observing the regression data $(x_1, y_1), \ldots, (x_n, y_n)$ and estimating the parameters α, β, and σ^2, perhaps the experimenter is going to set $x = x_0$ and obtain a new observation, call it Y_0. There might be interest in estimating the mean of the population from which this observation will be drawn, or even predicting what this observation will be. We will now discuss these types of inferences.

We assume that $(x_1, Y_1), \ldots, (x_n, Y_n)$ satisfy the conditional normal regression model, and based on these n observations we have the estimates $\hat{\alpha}, \hat{\beta}$, and S^2. Let x_0 be a specified value of the predictor variable. First, consider estimating the mean of the Y population associated with x_0, that is, $\mathrm{E}(Y|x_0) = \alpha + \beta x_0$. The obvious choice for our point estimator is $\hat{\alpha} + \hat{\beta}x_0$. This is an unbiased estimator since $\mathrm{E}(\hat{\alpha} + \hat{\beta}x_0) = \mathrm{E}\hat{\alpha} + (\mathrm{E}\hat{\beta})x_0 = \alpha + \beta x_0$. Using the moments given in Theorem 11.3.3, we can also calculate

$$\begin{aligned}
\operatorname{Var}(\hat{\alpha} + \hat{\beta}x_0) &= \operatorname{Var}\hat{\alpha} + (\operatorname{Var}\hat{\beta})x_0^2 + 2x_0\operatorname{Cov}(\hat{\alpha}, \hat{\beta}) \\
&= \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2 + \frac{\sigma^2 x_0^2}{S_{xx}} - \frac{2\sigma^2 x_0 \bar{x}}{S_{xx}} \\
&= \frac{\sigma^2}{S_{xx}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 + \bar{x}^2 - 2x_0\bar{x} + x_0^2\right) \qquad (\pm\bar{x})
\end{aligned}$$

$$= \frac{\sigma^2}{S_{xx}}\left(\frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] + (x_0 - \bar{x})^2\right) \qquad \binom{\text{recombine}}{\text{terms}}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right). \qquad \left(\sum x_i^2 - \tfrac{1}{n}(\sum x_i)^2 = S_{xx}\right)$$

Finally, since $\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of $Y_1, \ldots, Y_n$, so is $\hat{\alpha} + \hat{\beta}x_0$. Thus $\hat{\alpha} + \hat{\beta}x_0$ has a normal distribution, specifically,

$$\hat{\alpha} + \hat{\beta}x_0 \sim \mathrm{n}\left(\alpha + \beta x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right). \tag{11.3.39}$$

By Theorem 11.3.3, $(\hat{\alpha}, \hat{\beta})$ and S^2 are independent. Thus S^2 is also independent of $\hat{\alpha} + \hat{\beta}x_0$ (Theorem 4.6.12) and

$$\frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

This pivot can be inverted to give the $100(1-\alpha)\%$ confidence interval for $\alpha + \beta x_0$,

$$\hat{\alpha} + \hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$\leq \quad \alpha + \beta x_0 \quad \leq \quad \hat{\alpha} + \hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \tag{11.3.40}$$

The length of the confidence interval for $\alpha + \beta x_0$ depends on the values of $x_1, \ldots, x_n$ through the value of $(x_0 - \bar{x})^2/S_{xx}$. It is clear that the length of the interval is shorter if x_0 is near $\bar{x}$ and minimized at $x_0 = \bar{x}$. Thus, in designing the experiment, the experimenter should choose the values $x_1, \ldots, x_n$ so that the value x_0, at which the mean is to be estimated, is at or near $\bar{x}$. It is only reasonable that we can estimate more precisely near the center of the data we observed.

A type of inference we have not discussed until now is *prediction* of an, as yet, unobserved random variable Y, a type of inference that is of interest in a regression setting. For example, suppose that x is a college applicant's measure of high school performance. A college admissions officer might want to use x to predict Y, the student's grade point average after one year of college. Clearly, Y has not been observed yet since the student has not even been admitted! The college has data on former students, $(x_1, y_1), \ldots, (x_n, y_n)$, giving their high school performances and one-year GPAs. These data might be used to predict the new student's GPA.

Definition 11.3.5 A $100(1-\alpha)\%$ *prediction interval* for an unobserved random variable Y based on the observed data $\mathbf{X}$ is a random interval $[L(\mathbf{X}), U(\mathbf{X})]$ with the property that

$$P_\theta(L(\mathbf{X}) \leq Y \leq U(\mathbf{X})) \geq 1 - \alpha$$

for all values of the parameter θ.

Note the similarity in the definitions of a prediction interval and a confidence interval. The difference is that a prediction interval is an interval on a random variable, rather than a parameter. Intuitively, since a random variable is more variable than a parameter (which is constant), we expect a prediction interval to be wider than a confidence interval of the same level. In the special case of linear regression, we see that this is the case.

We assume that the new observation Y_0 to be taken at $x = x_0$ has a $\mathrm{n}(\alpha + \beta x_0, \sigma^2)$ distribution, independent of the previous data, $(x_1, Y_1), \ldots, (x_n, Y_n)$. The estimators $\hat{\alpha}, \hat{\beta}$, and S^2 are calculated from the previous data and, thus, Y_0 is independent of $\hat{\alpha}$, $\hat{\beta}$, and S^2. Using (11.3.39), we find that $Y_0 - (\hat{\alpha} + \hat{\beta}x_0)$ has a normal distribution with mean $\mathrm{E}(Y_0 - (\hat{\alpha} + \hat{\beta}x_0)) = \alpha + \beta x_0 - (\alpha + \beta x_0) = 0$ and variance

$$\mathrm{Var}\,(Y_0 - (\hat{\alpha} + \hat{\beta}x_0)) = \mathrm{Var}\,Y_0 + \mathrm{Var}\,(\hat{\alpha} + \hat{\beta}x_0) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right).$$

Using the independence of S^2 and $Y_0 - (\hat{\alpha} + \hat{\beta}x_0)$, we see that

$$T = \frac{Y_0 - (\hat{\alpha} + \hat{\beta}x_0)}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

which can be rearranged in the usual way to obtain the $100(1 - \alpha)\%$ prediction interval,

$$\begin{aligned} &\hat{\alpha} + \hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \\ (11.3.41) \qquad &< Y_0 < \hat{\alpha} + \hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \end{aligned}$$

Since the endpoints of this interval depend only on the observed data, (11.3.41) defines a prediction interval for the new observation Y_0.

11.3.6 Simultaneous Estimation and Confidence Bands

In the previous section we looked at prediction at a single value x_0. In some circumstances, however, there may be interest in prediction at many x_0s. For example, in the previously mentioned grade point average prediction problem, an admissions officer probably has interest in predicting the grade point average of many applicants, which naturally leads to prediction at many x_0s.

The problem encountered is the (by now) familiar problem of simultaneous inference. That is, how do we control the overall confidence level for the simultaneous inference? In the previous section, we saw that a $1 - \alpha$ confidence interval for the mean of the Y population associated with x_0, that is, $\mathrm{E}(Y|x_0) = \alpha + \beta x_0$, is given by

$$\hat{\alpha}+\hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$$

$$< \quad \alpha+\beta x_0 \quad < \quad \hat{\alpha}+\hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}.$$

Now suppose that we want to make an inference about the Y population mean at a number of x_0 values. For example, we might want intervals for $\mathrm{E}(Y|x_{0i}), i = 1, \ldots, m$. We know that if we set up m intervals as above, each at level $1-\alpha$, the overall inference will not be at the $1-\alpha$ level.

A simple and reasonably good solution is to use the Bonferroni Inequality, as used in Example 11.2.9. Using the inequality, we can state that the probability is at least $1-\alpha$ that

$$\hat{\alpha}+\hat{\beta}x_{0i} - t_{n-2,\alpha/(2m)}S\sqrt{\frac{1}{n}+\frac{(x_{0i}-\bar{x})^2}{S_{xx}}}$$

$$< \quad \alpha+\beta x_{0i} \quad < \quad \hat{\alpha}+\hat{\beta}x_{0i} + t_{n-2,\alpha/(2m)}S\sqrt{\frac{1}{n}+\frac{(x_{0i}-\bar{x})^2}{S_{xx}}} \tag{11.3.42}$$

simultaneously for $i = 1, \ldots, m$. (See Exercise 11.39.)

We can take simultaneous inference in regression one step further. Realize that our assumption about the population regression line implies that the equation $\mathrm{E}(Y|x) = \alpha+\beta x$ holds *for all* x; hence, we should be able to make inferences at all x. Thus, we want to make a statement like (11.3.42), but we want it to hold for all x. As might be expected, as he did for the ANOVA, Scheffé derived a solution for this problem. We summarize the result for the case of simple linear regression in the following theorem.

Theorem 11.3.6 *Under the conditional normal regression model (11.3.22), the probability is at least* $1-\alpha$ *that*

$$\hat{\alpha}+\hat{\beta}x - M_\alpha S\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$$

$$< \quad \alpha+\beta x \quad < \quad \hat{\alpha}+\hat{\beta}x + M_\alpha S\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}} \tag{11.3.43}$$

simultaneously for all x, *where* $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$.

Proof: If we rearrange terms, it should be clear that the conclusion of the theorem is true if we can find a constant M_α that satisfies

$$P\left(\frac{\Big((\hat{\alpha}+\hat{\beta}x)-(\alpha+\beta x)\Big)^2}{S^2\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right]} \le M_\alpha^2 \text{ for all } x\right) = 1-\alpha$$

or, equivalently,

$$P\left(\max_x \frac{\Big((\hat\alpha+\hat\beta x)-(\alpha+\beta x)\Big)^2}{S^2\left[\frac{1}{n}+\frac{(x-\bar x)^2}{S_{xx}}\right]} \le M_\alpha^2\right) = 1-\alpha.$$

The parameterization given in Exercise 11.32, which results in independent estimators for α and β, makes the above maximization easier. Write

$$\hat\alpha+\hat\beta x = \bar Y + \hat\beta(x-\bar x),$$
$$\alpha+\beta x = \mu_{\bar Y} + \beta(x-\bar x), \qquad (\mu_{\bar Y} = \mathrm{E}\,\bar Y = \alpha+\beta\bar x)$$

and, for notational convenience, define $t = x-\bar x$. We then have

$$\frac{\Big((\hat\alpha+\hat\beta x)-(\alpha+\beta x)\Big)^2}{S^2\left[\frac{1}{n}+\frac{(x-\bar x)^2}{S_{xx}}\right]} = \frac{\Big((\bar Y-\mu_{\bar Y})+(\hat\beta-\beta)t\Big)^2}{S^2\left[\frac{1}{n}+\frac{t^2}{S_{xx}}\right]},$$

and we want to find M_α to satisfy

$$P\left(\max_t \frac{\Big((\bar Y-\mu_{\bar Y})+(\hat\beta-\beta)t\Big)^2}{S^2\left[\frac{1}{n}+\frac{t^2}{S_{xx}}\right]} \le M_\alpha^2\right) = 1-\alpha.$$

Note that S^2 plays no role in the maximization, merely being a constant. Applying the result of Exercise 11.40, a direct application of calculus, we obtain

$$\max_t \frac{\Big((\bar Y-\mu_{\bar Y})+(\hat\beta-\beta)t\Big)^2}{S^2\left[\frac{1}{n}+\frac{t^2}{S_{xx}}\right]} = \frac{n(\bar Y-\mu_{\bar Y})^2 + S_{xx}(\hat\beta-\beta)^2}{S^2}$$

$$= \frac{\frac{(\bar Y-\mu_{\bar Y})^2}{\sigma^2/n} + \frac{(\hat\beta-\beta)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2}. \qquad \text{(multiply by } \sigma^2/\sigma^2\text{)} \tag{11.3.44}$$

From Theorem 11.3.3 and Exercise 11.32, we see that this last expression is the quotient of independent chi squared random variables, the denominator being divided by its degrees of freedom. The numerator is the sum of two independent random variables, each of which has a χ^2_1 distribution. Thus the numerator is distributed as χ^2_2, the distribution of the quotient is

$$\frac{\frac{(\bar Y-\mu_{\bar Y})^2}{\sigma^2/n} + \frac{(\hat\beta-\beta)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2} \sim 2F_{2,n-2},$$

and

$$P\left(\max_t \frac{\Big((\bar Y-\mu_{\bar Y})+(\hat\beta-\beta)t\Big)^2}{S^2\left[\frac{1}{n}+\frac{t^2}{S_{xx}}\right]} \le M_\alpha^2\right) = 1-\alpha$$

if $M_\alpha = \sqrt{2F_{2,n-2}}$, proving the theorem. □

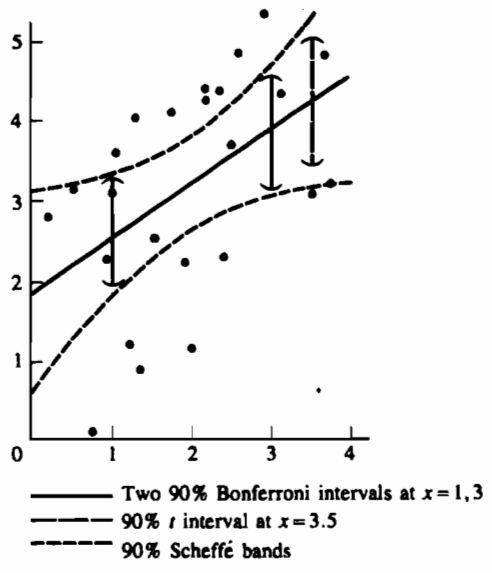

Figure 11.3.3. *Scheffé bands, t interval (at $x = 3.5$), and Bonferroni intervals (at $x = 1$ and $x = 3$) for data in Table 11.3.1*

Since (11.3.43) is true for all x, it actually gives a *confidence band* on the entire population regression line. That is, as a confidence interval covers a single-valued parameter, a confidence band covers an entire line with a band. An example of the Scheffé band is given in Figure 11.3.3, along with two Bonferroni intervals and a single t interval. Notice that, although it is not the case in Figure 11.3.3, it is possible for the Bonferroni intervals to be *wider* than the Scheffé bands, even though the Bonferroni inference (necessarily) pertains to fewer intervals. This will be the case whenever

$$t_{n-2,\alpha/(2m)} > 2F_{2,n-2,\alpha},$$

where m is defined as in (11.3.42). The inequality will always be satisfied for large enough m, so there will always be a point where it pays to switch from Bonferroni to Scheffé, even if there is interest in only a finite number of xs. This "phenomenon," that we seem to get something for nothing, occurs because the Bonferroni Inequality is an all-purpose bound while the Scheffé band is an exact solution for the problem at hand. (The actual coverage probability for the Bonferroni intervals is higher than $1 - \alpha$.) There are many variations on the Scheffé band. Some variations have different shapes and some guarantee coverage for only a particular interval of x values. See the Miscellanea section for a discussion of these alternative bands.

In theory, the proof of Theorem 11.3.6, with suitable modifications, can result in simultaneous prediction intervals. (In fact, the maximization of the function in Exercise 11.40 gives the result almost immediately.) The problem, however, is that the resulting statistic does not have a particularly nice distribution.

Finally, we note a problem about using procedures like the Scheffé band to make inferences at x values that are outside the range of the observed xs. Such procedures

are based on the assumption that we *know* the population regression function is linear for all x. Although it may be reasonable to assume the regression function is linear over the range of xs observed, *extrapolation* to xs outside the observed range is usually unwise. (Since there are no data outside the observed range, we cannot check whether the regression becomes nonlinear.) This caveat also applies to the procedures in Section 11.3.5.

11.4 Exercises

11.1 An ANOVA variance-stabilizing transformation stabilizes variances in the following approximate way. Let Y have mean θ and variance $v(\theta)$.

(a) Use arguments as in Section 10.1.3 to show that a one-term Taylor series approximation of the variance of $g(y)$ is given by $\text{Var}\,(g(Y)) = [\frac{d}{d\theta}g(\theta)]^2 v(\theta)$.

(b) Show that the approximate variance of $g^*(Y)$ is independent of θ, where $g^*(y) = \int [1/\sqrt{v(y)}]dy$.

11.2 Verify that the following transformations are approximately variance-stabilizing in the sense of Exercise 11.1.

(a) $Y \sim$ Poisson, $g^*(y) = \sqrt{y}$

(b) $Y \sim$ binomial(n, p), $g^*(y) = \sin^{-1}(\sqrt{y/n})$

(c) Y has variance $v(\theta) = K\theta^2$ for some constant K, $g^*(y) = \log(y)$.

(Conditions for the existence of variance-stabilizing transformations go back at least to Curtiss 1943, with refinements given by Bar-Lev and Enis 1988, 1990.)

11.3 The Box–Cox family of power transformations (Box and Cox 1964) is defined by

$$g^*_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0, \end{cases}$$

where λ is a free parameter.

(a) Show that, for each y, $g^*_\lambda(y)$ is continuous in λ. In particular, show that

$$\lim_{\lambda \to 0} (y^\lambda - 1)/\lambda = \log y.$$

(b) Find the function $v(\theta)$, the approximate variance of Y, that $g^*_\lambda(y)$ stabilizes. (Note that $v(\theta)$ will most likely also depend on λ.)

Analysis of transformed data in general and the Box–Cox power transformation in particular has been the topic of some controversy in the statistical literature. See Bickel and Doksum (1981), Box and Cox (1982), and Hinkley and Runger (1984).

11.4 A most famous (and useful) variance-stabilizing transformation is Fisher's z-transformation, which we have already encountered in Exercise 10.17. Here we will look at a few more details. Suppose that (X, Y) are bivariate normal with correlation coefficient ϱ and sample correlation r.

(a) Starting from Exercise 10.17, part (d), use the Delta Method to show that

$$\frac{1}{2}\left[\log\left(\frac{1+r}{1-r}\right) - \log\left(\frac{1+\varrho}{1-\varrho}\right)\right]$$

is approximately normal with mean 0 and variance $1/n$.

(b) Fisher actually used a somewhat more accurate expansion (Stuart and Ord 1987, Section 16.33) and established that the quantity in part (a) is approximately normal with

$$\text{mean} = \frac{\varrho}{2(n-1)} \quad \text{and} \quad \text{variance} = \frac{1}{n-1} + \frac{4-\varrho^2}{2(n-1)^2}.$$

Show that for small ϱ and moderate n, we can approximate this mean and variance by 0 and $1/(n-3)$, which is the most popular form of Fisher's z-transformation.

11.5 Suppose that random variables Y_{ij} are observed according to the overparameterized oneway ANOVA model in (11.2.2). Show that, without some restriction on the parameters, this model is not identifiable by exhibiting two distinct collections of parameters that lead to exactly the same distribution of the Y_{ij}s.

11.6 Under the oneway ANOVA assumptions:

(a) Show that the set of statistics $(\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{k\cdot}, S_p^2)$ is sufficient for $(\theta_1, \theta_2, \ldots, \theta_k, \sigma^2)$.

(b) Show that $S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2$ is independent of each $\bar{Y}_{i\cdot}, i = 1, \ldots, k$. (See Lemma 5.3.3).

(c) If σ^2 is known, explain how the ANOVA data are equivalent to their canonical version in Miscellanea 11.5.6.

11.7 Complete the proof of Theorem 11.2.8 by showing that

$$\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i \left((\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta})\right)^2 \sim \chi^2_{k-1}.$$

(*Hint*: Define $\bar{U}_i = \bar{Y}_{i\cdot} - \theta_i, i = 1, \ldots, k$. Show that $\bar{U}_i$ are independent $\mathrm{n}(0, \sigma^2/n_i)$. Then adapt the induction argument of Lemma 5.3.2 to show that $\sum n_i(\bar{U}_i - \bar{\bar{U}})^2/\sigma^2 \sim \chi^2_{k-1}$, where $\bar{\bar{U}} = \sum n_i \bar{U}_i / \sum n_i$.)

11.8 Show that under the oneway ANOVA assumptions, for any set of constants $\mathbf{a} = (a_1, \ldots, a_k)$, the quantity $\sum a_i \bar{Y}_{i\cdot}$ is normally distributed with mean $\sum a_i \theta_i$ and variance $\sigma^2 \sum a_i^2/n_i$. (See Corollary 4.6.10.)

11.9 Using an argument similar to that which led to the t test in (11.2.7), show how to construct a t test for

(a) $H_0: \sum a_i\theta_i = \delta$ versus $H_1: \sum a_i\theta_i \neq \delta$.

(b) $H_0: \sum a_i\theta_i \leq \delta$ versus $H_1: \sum a_i\theta_i > \delta$, where δ is a specified constant.

11.10 Suppose we have a oneway ANOVA with five treatments. Denote the treatment means by $\theta_1, \ldots, \theta_5$, where θ_1 is a control and $\theta_2, \ldots, \theta_5$ are alternative new treatments, and assume that an equal number of observations per treatment is taken. Consider the four contrasts $\sum a_i\theta_i$ defined by

$$\mathbf{a}_1 = \left(1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}\right),$$

$$\mathbf{a}_2 = \left(0, 1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}\right),$$

$$\mathbf{a}_3 = \left(0, 0, 1, -\frac{1}{2}, -\frac{1}{2}\right),$$

$$\mathbf{a}_4 = (0, 0, 0, 1, -1).$$

(a) Argue that the results of the four t tests using these contrasts can lead to conclusions about the ordering of $\theta_1, \ldots, \theta_5$. What conclusions might be made?

(b) Show that any two contrasts $\sum a_i \bar{Y}_{i\cdot}$ formed from the four $\mathbf{a}_i$s in part (a) are uncorrelated. (Recall that these are called orthogonal contrasts.)

(c) For the fertilizer experiment of Example 11.2.3, the following contrasts were planned:

$$\mathbf{a}_1 = (-1, 1, 0, 0, 0),$$
$$\mathbf{a}_2 = \left(0, -1, \frac{1}{2}, \frac{1}{2}, 0\right),$$
$$\mathbf{a}_3 = (0, 0, 1, -1, 0),$$
$$\mathbf{a}_4 = (0, -1, 0, 0, 1,).$$

Show that these contrasts are not orthogonal. Interpret these contrasts in the context of the fertilizer experiment, and argue that they are a sensible set of contrasts.

11.11 For any sets of constants $\mathbf{a} = (a_1, \ldots, a_k)$ and $\mathbf{b} = (b_1, \ldots, b_k)$, show that under the oneway ANOVA assumptions,

$$\mathrm{Cov}\left(\sum a_i \bar{Y}_{i\cdot}, \sum b_i \bar{Y}_{i\cdot}\right) = \sigma^2 \sum \frac{a_i b_i}{n_i}.$$

Hence, in the oneway ANOVA, contrasts are uncorrelated (orthogonal) if $\sum a_i b_i / n_i = 0$.

11.12 Suppose that we have a oneway ANOVA with equal numbers of observations on each treatment, that is, $n_i = n, i = 1, \ldots, k$. In this case the F test can be considered an average t test.

(a) Show that a t test of $H_0 : \theta_i = \theta_{i'}$ versus $H_1 : \theta_i \neq \theta_{i'}$ can be based on the statistic

$$t^2_{ii'} = \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot})^2}{S_p^2(2/n)}.$$

(b) Show that

$$\frac{1}{k(k-1)} \sum_{i,i'} t^2_{ii'} = F,$$

where F is the usual ANOVA F statistic. (*Hint*: See Exercise 5.8(a).) (*Communicated by George McCabe, who learned it from John Tukey.*)

11.13 Under the oneway ANOVA assumptions, show that the likelihood ratio test of $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$ is given by the F test of (11.2.14).

11.14 The Scheffé simultaneous interval procedure actually works for all linear combinations, not just contrasts. Show that under the oneway ANOVA assumptions, if $\mathbf{M} = \sqrt{kF_{k,N-k,\alpha}}$ (note the change in the numerator degrees of freedom), then the probability is $1 - \alpha$ that

$$\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \mathbf{M}\sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}} \leq \sum_{i=1}^k a_i \theta_i \leq \sum_{i=1}^k a_i \bar{Y}_{i\cdot} + \mathbf{M}\sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}$$

simultaneously for all $\mathbf{a} = (a_1, \ldots, a_k)$. It is probably easiest to proceed by first establishing, in the spirit of Lemma 11.2.7, that if $v_1, \ldots, v_k$ are constants and $c_1, \ldots, c_k$ are positive constants, then

$$\max_{\mathbf{a}} \left\{ \frac{\left(\sum_{i=1}^k a_i v_i\right)^2}{\sum_{i=1}^k a_i^2/c_i} \right\} = \sum_{i=1}^k c_i v_i^2.$$

The proof of Theorem 11.2.10 can then be adapted to establish the result.

11.15 (a) Show that for the t and F distributions, for any ν, α, and k,

$$t_{\nu,\alpha/2} \leq \sqrt{(k-1)F_{k-1,\nu,\alpha}}.$$

(Recall the relationship between the t and the F. This inequality is a consequence of the fact that the distributions $kF_{k,\nu}$ are stochastically increasing in k for fixed ν but is actually a weaker statement. See Exercise 5.19.)

(b) Explain how the above inequality shows that the simultaneous Scheffé intervals are always wider than the single-contrast intervals.

(c) Show that it also follows from the above inequality that Scheffé tests are less powerful than t tests.

11.16 In Theorem 11.2.5 we saw that the ANOVA null is equivalent to all contrasts being 0. We can also write the ANOVA null as the intersection over another set of hypotheses.

(a) Show that the hypotheses

$$H_0\colon \quad \theta_1 = \theta_2 = \cdots = \theta_k \qquad \text{versus} \qquad H_1\colon \quad \theta_i \neq \theta_j \text{ for some } i, j$$

and the hypotheses

$$H_0\colon \quad \theta_i - \theta_j = 0 \text{ for all } i, j \qquad \text{versus} \qquad H_1\colon \quad \theta_i - \theta_j \neq 0 \text{ for some } i, j$$

are equivalent.

(b) Express H_0 and H_1 of the ANOVA test as unions and intersections of the sets

$$\Theta_{ij} = \{\theta = (\theta_1, \ldots, \theta_k)\colon \theta_i - \theta_j = 0\}.$$

Describe how these expressions can be used to construct another (different) union–intersection test of the ANOVA null hypothesis. (See Miscellanea 11.5.2.)

11.17 A multiple comparison procedure called the *Protected LSD* (Protected Least Significant Difference) is performed as follows. If the ANOVA F test rejects H_0 at level α, then for each pair of means θ_i and $\theta_{i'}$, declare the means different if

$$\frac{|\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}|}{\sqrt{S_p^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}} > t_{\alpha/2, N-k}.$$

Note that each t test is done at the same α level as the ANOVA F test. Here we are using an *experimentwise* α level, where

$$\text{experimentwise } \alpha = P\left(\begin{array}{c} \text{at least one false} \\ \text{assertion of difference} \end{array} \middle| \begin{array}{c} \text{all the means} \\ \text{are equal} \end{array} \right).$$

(a) Prove that no matter how many means are in the experiment, simultaneous inference from the Protected LSD is made at level α.

(b) The *ordinary* (or *unprotected*) LSD simply does the individual t tests, at level α, no matter what the outcome of the ANOVA F test. Show that the ordinary LSD can have an experimentwise error rate greater than α. (The unprotected LSD does maintain a *comparisonwise* error rate of α.)

(c) Perform the LSD procedure on the fish toxin data of Example 11.2.1. What are the conclusions?

11.18 Demonstrate that "data snooping," that is, testing hypotheses that are suggested by the data, is generally not a good practice.

(a) Show that, for any random variable Y and constants a and b with $a > b$ and $P(Y > b) < 1$, $P(Y > a|Y > b) > P(Y > a)$.

(b) Apply the inequality in part (a) to the size of a data-suggested hypothesis test by letting Y be a test statistic and a be a cutoff point.

11.19 Let $X_i \sim \text{gamma}(\lambda_i, 1)$ independently for $i = 1, \ldots, n$. Define $Y_i = X_{i+1} / \left(\sum_{j=1}^{i} X_j\right)$, $i = 1, \ldots, n-1$, and $Y_n = \sum_{i=1}^{n} X_i$.

(a) Find the joint and marginal distributions of $Y_i, i = 1, \ldots, n$.

(b) Connect your results to any distributions that are commonly employed in the ANOVA.

11.20 Assume the oneway ANOVA null hypothesis is true.

(a) Show that $\sum n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2/(k-1)$ gives an unbiased estimate of σ^2.

(b) Show how to use the method of Example 5.3.5 to derive the ANOVA F test.

11.21 (a) Illustrate the partitioning of the sums of squares in the ANOVA by calculating the complete ANOVA table for the following data. To determine diet quality, male weanling rats were fed diets with various protein levels. Each of 15 rats was randomly assigned to one of three diets, and their weight gain in grams was recorded.

Diet protein level		
Low	Medium	High
3.89	8.54	20.39
3.87	9.32	24.22
3.26	8.76	30.91
2.70	9.30	22.78
3.82	10.45	26.33

(b) Analytically verify the partitioning of the ANOVA sums of squares by completing the proof of Theorem 11.2.11.

(c) Illustrate the relationship between the t and F statistics, given in Exercise 11.12(b), using the data of part (a).

11.22 Calculate the expected values of MSB and MSW given in the oneway ANOVA table. (Such expectations are formally known as *expected mean squares* and can be used to help identify F tests in complicated ANOVAs. An algorithm exists for calculating expected mean squares. See, for example, Kirk 1982 for details about the algorithm.)

11.23 Use the model in Miscellanea 11.5.3.

(a) Show that the mean and variance of Y_{ij} are $EY_{ij} = \mu + \tau_i$ and $\text{Var}\, Y_{ij} = \sigma_B^2 + \sigma^2$.

(b) If $\sum a_i = 0$, show that the unconditional variance of $\sum a_i \bar{Y}_{i\cdot}$ is $\text{Var}\,(\sum a_i \bar{Y}_{i\cdot}) = \frac{1}{r}(\sigma^2 + \sigma_B^2)(1-\rho)\sum a_i^2$, where $\rho =$ intraclass correlation.

11.24 The form of the Stein estimator of Miscellanea 11.5.6 can be justified somewhat by an *empirical Bayes* argument given in Efron and Morris (1972), which can be quite useful in data analysis. Such an argument may have been known by Stein (1956), although he makes no mention of it. Let $X_i \sim \text{n}(\theta_i, 1), i = 1, \ldots, p$, and θ_i be iid $\text{n}(0, \tau^2)$.

(a) Show that the X_is, marginally, are iid $\text{n}(0, \tau^2 + 1)$, and, hence, $\sum X_i^2/(\tau^2+1) \sim \chi_p^2$.

(b) Using the marginal distribution, show that $\text{E}(1-((p-2)/\sum_{j=1}^p X_j^2)) = \tau^2/(\tau^2+1)$ if $p \geq 3$. Thus, the Stein estimator of Miscellanea 11.5.6 is an empirical Bayes version of the Bayes estimator $\delta_i^\pi(\mathbf{X}) = [\tau^2/(\tau^2+1)]X_i$.

(c) Show that the argument fails if $p < 3$ by showing that $\text{E}(1/Y) = \infty$ if $Y \sim \chi_p^2$ with $p < 3$.

11.25 In Section 11.3.1, we found the least squares estimators of α and β by a two-stage minimization. This minimization can also be done using partial derivatives.

(a) Compute $\frac{\partial \text{RSS}}{\partial c}$ and $\frac{\partial \text{RSS}}{\partial d}$ and set them equal to 0. Show that the resulting two equations can be written as

$$nc + \left(\sum_{i=1}^n x_i\right) d = \sum_{i=1}^n y_i \quad \text{and} \quad \left(\sum_{i=1}^n x_i\right) c + \left(\sum_{i=1}^n x_i^2\right) d = \sum_{i=1}^n x_i y_i.$$

(These equations are called the *normal equations* for this minimization problem.)

(b) Show that $c = a$ and $d = b$ are the solutions to the normal equations.

(c) Check the second partial derivative condition to verify that the point $c = a$ and $d = b$ is indeed the minimum of RSS.

11.26 Suppose n is an even number. The values of the predictor variable, $x_1, \ldots, x_n$, all must be chosen to be in the interval $[e, f]$. Show that the choice that maximizes S_{xx} is for half of the x_i equal to e and the other half equal to f. (This was the choice mentioned in Section 11.3.2 that minimizes $\text{Var}\, b$.)

11.27 Observations (x_i, Y_i), $i = 1, \ldots, n$, follow the model $Y_i = \alpha + \beta x_i + \epsilon_i$, where $\text{E}\,\epsilon_i = 0$, $\text{Var}\,\epsilon_i = \sigma^2$, and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$. Find the best linear unbiased estimator of α.

11.28 Show that in the conditional normal model for simple linear regression, the MLE of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

11.29 Consider the residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ defined in Section 11.3.4 by $\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$.

(a) Show that $\text{E}\hat{\epsilon}_i = 0$.

(b) Verify that

$$\text{Var}\,\hat{\epsilon}_i = \text{Var}\, Y_i + \text{Var}\,\hat{\alpha} + x_i^2 \text{Var}\,\hat{\beta} - 2\text{Cov}(Y_i, \hat{\alpha}) - 2x_i \text{Cov}(Y_i, \hat{\beta}) + 2x_i \text{Cov}(\hat{\alpha}, \hat{\beta}).$$

(c) Use Lemma 11.3.2 to show that

$$\text{Cov}(Y_i, \hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right) \quad \text{and} \quad \text{Cov}(Y_i, \hat{\beta}) = \sigma^2 \frac{x_i - \bar{x}}{S_{xx}},$$

and use these to verify (11.3.28).

11.30 Fill in the details about the distribution of $\hat{\alpha}$ left out of the proof of Theorem 11.3.3.

(a) Show that the estimator $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ can be expressed as $\hat{\alpha} = \sum_{i=1}^n c_i Y_i$, where

$$c_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}.$$

(b) Verify that

$$\text{E}\hat{\alpha} = \alpha \quad \text{and} \quad \text{Var}\,\hat{\alpha} = \sigma^2 \left[\frac{1}{nS_{xx}} \sum_{i=1}^n x_i^2 \right].$$

(c) Verify that

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}}{S_{xx}}.$$

11.31 Verify the claim in Theorem 11.3.3, that $\hat{\epsilon}_i$ is uncorrelated with $\hat{\alpha}$ and $\hat{\beta}$. (Show that $\hat{\epsilon}_i = \sum e_j Y_j$, where the e_js are given by (11.3.30). Then, using the facts that we can write $\hat{\alpha} = \sum c_j Y_j$ and $\hat{\beta} = \sum d_j Y_j$, verify that $\sum e_j c_j = \sum e_j d_j = 0$ and apply Lemma 11.3.2.)

11.32 Observations $(x_i, Y_i), i = 1, \ldots, n$, are made according to the model

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

where $x_1, \ldots, x_n$ are fixed constants and $\epsilon_1, \ldots, \epsilon_n$ are iid $\text{n}(0, \sigma^2)$. The model is then reparameterized as

$$Y_i = \alpha' + \beta'(x_i - \bar{x}) + \epsilon_i.$$

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the MLEs of α and β, respectively, and $\hat{\alpha}'$ and $\hat{\beta}'$ denote the MLEs of α' and β', respectively.

(a) Show that $\hat{\beta}' = \hat{\beta}$.

(b) Show that $\hat{\alpha}' \neq \hat{\alpha}$. In fact, show that $\hat{\alpha}' = \bar{Y}$. Find the distribution of $\hat{\alpha}'$.

(c) Show that $\hat{\alpha}'$ and $\hat{\beta}'$ are uncorrelated and, hence, independent under normality.

11.33 Observations $(X_i, Y_i), i = 1, \ldots, n$, are made from a bivariate normal population with parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, and the model $Y_i = \alpha + \beta x_i + \epsilon_i$ is going to be fit.

(a) Argue that the hypothesis $H_0 : \beta = 0$ is true if and only if the hypothesis $H_0 : \rho = 0$ is true. (See (11.3.25).)

(b) Show algebraically that

$$\frac{\hat{\beta}}{S/\sqrt{S_{xx}}} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}},$$

where r is the sample correlation coefficient, the MLE of ρ.

(c) Show how to test $H_0: \rho = 0$, given only r^2 and n, using Student's t with $n-2$ degrees of freedom (see (11.3.33)). (Fisher derived an approximate confidence interval for ρ, using a variance-stabilizing transformation. See Exercise 11.4.)

11.34 (a) Illustrate the partitioning of the sum of squares for simple linear regression by calculating the regression ANOVA table for the following data. Parents are often interested in predicting the eventual heights of their children. The following is a portion of the data taken from a study that might have been suggested by Galton's analysis.

Height (inches) at age 2 (x)	39	30	32	34	35	36	36	30
Height (inches) as an adult (y)	71	63	63	67	68	68	70	64

(b) Analytically establish the partitioning of the sum of squares for simple linear regression by verifying (11.3.36).

(c) Prove that the two expressions for the regression sum of squares are, in fact, equal; that is, show that

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}.$$

(d) Show that the *coefficient of determination*, r^2, given by

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

can be derived as the square of the sample correlation coefficient either of the n pairs $(y_1, x_1), \ldots, (y_n, x_n)$ or of the n pairs $(y_1, \hat{y}_1), \ldots, (y_n, \hat{y}_n)$.

11.35 Observations $Y_1, \ldots, Y_n$ are described by the relationship $Y_i = \theta x_i^2 + \epsilon_i$, where $x_1, \ldots, x_n$ are fixed constants and $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathrm{n}(0, \sigma^2)$.

(a) Find the least squares estimator of θ.

(b) Find the MLE of θ.

(c) Find the best unbiased estimator of θ.

11.36 Observations $Y_1, \ldots, Y_n$ are made according to the model $Y_i = \alpha + \beta x_i + \epsilon_i$, where $x_1, \ldots, x_n$ are fixed constants and $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathrm{n}(0, \sigma^2)$. Let $\hat{\alpha}$ and $\hat{\beta}$ denote MLEs of α and β.

(a) Assume that $x_1, \ldots, x_n$ are observed values of iid random variables $X_1, \ldots, X_n$ with distribution $\mathrm{n}(\mu_X, \sigma_X^2)$. Prove that when we take expectations over the joint distribution of X and Y, we still get $\mathrm{E}\hat{\alpha} = \alpha$ and $\mathrm{E}\hat{\beta} = \beta$.

(b) The phenomenon of part (a) does not carry over to the covariance. Calculate the unconditional covariance of $\hat{\alpha}$ and $\hat{\beta}$ (using the joint distribution of X and Y).

11.37 We observe random variables $Y_1, \ldots, Y_n$ that are mutually independent, each with a normal distribution with variance σ^2. Furthermore, $\mathrm{E}Y_i = \beta x_i$, where β is an unknown parameter and $x_1, \ldots, x_n$ are fixed constants not all equal to 0.

(a) Find the MLE of β. Compute its mean and variance.

(b) Compute the Cramér–Rao Lower Bound for the variance of an unbiased estimator of β.

(c) Find a best unbiased estimator of β.

(d) If you could place the values $x_1, \ldots, x_n$ anywhere within a given nondegenerate closed interval $[A, B]$, where would you place these values? Justify your answer.

(e) For a given positive value r, the *maximum probability estimator of β with respect to r* is the value of D that maximizes the integral

$$\int_{D-r}^{D+r} f(y_1, \ldots, y_n|\beta)d\beta,$$

where $f(y_1, \ldots, y_n|\beta)$ is the joint pdf of $Y_1, \ldots, Y_n$. Find this estimator.

11.38 An ecologist takes data (x_i, Y_i), $i = 1, \ldots, n$, where x_i is the size of an area and Y_i is the number of moss plants in the area. We model the data by $Y_i \sim \text{Poisson}(\theta x_i)$, Y_is independent.

(a) Show that the least squares estimator of θ is $\sum x_i Y_i / \sum x_i^2$. Show that this estimator has variance $\theta \sum x_i^3 / (\sum x_i^2)^2$. Also, compute its bias.

(b) Show that the MLE of θ is $\sum Y_i / \sum x_i$ and has variance $\theta / \sum x_i$. Compute its bias.

(c) Find a best unbiased estimator of θ and show that its variance attains the Cramér–Rao Lower Bound.

11.39 Verify that the simultaneous confidence intervals in (11.3.42) have the claimed coverage probability.

11.40 (a) Prove that if a, b, c, and d are constants, with $c > 0$ and $d > 0$, then

$$\max_t \frac{(a+bt)^2}{c+dt^2} = \frac{a^2}{c} + \frac{b^2}{d}.$$

(b) Use part (a) to verify equation (11.3.44) and hence fill in the gap in Theorem 11.3.6.

(c) Use part (a) to find a Scheffé-type simultaneous band using the prediction intervals of (11.3.41). That is, rewriting the prediction intervals as was done in Theorem 11.3.6, show that

$$\max_t \frac{\left((\bar{Y} - \mu_{\bar{Y}}) + (\hat{\beta} - \beta)t\right)^2}{S^2\left[1 + \frac{1}{n} + \frac{t^2}{S_{xx}}\right]} = \frac{\frac{n}{n+1}(\bar{Y} - \mu_{\bar{Y}})^2 + S_{xx}(\hat{\beta} - \beta)^2}{S^2}.$$

(d) The distribution of the maximum is not easy to write down, but we could approximate it. Approximate the statistic by using moment matching, as done in Example 7.2.3.

11.41 In the discussion in Example 12.4.2, note that there was one observation from the potoroo data that had a missing value. Suppose that on the 24th animal it was observed that $O_2 = 16.3$.

(a) Write down the observed data and expected complete data log likelihood functions.

(b) Describe the E step and the M step of an EM algorithm to find the MLEs.

(c) Find the MLEs using all 24 observations.

(d) Actually, the O_2 reading on the 24th animal was not observed, but rather the CO_2 was observed to be 4.2 (and the O_2 was missing). Set up the EM algorithm in this case and find the MLEs. (This is a much harder problem, as you now have to take expectations over the xs. This means you have to formulate the regression problem using the bivariate normal distribution.)

11.5 Miscellanea

11.5.1 Cochran's Theorem

Sums of squares of normal random variables, when properly scaled and centered, are distributed as chi squared random variables. This type of result is first due to Cochran (1934). Cochran's Theorem gives necessary and sufficient conditions on the scaling required for squared and summed iid normal random variables to be distributed as a chi squared random variable. The conditions are not difficult, but they are best stated in terms of properties of matrices and will not be treated here. It is an immediate consequence of Cochran's Theorem that in the oneway ANOVA, the χ^2 random variables partition as discussed in Section 11.2.6. Furthermore, another consequence is that in the Randomized Complete Blocks ANOVA (see Miscellanea 11.5.3), the mean squares all have chi squared distributions.

Cochran's Theorem has been generalized to the extent that necessary and sufficient conditions are known for the distribution of squared normals (not necessarily iid) to be chi squared. See Stuart and Ord (1987, Chapter 15) for details.

11.5.2 Multiple Comparisons

We have seen two ways of doing simultaneous inference in this chapter: the Scheffé procedure and use of the Bonferroni Inequality. There is a plethora of other simultaneous inference procedures. Most are concerned with inference on pairwise comparisons, that is, differences between means. These procedures can be applied to estimate treatment means in the oneway ANOVA.

A method due to Tukey (see Miller 1981), sometimes known as the Q method, applies a Scheffé-type maximization argument but over only pairwise differences, not all contrasts. The Q distribution is the distribution of

$$Q = \max_{i,j} \left| \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) - (\theta_i - \theta_j)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{n}\right)}} \right|,$$

where $n_i = n$ for all i. (Hayter 1984 has shown that if $n_i \neq n_j$ and the n above is replaced by the harmonic mean n_h, where $1/n_h = \frac{1}{2}((1/n_i) + (1/n_j))$, the resulting procedure is conservative.) The Q method is an improvement over Scheffé's S method in that if there is interest only in pairwise differences, the Q method is more powerful (shorter intervals). This is easy to see because, by definition, the Q maximization will produce a smaller maximum than the S method.

Other types of multiple comparison procedures that deal with pairwise differences are more powerful than the S method. Some procedures are the LSD (Least Significant Difference) Procedure, Protected LSD, Duncan's Procedure, and Student–Neumann–Keuls' Procedure. These last two are *multiple range* procedures. The cutoff point to which comparisons are made changes between comparisons.

One difficulty in fully understanding multiple comparison procedures is that the definition of Type I Error is not inviolate. Some of these procedures have changed the definition of Type I Error for multiple comparisons, so exactly what is meant

by "α level" is not always clear. Some of the types of error rates considered are called *experimentwise error rate*, *comparisonwise error rate*, and *familywise error rate*. Miller (1981) and Hsu (1996) are good references for this topic. A humorous but illuminating treatment of this subject is given in Carmer and Walker (1982).

11.5.3 Randomized Complete Block Designs

Section 11.2 was concerned with a *oneway* classification of the data; that is, there was only one categorization (treatment) in the experiment. In general, the ANOVA allows for many types of categorization, with one of the most commonly used ANOVAs being the Randomized Complete Block (RCB) ANOVA.

A *block* (or *blocking factor*) is categorization that is in an experiment for the express purpose of removing variation. In contrast to a treatment, there is usually no interest in finding block differences. The practice of blocking originated in agriculture, where experimenters took advantage of similar growing conditions to control experimental variances. To model this, the actual blocks in the experiment were considered to be a random sample from a large population of blocks (which makes them a *random* factor).

RCB ANOVA assumptions

Random variables Y_{ij} are observed according to the model

$$Y_{ij}|\mathbf{b} = \mu + \tau_i + b_j + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, r,$$

where:

(i) The random variables $\epsilon_{ij} \sim$ iid $\mathrm{n}(0, \sigma^2)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, r$ (normal errors with equal variances).

(ii) The random variables $B_1, \ldots, B_r$, whose realized (but unobserved) values are the blocks $b_1, \ldots, b_r$, are iid $\mathrm{n}(0, \sigma_B^2)$ and are independent of ϵ_{ij} for all i, j.

The mean and variance of Y_{ij} are

$$\mathrm{E}\, Y_{ij} = \mu + \tau_i \quad \text{and} \quad \mathrm{Var}\, Y_{ij} = \sigma_B^2 + \sigma^2.$$

Moreover, although the Y_{ij}s are uncorrelated conditionally, there is correlation in the blocks unconditionally. The correlation between Y_{ij} and $Y_{i'j}$ in block j, with $i \neq i'$, is

$$\frac{\mathrm{Cov}(Y_{ij}, Y_{i'j})}{\sqrt{(\mathrm{Var}\, Y_{ij})(\mathrm{Var}\, Y_{i'j})}} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma^2},$$

a quantity called the *intraclass correlation*. Thus, the model implies not only that there is correlation in the blocks but also that there is positive correlation. This is a consequence of the additive model and the assumption that the ϵs and Bs are independent (see Exercise 11.23). Even though the Y_{ij}s are not independent, the intraclass correlation structure still results in an analysis of variance where ratios of mean squares have the F distribution (see Miscellanea 11.5.1).

11.5.4 Other Types of Analyses of Variance

The two types of ANOVAs that we have considered, oneway ANOVAs and RCB ANOVAs, are the simplest types. For example, an extension of a complete block design is an *incomplete* block design. Sometimes there are physical constraints that prohibit putting all treatments in each block and an incomplete block design is needed. Deciding how to arrange the treatments in such a design is both difficult and critical. Of course, as the design gets more complicated, so does the analysis.

Study of the subject of statistical design, which is concerned with getting the most information from the fewest observations, leads to more complicated and more efficient ANOVAs in many situations. ANOVAs based on designs such as *fractional factorials*, *Latin squares*, and *balanced incomplete blocks* can be efficient methods of gathering much information about a phenomenon. Good overall references for this subject are Cochran and Cox (1957), Dean and Voss (1999), and Kuehl (2000).

11.5.5 Shapes of Confidence Bands

Confidence bands come in many shapes, not just the *hyperbolic* shape defined by the Scheffé band. For example, Gafarian (1964) showed how to construct a *straight-line* band over a finite interval. Gafarian-type bands allow statements of the form

$$P(\hat{\alpha} + \hat{\beta}x - d_\alpha \le \alpha + \beta x \le \hat{\alpha} + \hat{\beta}x + d_\alpha \text{ for all } x \in [a, b]) = 1 - \alpha.$$

Gafarian gave tables of d_α. A finite-width band must, necessarily, apply only to a finite range of x. Any band of level $1 - \alpha$ must have infinite length as $|x| \to \infty$.

Casella and Strawderman (1980), among others, showed how to construct Scheffé-type bands over finite intervals, thereby reducing width while maintaining the same confidence as the infinite Scheffé band. Naiman (1983) compared performance of straight-line and Scheffé bands over finite intervals. Under his criterion, one of average width, the Scheffé band is superior. In some cases, an experimenter might be more comfortable with the interpretation of a straight-line band, however.

Shapes other than straight-line and hyperbolic are possible. Piegorsch (1985) investigated and characterized the shapes that are admissible in the sense that their probability statements cannot be improved upon. He obtained "growth conditions" that must be satisfied by an admissible band. Naiman (1983, 1984, 1987) and Naiman and Wynn (1992, 1997) have developed this theory to a very high level, establishing useful inequalities and geometric identities to further improve inferences.

11.5.6 Stein's Paradox

One part of the analysis of variance is concerned with the simultaneous estimation of a collection of normal means. Developments in this particular problem, starting with Stein (1956), have had a profound effect on both the theory and applications of point estimation.

A canonical version of the analysis of variance is to observe $\mathbf{X} = (X_1, \ldots, X_p)$, independent normal random variables with $X_i \sim \mathrm{n}(\theta_i, 1)$, $i = 1, \ldots, p$, with the objective being the estimation of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Our usual estimate of θ_i would

be X_i, but Stein (1956) established the surprising result that, if $p \geq 3$, the estimator of θ_i given by

$$\delta_i^S(\mathbf{X}) = \left(1 - \frac{p-2}{\sum_{i=1}^p X_i^2}\right) X_i$$

is a better estimator of θ_i in the sense that

$$\sum_{i=1}^p \mathrm{E}_\theta \left(X_i - \theta_i\right)^2 \geq \sum_{i=1}^p \mathrm{E}_\theta \left(\delta_i^S(\mathbf{X}) - \theta_i\right)^2.$$

That is, the summed mean squared of Stein's estimator is always smaller, and usually strictly smaller, than that of $\mathbf{X}$.

Notice that the estimators are being compared using the sum of the component-wise mean squared errors, and each δ_i^S can be a function of the entire vector $(X_1, \ldots, X_p)$. Thus, all of the data can be used in estimating each mean. Since the X_is are independent, we might think that restricting δ_i^S to be just a function of X_i would be enough. However, by summing the mean squared errors, we tie the components together.

In the oneway ANOVA we observe

$$\bar{Y}_{i\cdot} \sim \mathrm{n}\left(\theta_i, \frac{\sigma^2}{n_i}\right), \quad i = 1, \ldots, k, \quad \text{independent},$$

where the $\bar{Y}_{i\cdot}$s are the cell means. The Stein estimator takes the form

$$\delta_i^S(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot}) = \left(1 - \frac{(k-2)\sigma^2}{\sum n_j \bar{Y}_{j\cdot}^2}\right)^+ \bar{Y}_{i\cdot}, \quad i = 1, \ldots, k.$$

This Stein-type estimator can further be improved by choosing a meaningful place toward which to shrink (the above estimator shrinks toward 0). One such estimator, due to Lindley (1962), shrinks toward the grand mean of the observations. It is given by

$$\delta_i^{\mathrm{L}}(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot}) = \bar{\bar{Y}} + \left(1 - \frac{(k-3)\sigma^2}{\sum n_j (\bar{Y}_{j\cdot} - \bar{\bar{Y}})^2}\right)^+ (\bar{Y}_{i\cdot} - \bar{\bar{Y}}), \quad i = 1, \ldots, k.$$

Other choices of a shrinkage target might be even more appropriate. Discussion of this, including methods for improving on confidence statements, such as the Scheffé S method, is given in Casella and Hwang (1987). Morris (1983) also discusses applications of these types of estimators.

There have been many theoretical developments using Stein-type estimators, not only in point estimation but also in confidence set estimation, where it has been shown that recentering at a Stein estimator can result in increased coverage probability and reduced size. There is also a strong connection between Stein estimators and empirical Bayes estimators (see Miscellanea 7.5.6), first uncovered in a series

of papers by Efron and Morris (1972, 1973, 1975), where the components of $\boldsymbol{\theta}$ are tied together using a common prior distribution. An introduction to the theory and some applications of Stein estimators is given in Lehmann and Casella (1998, Chapter 5).

Chapter 12

Regression Models

"So startling would his results appear to the uninitiated that until they learned the processes by which he had arrived at them they might well consider him as a necromancer."

Dr. Watson, speaking about Sherlock Holmes
A Study in Scarlet

12.1 Introduction

Chapter 11 was concerned with what could be called "classic linear models." Both the ANOVA and simple linear regression are based on an underlying linear model with normal errors. In this chapter we look at some extensions of this model that have proven to be useful in practical problems.

In Section 12.2, the linear model is extended to models with errors in the predictor, which is called regression with errors in variables (EIV). In this model the predictor variable X now becomes a random variable like the response variable Y. Estimation in this model encounters many unforeseen difficulties and can be very different from the simple linear regression model.

The linear model is further generalized in Section 12.3, where we look at logistic regression. Here, the response variable is discrete, a Bernoulli variable. The Bernoulli mean is a bounded function, and a linear model on a bounded function can run into problems (especially at the boundaries). Because of this we transform the mean to an unbounded parameter (using the logit transformation) and model the transformed parameter as a linear function of the predictor variable. When a linear model is put on a function of a response mean, it becomes a *generalized linear model.*

Lastly, in Section 12.4, we look at robustness in the setting of linear regression. In contrast to the other sections in this chapter where we change the model, now we change the fitting criterion. The development parallels that of Section 10.2.2, where we looked at robust point estimates. That is, we replace the least squares criterion with one based on a ρ-function that results in estimates that are less sensitive to underlying observations (but retain some efficiency).

12.2 Regression with Errors in Variables

Regression with *errors in variables* (EIV), also known as the *measurement error model*, is so fundamentally different from the simple linear regression of Section 11.3

that it is probably best thought of as a completely different topic. It is presented as a generalization of the usual regression model mainly for traditional reasons. However, the problems that arise with this model are very different.

The models of this section are generalizations of simple linear regression in that we will work with models of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i, \tag{12.2.1}$$

but now we do not assume that the xs are known. Instead, we can measure a random variable whose mean is x_i. (In keeping with our notational conventions, we will speak of measuring a random variable X_i whose mean is not x_i but ξ_i.)

The intention here is to illustrate different approaches to the EIV model, showing some of the standard solutions and the (sometimes) unexpected difficulties that arise. For a more thorough introduction to this problem, there are the review article by Gleser (1991); books by Fuller (1987) and Carroll, Ruppert, and Stefanski (1995); and a volume edited by Brown and Fuller (1991). Kendall and Stuart (1979, Chapter 29) also treat this topic in some detail.

In the general EIV model we assume that we observe pairs (x_i, y_i) sampled from random variables (X_i, Y_i) whose means satisfy the linear relationship

$$\mathrm{E}Y_i = \alpha + \beta(\mathrm{E}X_i). \tag{12.2.2}$$

If we define

$$\mathrm{E}Y_i = \eta_i \quad \text{and} \quad \mathrm{E}X_i = \xi_i,$$

then the relationship (12.2.2) becomes

$$\eta_i = \alpha + \beta\xi_i, \tag{12.2.3}$$

a linear relationship between the means of the random variables.

The variables ξ_i and η_i are sometimes called *latent variables*, a term that refers to quantities that cannot be directly measured. Latent variables may be not only impossible to measure directly but impossible to measure *at all*. For example, the IQ of a person is impossible to measure. We can measure a score on an IQ test, but we can never measure the variable IQ. Relationships between IQ and other variables, however, are often hypothesized.

The model specified in (12.2.2) really makes no distinction between X and Y. If we are interested in a regression, however, there should be a reason for choosing Y as the response and X as the predictor. Keeping this specification in mind, of regressing Y on X, we define the *errors in variables model* or *measurement error model* as this. Observe independent pairs (X_i, Y_i), $i = 1, \ldots, n$, according to

$$\begin{aligned} Y_i &= \alpha + \beta\xi_i + \epsilon_i, \quad & \epsilon_i &\sim \mathrm{n}(0, \sigma_\epsilon^2), \\ X_i &= \xi_i + \delta_i, & \delta_i &\sim \mathrm{n}(0, \sigma_\delta^2). \end{aligned} \tag{12.2.4}$$

Note that the assumption of normality, although common, is not necessary. Other distributions can be used. In fact, some of the problems encountered with this model are caused by the normality assumption. (See, for example, Solari 1969.)

Example 12.2.1 (Estimating atmospheric pressure) The EIV regression model arises fairly naturally in situations where the x variable is observed along with the y variable (rather than being controlled). For example, in the 1800s the Scottish physicist J. D. Forbes tried to use measurements on the boiling temperature of water to estimate altitude above sea level. To do this, he simultaneously measured boiling temperature and atmospheric pressure (from which altitude can be obtained). Since barometers were quite fragile in the 1800s, it would be useful to estimate pressure, or more precisely log(pressure), from temperature. The data observed at nine locales are

Boiling point (°F)	log(pressure) (log(Hg))
194.5	1.3179
197.9	1.3502
199.4	1.3646
200.9	1.3782
201.4	1.3806
203.6	1.4004
209.5	1.4547
210.7	1.4630
212.2	1.4780

and an EIV model is reasonable for this situation. ||

A number of special cases of the model (12.2.4) have already been seen. If $\delta_i = 0$, then the model becomes simple linear regression (since there is now no measurement error, we can directly observe the ξ_is). If $\alpha = 0$, then we have

$$Y_i \sim \mathrm{n}(\eta_i, \sigma_\epsilon^2), \quad i = 1, \ldots, n,$$
$$X_i \sim \mathrm{n}(\xi_i, \sigma_\delta^2), \quad i = 1, \ldots, n,$$

where, possibly, $\sigma_\delta^2 \neq \sigma_\epsilon^2$, a version of the Behrens–Fisher problem.

12.2.1 Functional and Structural Relationships

There are two different types of relationship that can be specified in the EIV model: one that specifies a *functional* linear relationship and one describing a *structural* linear relationship. The different relationship specifications can lead to different estimators with different properties. As said by Moran (1971), "This is not very happy terminology, but we will stick to it because the distinction is essential. . . ." Some interpretations of this terminology are given in the Miscellanea section. For now we merely present the two models.

Linear functional relationship model

This is the model as presented in (12.2.4) where we have random variables X_i and Y_i, with $\mathrm{E}X_i = \xi_i$, $\mathrm{E}Y_i = \eta_i$, and we assume the *functional relationship*

$$\eta_i = \alpha + \beta\xi_i.$$

We observe pairs $(X_i, Y_i), i = 1, \ldots, n$, according to

$$\begin{aligned} Y_i &= \alpha + \beta\xi_i + \epsilon_i, \quad & \epsilon_i &\sim \mathrm{n}(0, \sigma_\epsilon^2), \\ X_i &= \xi_i + \delta_i, & \delta_i &\sim \mathrm{n}(0, \sigma_\delta^2), \end{aligned} \tag{12.2.5}$$

where the ξ_is are fixed, unknown parameters and the ϵ_is and δ_is are independent. The parameters of main interest are α and β, and inference on these parameters is made using the joint distribution of $((X_1, Y_1), \ldots, (X_n, Y_n))$, *conditional on* $\xi_1, \ldots, \xi_n$.

Linear structural relationship model

This model can be thought of as an extension of the functional relationship model, extended through the following hierarchy. As in the functional relationship model, we have random variables X_i and Y_i, with $\mathrm{E}X_i = \xi_i$, $\mathrm{E}Y_i = \eta_i$, and we assume the *functional relationship* $\eta_i = \alpha + \beta\xi_i$. But now we assume that the parameters $\xi_1, \ldots, \xi_n$ are themselves a random sample from a common population. Thus, conditional on $\xi_1, \ldots, \xi_n$, we observe pairs $(X_i, Y_i), i = 1, \ldots, n$, according to

$$\begin{aligned} Y_i &= \alpha + \beta\xi_i + \epsilon_i, \quad & \epsilon_i &\sim \mathrm{n}(0, \sigma_\epsilon^2), \\ X_i &= \xi_i + \delta_i, & \delta_i &\sim \mathrm{n}(0, \sigma_\delta^2), \end{aligned} \tag{12.2.6}$$

and also

$$\xi_i \sim \text{iid } \mathrm{n}(\xi, \sigma_\xi^2).$$

As before, the ϵ_is and δ_is are independent and they are also independent of the ξ_is. As in the functional relationship model, the parameters of main interest are α and β. Here, however, the inference on these parameters is made using the joint distribution of $((X_1, Y_1), \ldots, (X_n, Y_n))$, *unconditional on* $\xi_1, \ldots, \xi_n$. (That is, $\xi_1, \ldots, \xi_n$ are integrated out according to the distribution in (12.2.6).)

The two models are quite similar in that statistical properties of estimators in one model (for example, consistency) often carry over into the other model. More precisely, estimators that are consistent in the functional model are also consistent in the structural model (Nussbaum 1976 or Gleser 1983). This makes sense, as the functional model is a "conditional version" of the structural model. Estimators that are consistent in the functional model must be so for all values of the ξ_is so are necessarily consistent in the structural model, which averages over the ξ_is. The converse implication is false. However, there is a useful implication that goes from the structural to the functional relationship model. If a parameter is not *identifiable* in the structural model, it is also not identifiable in the functional model. (See Definition 11.2.2.)

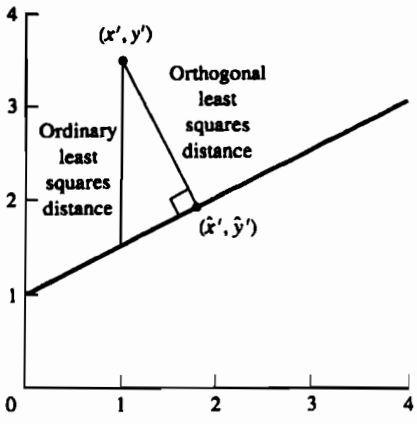

Figure 12.2.1. *Distance minimized by orthogonal least squares*

As we shall see, the models share similar problems and, in certain situations, similar likelihood solutions. It is probably easier to do statistical theory in the structural model, while the functional model often seems to be the more reasonable model for many situations. Thus, the underlying similarities come in handy.

As already mentioned, one of the major differences in the models is in the inferences about α and β, the parameters that describe the regression relationship. This difference is of utmost importance and cannot be stressed too often. In the functional relationship model, this inference is made conditional on $\xi_1, \ldots, \xi_n$, using the joint distribution of X and Y conditional on $\xi_1, \ldots, \xi_n$. On the other hand, in the structural relationship model, this inference is made unconditional on $\xi_1, \ldots, \xi_n$, using the marginal distribution of X and Y with $\xi_1, \ldots, \xi_n$ integrated out.

12.2.2 A Least Squares Solution

As in Section 11.3.1, we forget statistics for a while and try to find the "best" line through the observed points (x_i, y_i), $i = 1, \ldots, n$. Previously, when it was assumed that the xs were measured without error, it made sense to consider minimization of vertical distances. This distance measure implicitly assumes that the x value is correct and results in *ordinary* least squares. Here, however, there is no reason to consider vertical distances, since the xs now have error associated with them. In fact, statistically speaking, ordinary least squares has some problems in EIV models (see the Miscellanea section).

One way to take account of the fact that the xs also have error in their measurement is to perform *orthogonal least squares*, that is, to find the line that minimizes orthogonal (perpendicular to the line) distances rather than vertical distances (see Figure 12.2.1). This distance measure does not favor the x variable, as does ordinary least squares, but rather treats both variables equitably. It is also known as the method of *total least squares*. From Figure 12.2.1, for a particular data point (x', y'), the point on a line $y = a + bx$ that is closest when we measure distance orthogonally is given

by (see Exercise 12.1)

$$\text{(12.2.7)} \qquad \hat{x}' = \frac{by' + x' - ab}{1 + b^2}, \qquad \text{and} \qquad \hat{y}' = a + \frac{b}{1 + b^2}(by' + x' - ab).$$

Now assume that we have data $(x_i, y_i), i = 1, \ldots, n$. The squared distance between an observed point (x_i, y_i) and the closest point on the line $y = a + bx$ is $(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2$, where $\hat{x}_i$ and $\hat{y}_i$ are defined by (12.2.7). The *total least squares problem* is to minimize, over all a and b, the quantity

$$\sum_{i=1}^{n} \left((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right).$$

It is straightforward to establish that we have

$$\begin{aligned} &\sum_{i=1}^{n} \left((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right) \\ &\qquad = \sum_{i=1}^{n} \left(\frac{b^2}{(1 + b^2)^2} [y_i - (a + bx_i)]^2 + \frac{1}{(1 + b^2)^2} [y_i - (a + bx_i)]^2 \right) \\ &\qquad = \frac{1}{1 + b^2} \sum_{i=1}^{n} (y_i - (a + bx_i))^2 . \end{aligned} \qquad \text{(12.2.8)}$$

For fixed b, the term in front of the sum is a constant. Thus, the minimizing choice of a in the sum is $a = \bar{y} - b\bar{x}$, just as in (11.3.9). If we substitute back into (12.2.8), the total least squares solution is the one that minimizes, over all b,

$$\text{(12.2.9)} \qquad \frac{1}{1 + b^2} \sum_{i=1}^{n} ((y_i - \bar{y}) - b(x_i - \bar{x}))^2 .$$

As in (11.3.6) and (11.3.6), we define the sums of squares and cross-products by

$$\text{(12.2.10)} \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Expanding the square and summing show that (12.2.9) becomes

$$\frac{1}{1 + b^2} [S_{yy} - 2bS_{xy} + b^2 S_{xx}].$$

Standard calculus methods will give the minimum (see Exercise 12.2), and we find the orthogonal least squares line given by $y = a + bx$, with

$$\text{(12.2.11)} \qquad a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{-(S_{xx} - S_{yy}) + \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

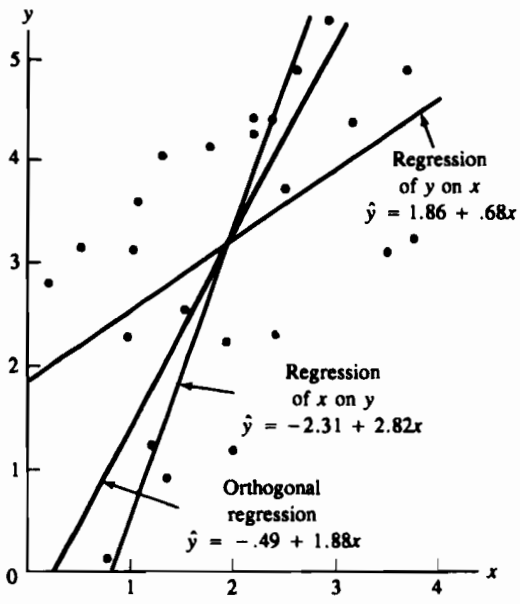

Figure 12.2.2. *Three regression lines for data in Table 11.3.1*

As might be expected, this line is different from the least squares line. In fact, as we shall see, this line always lies between the ordinary regression of y on x and the ordinary regression of x on y. This is illustrated in Figure 12.2.2, where the data in Table 11.3.1 were used to calculate the orthogonal least squares line $\hat{y} = -.49 + 1.88x$.

In simple linear regression we saw that, under normality, the ordinary least squares solutions for α and β were the same as the MLEs. Here, the orthogonal least squares solution is the MLE only in a special case, when we make certain assumptions about the parameters.

The difficulties to be encountered with likelihood estimation once again illustrate the differences between a mathematical solution and a statistical solution. We obtained a mathematical least squares solution to the line fitting problem without much difficulty. This will not happen with the likelihood solution.

12.2.3 Maximum Likelihood Estimation

We first consider the maximum likelihood solution of the functional linear relationship model, the situation for the structural relationship model being similar and, in some respects, easier. With the normality assumption, the functional relationship model can be expressed as

$$Y_i \sim \mathrm{n}(\alpha + \beta\xi_i, \sigma_\epsilon^2) \quad \text{and} \quad X_i \sim \mathrm{n}(\xi_i, \sigma_\delta^2), \quad i = 1, \ldots, n,$$

where the X_is and Y_is are independent. Given observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$, the likelihood function is

(12.2.12)

$$L\left(\alpha, \beta, \xi_1, \ldots, \xi_n, \sigma_\delta^2, \sigma_\epsilon^2 | (\mathbf{x}, \mathbf{y})\right) = \frac{1}{(2\pi)^n} \frac{1}{(\sigma_\delta^2 \sigma_\epsilon^2)^{n/2}} \exp\left[-\sum_{i=1}^n \frac{(x_i - \xi_i)^2}{2\sigma_\delta^2}\right] \exp\left[-\sum_{i=1}^n \frac{(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\epsilon^2}\right].$$

The problem with this likelihood function is that it does not have a finite maximum. To see this, take the parameter configuration $\xi_i = x_i$ and then let $\sigma_\delta^2 \to 0$. The value of the function goes to infinity, showing that there is no maximum likelihood solution. In fact, Solari (1969) has shown that if the equations defining the first derivative of L are set equal to 0 and solved, the result is a saddle point, not a maximum. Notice that as long as we have total control over the parameters, we can always force the likelihood function to infinity. In particular, we can always take a variance to 0, while keeping the exponential term bounded.

We will make the common assumption, which not only is reasonable but also alleviates many problems, that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, where $\lambda > 0$ is fixed and known. (See Kendall and Stuart 1979, Chapter 29, for a discussion of other assumptions on the variances.) This assumption is one of the least restrictive, saying that we know only the ratio of the variances, not the individual values. Moreover, the resulting model is relatively well behaved.

Under this assumption, we can write the likelihood function as

(12.2.13)

$$L\left(\alpha, \beta, \xi_1, \ldots, \xi_n, \sigma_\delta^2 | (\mathbf{x}, \mathbf{y})\right) = \frac{1}{(2\pi)^n} \frac{\lambda^{n/2}}{(\sigma_\delta^2)^n} \exp\left[-\sum_{i=1}^n \frac{(x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\delta^2}\right],$$

which we can now maximize. We will perform the maximization in stages, making sure that, at each step, we have a maximum before proceeding to the next step. By examining the function (12.2.13), we can determine a reasonable order of maximization.

First, for each value of α, β, and σ_δ^2, to maximize L with respect to $\xi_1, \ldots, \xi_n$ we minimize $\sum_{i=1}^n \left((x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2\right)$. (See Exercise 12.3 for details.) For each i, we have a quadratic in ξ_i and the minimum is attained at

$$\xi_i^* = \frac{x_i + \lambda\beta(y_i - \alpha)}{1 + \lambda\beta^2}.$$

On substituting back we get

$$\sum_{i=1}^n \left((x_i - \xi_i^*)^2 + \lambda(y_i - (\alpha + \beta\xi_i^*))^2\right) = \frac{\lambda}{1 + \lambda\beta^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2.$$

The likelihood function now becomes

$$\max_{\xi_1,\ldots,\xi_n} L(\alpha,\beta,\xi_1,\ldots,\xi_n,\sigma_\delta^2|(\mathbf{x},\mathbf{y}))$$

$$= \frac{1}{(2\pi)^n}\frac{\lambda^{n/2}}{(\sigma_\delta^2)^n}\exp\left\{-\frac{1}{2\sigma_\delta^2}\left[\frac{\lambda}{1+\lambda\beta^2}\sum_{i=1}^n (y_i-(\alpha+\beta x_i))^2\right]\right\}. \tag{12.2.14}$$

Now, we can maximize with respect to α and β, but a little work will show that we have already done this in the orthogonal least squares solution! Yes, there is somewhat of a correspondence between orthogonal least squares and maximum likelihood in the EIV model and we are about to exploit it. Define

$$\alpha^* = \sqrt{\lambda}\alpha, \quad \beta^* = \sqrt{\lambda}\beta, \quad y_i^* = \sqrt{\lambda}y_i, \quad i=1,\ldots,n. \tag{12.2.15}$$

The exponent of (12.2.14) becomes

$$\frac{\lambda}{1+\lambda\beta^2}\sum_{i=1}^n (y_i-(\alpha+\beta x_i))^2 = \frac{1}{1+\beta^{*2}}\sum_{i=1}^n (y_i^*-(\alpha^*+\beta^* x_i))^2,$$

which is identical to the expression in the orthogonal least squares problem. From (12.2.11) we know the minimizing values of α^* and β^*, and using (12.2.15) we obtain our MLEs for the slope and intercept:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{-(S_{xx}-\lambda S_{yy}) + \sqrt{(S_{xx}-\lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}}. \tag{12.2.16}$$

It is clear from the formula that, at $\lambda = 1$, the MLEs agree with the orthogonal least squares solutions. This makes sense. The orthogonal least squares solution treated x and y as having the same magnitude of error and this translates into a variance ratio of 1. Carrying this argument further, we can relate this solution to ordinary least squares or maximum likelihood when the xs are assumed to be fixed. If the xs are fixed, their variance is 0 and hence $\lambda = 0$. The maximum likelihood solution for general λ does reduce to ordinary least squares in this case. This relationship, among others, is explored in Exercise 12.4.

Putting (12.2.16) together with (12.2.14), we now have almost completely maximized the likelihood. We have

$$\max_{\alpha,\beta,\xi_1,\ldots,\xi_n} L(\alpha,\beta,\xi_1,\ldots,\xi_n,\sigma_\delta^2|(\mathbf{x},\mathbf{y}))$$

$$= \frac{1}{(2\pi)^n}\frac{\lambda^{n/2}}{(\sigma_\delta^2)^n}\exp\left[-\frac{1}{2\sigma_\delta^2}\frac{\lambda}{1+\lambda\hat{\beta}^2}\sum_{i=1}^n \left(y_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2\right]. \tag{12.2.17}$$

Now maximizing L with respect to σ_δ^2 is very similar to finding the MLE of σ^2 in ordinary normal sampling (see Example 7.2.11), the major difference being the exponent of n, rather than $n/2$, on σ_δ^2. The details are left to Exercise 12.5. The

resulting MLE for σ_δ^2 is

$$\hat{\sigma}_\delta^2 = \frac{1}{2n}\frac{\lambda}{1+\lambda\hat{\beta}^2}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha}+\hat{\beta}x_i)\right)^2. \tag{12.2.18}$$

From the properties of MLEs, it follows that the MLE of σ_ϵ^2 is given by $\hat{\sigma}_\epsilon^2 = \hat{\sigma}_\delta^2/\lambda$ and $\hat{\xi}_i = \hat{\alpha} + \hat{\beta}x_i$. Although the $\hat{\xi}_i$s are not usually of interest, they can sometimes be useful if prediction is desired. Also, the $\hat{\xi}_i$s are useful in examining the adequacy of the fit (see Fuller 1987).

It is interesting to note that although $\hat{\alpha}$ and $\hat{\beta}$ are consistent estimators, σ_δ^2 is not. More precisely, as $n \to \infty$,

$$\hat{\alpha} \to \alpha \text{ in probability,}$$

$$\hat{\beta} \to \beta \text{ in probability,}$$

but

$$\hat{\sigma}_\delta^2 \to \frac{1}{2}\sigma_\delta^2 \text{ in probability.}$$

General results on consistency in EIV functional relationship models have been obtained by Gleser (1981).

We now turn to the linear structural relationship model. Recall that here we assume that we observe pairs $(X_i, Y_i), i = 1, \ldots, n$, according to

$$Y_i \sim \mathrm{n}(\alpha + \beta\xi_i, \sigma_\epsilon^2),$$

$$X_i \sim \mathrm{n}(\xi_i, \sigma_\delta^2),$$

$$\xi_i \sim \mathrm{n}(\xi, \sigma_\xi^2),$$

where the ξ_is are independent and, given the ξ_is, the X_is and Y_is are independent. As mentioned before, inference about α and β will be made from the marginal distribution of X_i and Y_i, that is, the distribution obtained by integrating out ξ_i. If we integrate out ξ_i, we obtain the marginal distribution of (X_i, Y_i) (see Exercise 12.6):

$$(X_i, Y_i) \sim \text{ bivariate normal}(\xi, \alpha + \beta\xi, \sigma_\delta^2 + \sigma_\xi^2, \sigma_\epsilon^2 + \beta^2\sigma_\xi^2, \beta\sigma_\xi^2). \tag{12.2.19}$$

Notice the similarity of the correlation structure to that of the RCB ANOVA (see Miscellanea 11.5.3). There, conditional on blocks, the observations were uncorrelated, but unconditionally, there was correlation (the intraclass correlation). Here, the functional relationship model, which is conditional on the ξ_is, has uncorrelated observations, but the structural relationship model, where we infer unconditional on the ξ_is, has correlated observations. The ξ_is are playing a role similar to blocks and the correlation that appears here is similar to the intraclass correlation. (In fact, it is identical to the intraclass correlation if $\beta = 1$ and $\sigma_\delta^2 = \sigma_\epsilon^2$.)

To proceed with likelihood estimation in this case, given observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$, the likelihood function is that of a bivariate normal, as was encountered in Exercise 7.18. There, it was seen that the likelihood estimators in the bivariate normal could be found by equating sample quantities to population quantities. Hence, to find the MLEs of α, β, ξ, σ_ϵ^2, σ_δ^2, and σ_ξ^2, we solve

$$\begin{aligned}
\bar{y} &= \hat{\alpha} + \hat{\beta}\hat{\xi}, \\
\bar{x} &= \hat{\xi}, \\
\frac{1}{n}S_{yy} &= \hat{\sigma}_\epsilon^2 + \hat{\beta}^2\hat{\sigma}_\xi^2, \\
\frac{1}{n}S_{xx} &= \hat{\sigma}_\delta^2 + \hat{\sigma}_\xi^2, \\
\frac{1}{n}S_{xy} &= \hat{\beta}\hat{\sigma}_\xi^2.
\end{aligned} \tag{12.2.20}$$

Note that we have five equations, but there are six unknowns, so the system is indeterminate. That is, the system of equations does not have a unique solution and there is no unique value of the parameter vector $(\alpha, \beta, \xi, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$ that maximizes the likelihood.

Before we go on, realize that the variances of X_i and Y_i here are different from the variances in the functional relationship model. There we were working conditional on $\xi_1, \ldots, \xi_n$, and here we are working marginally with respect to the ξ_is. So, for example, in the functional relationship model we write $\operatorname{Var} X_i = \sigma_\delta^2$ (where it is understood that this variance is conditional on $\xi_1, \ldots, \xi_n$), while in the structural model we write $\operatorname{Var} X_i = \sigma_\delta^2 + \sigma_\xi^2$ (where it is understood that this variance is unconditional on $\xi_1, \ldots, \xi_n$). This should not be a source of confusion.

A solution to the equations in (12.2.20) implies a restriction on $\hat{\beta}$, a restriction that we have already encountered in the functional relationship case (see Exercise 12.4). From the above equations involving the variances and covariance, it is straightforward to deduce that

$$\begin{aligned}
\hat{\sigma}_\delta^2 \geq 0 \quad &\text{only if } S_{xx} \geq \tfrac{1}{\hat{\beta}} S_{xy}, \\
\hat{\sigma}_\epsilon^2 \geq 0 \quad &\text{only if } S_{yy} \geq \hat{\beta} S_{xy},
\end{aligned}$$

which together imply that

$$\frac{|S_{xy}|}{S_{xx}} \leq |\hat{\beta}| \leq \frac{S_{yy}}{|S_{xy}|}.$$

(The bounds on $\hat{\beta}$ are established in Exercise 12.9.)

We now address the identifiability problem in the structural relationship case, a problem that can be expected since, in (12.2.19) we have more parameters than are needed to specify the distribution. To make the structural linear relationship model identifiable, we must make an assumption that reduces the number of parameters to five. It fortunately happens that the assumption about variances made for the functional relationship solves the identifiability problem here. Thus, we assume that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, where λ is known. This reduces the number of unknown parameters to five and makes the model identifiable. (See Exercise 12.8.) More restrictive assumptions, such as assuming that σ_δ^2 is known, may lead to MLEs of variances that have the value 0. Kendall and Stuart (1979, Chapter 29) have a full discussion of this.

Once we assume that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, the maximum likelihood estimates for $\hat{\alpha}$ and $\hat{\beta}$ in this model are the same as in the functional relationship model and are given by (12.2.16). The variance estimates are different, however, and are given by

$$\hat{\sigma}_\delta^2 = \frac{1}{n}\left(S_{xx} - \frac{S_{xy}}{\hat{\beta}}\right),$$

$$\hat{\sigma}_\epsilon^2 = \frac{\hat{\sigma}_\delta^2}{\lambda} = \frac{1}{n}(S_{yy} - \hat{\beta}S_{xy}), \tag{12.2.21}$$

$$\hat{\sigma}_\xi^2 = \frac{1}{n}\frac{S_{xy}}{\hat{\beta}}.$$

(Exercise 12.10 shows this and also explores the relationship between variance estimates here and in the functional model.) Note that, in contrast to what happened in the functional relationship model, these estimators are all consistent in the linear structural relationship model (when $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$).

12.2.4 Confidence Sets

As might be expected, the construction of confidence sets in the EIV model is a difficult task. A complete treatment of the subject needs machinery that we have not developed. In particular, we will concentrate here only on confidence sets for the slope, β.

As a first attack, we could use the approximate likelihood method of Section 10.4.1 to construct approximate confidence intervals. In practice this is probably what is most often done and is not totally unreasonable. However, these approximate intervals cannot maintain a nominal $1-\alpha$ confidence level. In fact, results of Gleser and Hwang (1987) yield the rather unsettling result that any interval estimator of the slope whose length is *always finite* will have confidence coefficient equal to 0!

For definiteness, in the remainder of this section we will assume that we are in the structural relationship case of the EIV model. The confidence set results presented are valid in both the structural and functional cases and, in particular, the formulas remain the same. We continue to assume that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, where λ is known.

Gleser and Hwang (1987) identify the parameter

$$\tau^2 = \frac{\sigma_\xi^2}{\sigma_\delta^2}$$

as determining the amount of information potentially available in the data to determine the slope β. They show that, as $\tau^2 \to 0$, the coverage probability of any finite-length confidence interval on β must also go to 0. To see why this is plausible, note that $\tau^2 = 0$ implies that the ξ_is do not vary and it would be impossible to fit a unique straight line.

An approximate confidence interval for β can be constructed by using the fact that the estimator

$$\hat{\sigma}_\beta^2 = \frac{(1+\lambda\hat{\beta}^2)^2(S_{xx}S_{yy} - S_{xy}^2)}{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}$$

is a consistent estimator of $\sigma_{\hat{\beta}}^2$, the true variance of $\hat{\beta}$. Hence, using the CLT together with Slutsky's Theorem (see Section 5.5), we can show that the interval

$$\hat{\beta} - \frac{z_{\alpha/2}\hat{\sigma}_{\hat{\beta}}}{\sqrt{n}} \leq \beta \leq \hat{\beta} + \frac{z_{\alpha/2}\hat{\sigma}_{\hat{\beta}}}{\sqrt{n}}$$

is an approximate $1-\alpha$ confidence interval for β. However, since it has finite length, it cannot maintain $1-\alpha$ coverage for all parameter values.

Gleser (1987) considers a modification of this interval and reports the infimum of its coverage probabilities as a function of τ^2. Gleser's modification, $C_G(\hat{\beta})$, is

$$\hat{\beta} - \frac{t_{n-2,\alpha/2}\hat{\sigma}_{\hat{\beta}}}{\sqrt{n-2}} \leq \beta \leq \hat{\beta} + \frac{t_{n-2,\alpha/2}\hat{\sigma}_{\hat{\beta}}}{\sqrt{n-2}}. \tag{12.2.22}$$

Again using the CLT together with Slutsky's Theorem, we can show that this is an approximate $1-\alpha$ confidence interval for β. Since this interval also has finite length, it also cannot maintain $1-\alpha$ coverage for all parameter values. Gleser does some finite-sample numerical calculations and gives bounds on the infima of the coverage probabilities as a function of τ^2. For reasonable values of n ($\geq$10), the coverage probability of a nominal 90% interval will be at least 80% if $\tau^2 \geq .25$. As τ^2 or n increases, this performance improves.

In contrast to $C_G(\hat{\beta})$ of (12.2.22), which has finite length but no guaranteed coverage probability, we now look at an exact confidence set that, as it must, has infinite length. The set, known as the Creasy–Williams confidence set, is due to Creasy (1956) and Williams (1959) and is based on the fact (see Exercise 12.11) that if $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$, then

$$\mathrm{Cov}(\beta\lambda Y_i + X_i, Y_i - \beta X_i) = 0.$$

Define $r_\lambda(\beta)$ to be the sample correlation coefficient between $\beta\lambda Y_i + X_i$ and $Y_i - \beta X_i$, that is,

$$\begin{aligned} r_\lambda(\beta) &= \frac{\sum_{i=1}^n \left((\beta\lambda y_i + x_i) - (\beta\lambda\bar{y} + \bar{x})\right)\left((y_i - \beta x_i) - (\bar{y} - \beta\bar{x})\right)}{\sqrt{\sum_{i=1}^n \left((\beta\lambda y_i + x_i) - (\beta\lambda\bar{y} + \bar{x})\right)^2 \sum_{i=1}^n \left((y_i - \beta x_i) - (\bar{y} - \beta\bar{x})\right)^2}} \\ &= \frac{\beta\lambda S_{yy} + (1-\beta^2\lambda)S_{xy} - \beta S_{xx}}{\sqrt{(\beta^2\lambda^2 S_{yy} + 2\beta\lambda S_{xy} + S_{xx})(S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx})}}. \end{aligned} \tag{12.2.23}$$

Since $\beta\lambda Y_i + X_i$ and $Y_i - \beta X_i$ are bivariate normal with correlation 0, it follows (see Exercise 11.33) that

$$\frac{\sqrt{n-2}\, r_\lambda(\beta)}{\sqrt{1 - r_\lambda^2(\beta)}} \sim t_{n-2}$$

for any value of β. Thus, we have identified a pivotal quantity and we conclude that

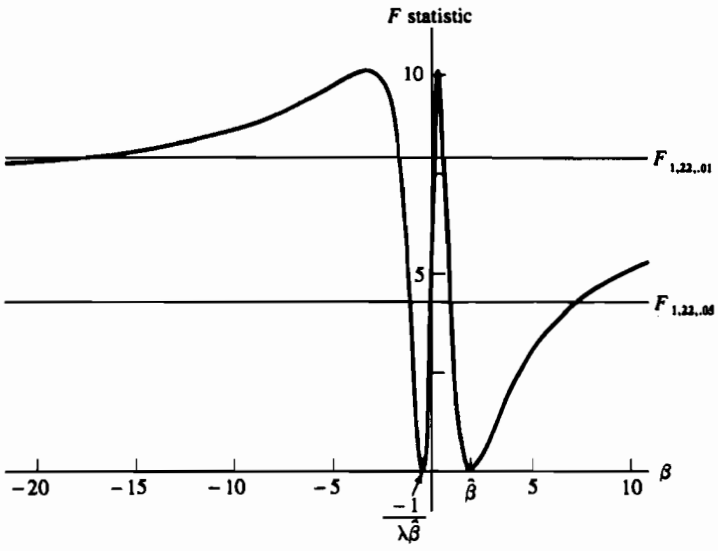

Figure 12.2.3. *F statistic defining Creasy–Williams confidence set,* $\lambda = 1$

the set

$$\left\{\beta : \frac{(n-2)r_\lambda^2(\beta)}{1-r_\lambda^2(\beta)} \le F_{1,n-2,\alpha}\right\} \tag{12.2.24}$$

is a $1-\alpha$ confidence set for β (see Exercise 12.11).

Although this confidence set is a $1-\alpha$ set, it suffers from defects similar to those of Fieller's intervals. The function describing the set (12.2.24) has two minima, where the function is 0. The confidence set can consist of two finite disjoint intervals, one finite and two infinite disjoint intervals, or the entire real line. For example, the graph of the F statistic function for the data in Table 11.3.1 with $\lambda = 1$ is in Figure 12.2.3. The confidence set is all the βs where the function is less than or equal to $F_{1,22,\alpha}$. For $\alpha = .05$ and $F_{1,22,.05} = 4.30$, the confidence set is $[-1.13, -.14] \cup [.89, 7.38]$. For $\alpha = .01$ and $F_{1,22,.01} = 7.95$, the confidence set is $(-\infty, -18.18] \cup [-1.68, .06] \cup [.60, \infty)$.

Furthermore, for every value of β, $-r_\lambda(\beta) = r_\lambda(-1/(\lambda\beta))$ (see Exercise 12.12) so that if β is in the confidence set, so is $-1/(\lambda\beta)$. Using this confidence set, we cannot distinguish β from $-1/(\lambda\beta)$ and this confidence set always contains both positive and negative values. We can never determine the sign of the slope from this confidence set!

The confidence set given in (12.2.24) is not exactly the one discussed by Creasy (1956) but rather a modification. She was actually interested in estimating ϕ, the angle that β makes with the x-axis, that is, $\beta = \tan(\phi)$, and confidence sets there have fewer problems. Estimation of ϕ is perhaps more natural in EIV models (see, for example, Anderson 1976), but we seem to be more inclined to estimate α and β.

Most of the other standard statistical analyses that can be done in the ordinary linear regression case have analogues in EIV models. For example, we can test hy-

potheses about β or estimate values of EY_i. More about these topics can be found in Fuller (1987) or Kendall and Stuart (1979, Chapter 29).

12.3 Logistic Regression

The conditional normal model of Section 11.3.3 is an example of a *generalized linear model* (GLM). A GLM describes a relationship between the mean of a response variable Y and an independent variable x. But the relationship may be more complicated than the $\mathrm{E}\,Y_i = \alpha + \beta x_i$ of (11.3.2). Many different models can be expressed as GLMs. In this section, we will concentrate on a specific GLM, the logistic regression model.

12.3.1 The Model

A GLM consists of three components: the random component, the systematic component, and the link function.

(1) The response variables $Y_1, \ldots, Y_n$ are the *random component.* They are assumed to be independent random variables, each with a distribution from a specified exponential family. The Y_is are not identically distributed, but they each have a distribution from the same family: binomial, Poisson, normal, etc.

(2) The *systematic component* is the model. It is the function of the predictor variable x_i, linear in the parameters, that is related to the mean of Y_i. So the systematic component could be $\alpha + \beta x_i$ or $\alpha + \beta/x_i$, for example. We will consider only $\alpha + \beta x_i$ here.

(3) Finally, the *link function* $g(\mu)$ links the two components by asserting that $g(\mu_i) = \alpha + \beta x_i$, where $\mu_i = \mathrm{E}\,Y_i$.

The conditional normal regression model of Section 11.3.3 is an example of a GLM. In that model, the responses Y_is all have normal distributions. Of course, the normal family is an exponential family, which is the random component. The form of the regression function is $\alpha + \beta x_i$ in this model, which is the systematic component. Finally, the relationship $\mu_i = \mathrm{E}\,Y_i = \alpha + \beta x_i$ is assumed. This means the link function is $g(\mu) = \mu$. This simple link function is called the *identity link.*

Another very useful GLM is the *logistic regression model.* In this model, the responses $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \text{Bernoulli}(\pi_i)$. (The Bernoulli family is an exponential family.) Recall, $\mathrm{E}\,Y_i = \pi_i = P(Y_i = 1)$. In this model, π_i is assumed to be related to x_i by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i. \tag{12.3.1}$$

The left-hand side is the log of the odds of success for Y_i. The model assumes this log-odds (or *logit*) is a linear function of the predictor x. The Bernoulli pmf can be written in exponential family form as

$$\pi^y(1-\pi)^{1-y} = (1-\pi)\exp\left\{y\log\left(\frac{\pi_i}{1-\pi_i}\right)\right\}.$$

The term $\log(\pi/(1-\pi))$ is the natural parameter of this exponential family, and in (12.3.1) the link function $g(\pi) = \log(\pi/(1-\pi))$ is used. When the natural parameter is used in this way, it is called the *canonical link.*

Equation (12.3.1) can be rewritten as

$$\pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

or, more generally,

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \tag{12.3.2}$$

We see that $0 < \pi(x) < 1$, which seems appropriate because $\pi(x)$ is a probability. But, if it is possible that $\pi(x) = 0$ or 1 for some x, then this model is not appropriate. If we examine $\pi(x)$ more closely, its derivative can be written

$$\frac{d\pi(x)}{dx} = \beta\pi(x)(1-\pi(x)). \tag{12.3.3}$$

As the term $\pi(x)(1-\pi(x))$ is always positive, the derivative of $\pi(x)$ is positive, 0, or negative according as β is positive, 0, or negative. If β is positive, $\pi(x)$ is a strictly increasing function of x; if β is negative, $\pi(x)$ is a strictly decreasing function of x; if $\beta = 0$, $\pi(x) = e^{\alpha}/(1+e^{\alpha})$ for all x. As in simple linear regression, if $\beta = 0$, there is no relationship between π and x. Also, in a logistic regression model, $\pi(-\alpha/\beta) = 1/2$. A logistic regression function exhibits this kind of symmetry; for any c, $\pi((-\alpha/\beta)+c) = 1-\pi((-\alpha/\beta)-c)$.

The parameters α and β have meanings similar to those in simple linear regression. Setting $x = 0$ in (12.3.1) yields that α is the log-odds of success at $x = 0$. Evaluating (12.3.1) at x and $x+1$ yields, for any x,

$$\log\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha+\beta(x+1)-\alpha-\beta(x) = \beta.$$

Thus, β is the change in the log-odds of success corresponding to a one-unit increase in x. In simple linear regression, β is the change in the mean of Y corresponding to a one-unit increase in x. Exponentiating both sides of this equality yields

$$e^{\beta} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}. \tag{12.3.4}$$

The right-hand side is the *odds ratio* comparing the odds of success at $x+1$ to the odds of success at x. (Recall that in Examples 5.5.19 and 5.5.22 we looked at estimating odds.) In a logistic regression model this ratio is constant as a function of x. Finally,

$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\beta}\frac{\pi(x)}{1-\pi(x)}; \tag{12.3.5}$$

that is, e^{β} is the multiplicative change in the odds of success corresponding to a one-unit increase in x.

Equation (12.3.2) suggests other ways of modeling the Bernoulli success probability $\pi(x)$ as a function of the predictor variable x. Recall that $F(w) = e^w/(1+e^w)$ is the cdf of a logistic(0, 1) distribution. In (12.3.2) we have assumed $\pi(x) = F(\alpha + \beta x)$. We can define other models for $\pi(x)$ by using other continuous cdfs. If $F(w)$ is the standard normal cdf, the model is called probit regression (see Exercise 12.17). If a Gumbel cdf is used, the link function is called the log-log link.

12.3.2 Estimation

In linear regression, where we use a model such as $Y_i = \alpha + \beta x_i + \varepsilon_i$, the technique of least squares was an option for calculating estimates of α and β. This is no longer the case here. In the model (12.3.1) with $Y_i \sim \text{Bernoulli}(\pi_i)$, we no longer have a direct connection between Y_i and $\alpha + \beta x_i$ (which is why we need a link function). Thus, least squares is no longer an option.

The estimation method that is most commonly used is maximum likelihood. In the general model we have $Y_i \sim \text{Bernoulli}(\pi_i)$, where $\pi(x) = F(\alpha + \beta x)$. If we let $F_i = F(\alpha + \beta x_i)$, then the likelihood function is

$$L(\alpha, \beta|\mathbf{y}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} = \prod_{i=1}^{n} F_i^{y_i}(1 - F_i)^{1-y_i}$$

with log likelihood

$$\log L(\alpha, \beta|\mathbf{y}) = \sum_{i=1}^{n} \left\{ \log(1 - F_i) + y_i \log\left(\frac{F_i}{1 - F_i}\right) \right\}.$$

We obtain the likelihood equations by differentiating the log likelihood with respect to α and β. Let $dF(w)/dw = f(w)$, the pdf corresponding to $F(w)$, and let $f_i = f(\alpha + \beta x_i)$. Then

$$\frac{\partial \log(1 - F_i)}{\partial \alpha} = -\frac{f_i}{1 - F_i} = -\frac{F_i f_i}{F_i(1 - F_i)}$$

and

$$\frac{\partial}{\partial \alpha} \log\left(\frac{F_i}{1 - F_i}\right) = \frac{f_i}{F_i(1 - F_i)}. \tag{12.3.6}$$

Hence,

$$\frac{\partial}{\partial \alpha} \log L(\alpha, \beta|\mathbf{y}) = \sum_{i=1}^{n} (y_i - F_i) \frac{f_i}{F_i(1 - F_i)}. \tag{12.3.7}$$

A similar calculation yields

$$\frac{\partial}{\partial \beta} \log L(\alpha, \beta|\mathbf{y}) = \sum_{i=1}^{n} (y_i - F_i) \frac{f_i}{F_i(1 - F_i)} x_i. \tag{12.3.8}$$

For logistic regression, with $F(w) = e^w/(1+e^w)$, $f_i/[F_i(1-F_i)] = 1$, and (12.3.7) and (12.3.8) are somewhat simpler.

The MLEs are obtained by setting (12.3.7) and (12.3.8) equal to 0 and solving for α and β. These are nonlinear equations in α and β and must be solved numerically. (This will be discussed later.) For logistic and probit regression, the log likelihood is strictly concave. Hence, if the likelihood equations have a solution, it is unique and it is the MLE. However, for some extreme data the likelihood equations do not have a solution. The maximum of the likelihood occurs in some limit as the parameters go to $\pm\infty$. See Exercise 12.16 for an example. This is because the logistic model assumes $0 < \pi(x) < 1$, but, for certain data sets, the maximum of the logistic likelihood occurs in a limit with $\pi(x) = 0$ or 1. The probability of obtaining such data converges to 0 if the logistic model is true.

Example 12.3.1 (Challenger data) A by now infamous data set is that of space shuttle O-ring failures, which have been linked to temperature. The data in Table 12.3.1 give the temperatures at takeoff and whether or not an O-ring failed.

Solving (12.3.6) and (12.3.7) using $F(\alpha+\beta x_i) = e^{\alpha+\beta x_i}/(1+e^{\alpha+\beta x_i})$ yields MLEs $\hat{\alpha} = 15.043$ and $\hat{\beta} = -.232$. Figure 12.3.1 shows the fitted curve along with the data.

The space shuttle Challenger exploded during takeoff, killing the seven astronauts aboard. The explosion was the result of an O-ring failure, believed to be caused by the unusually cold weather (31° F) at the time of launch. The MLE of the probability of O-ring failure at 31° is .9996. (See Dalal, Fowlkes, and Hoadley 1989 for the full story.)

Table 12.3.1. *Temperature at flight time (°F) and failure of O-rings (1 = failure, 0 = success)*

Flight no.	14	9	23	10	1	5	13	15	4	3	8	17
Failure	1	1	1	1	0	0	0	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70	70
Flight no.	2	11	6	7	16	21	19	22	12	20	18	
Failure	1	1	0	0	0	1	0	0	0	0	0	
Temp.	70	70	72	73	75	75	76	76	78	79	81	

We have, thus far, assumed that at each value of x_i, we observe the result of only one Bernoulli trial. Although this often the case, there are many situations in which there are multiple Bernoulli observations at each value of x. We now revisit the likelihood solution in this more general case.

Suppose there are J different values of the predictor x in the data set $x_1, \ldots, x_J$. Let n_j denote the number of Bernoulli observations at x_j, and let Y_j^* denote the number of successes in these n_j observations. Thus, $Y_j^* \sim$ binomial$(n_j, \pi(x_j))$. Then the likelihood is

$$L(\alpha, \beta|\mathbf{y}^*) = \prod_{j=1}^{J} \pi(x_i)^{y_j^*}(1-\pi(x_i))^{n_j-y_j^*} = \prod_{j=1}^{J} F_j^{y_j^*}(1-F_j)^{n_j-y_j^*},$$

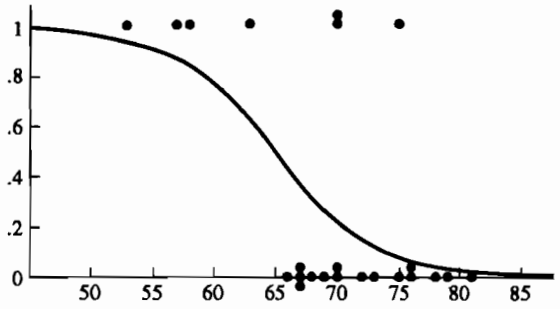

Figure 12.3.1. *The data of Table 12.3.1 along with the fitted logistic curve*

and the likelihood equations are

$$0 = \sum_{j=1}^{J} (y_j^* - n_j F_j) \frac{f_j}{F_j(1-F_j)}$$

$$0 = \sum_{j=1}^{J} (y_j^* - n_j F_j) \frac{f_j}{F_j(1-F_j)} x_j.$$

As we have estimated the parameters of the logistic regression using maximum likelihood, we can next get approximate variances using MLE asymptotics. However, we have to proceed in a more general way. In Section 10.1.3 we saw how to approximate the variance of the MLE using the information number. We use the same strategy here, but as there are two parameters, there is an *information matrix* given by the 2×2 matrix

$$(12.3.9) \quad I(\theta_1, \theta_2) = \begin{pmatrix} -\frac{\partial^2}{\partial \theta_1^2} \log L(\theta_1, \theta_2 | \mathbf{y}) & -\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L(\theta_1, \theta_2 | \mathbf{y}) \\ -\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L(\theta_1, \theta_2 | \mathbf{y}) & -\frac{\partial^2}{\partial \theta_2^2} \log L(\theta_1, \theta_2 | \mathbf{y}) \end{pmatrix}$$

For logistic regression, the information matrix is given by

$$(12.3.10) \quad I(\alpha, \beta) = \begin{pmatrix} \sum_{j=1}^{J} n_j F_j (1-F_j) & \sum_{j=1}^{J} x_j n_j F_j (1-F_j) \\ \sum_{j=1}^{J} x_j n_j F_j (1-F_j) & \sum_{j=1}^{J} x_j^2 n_j F_j (1-F_j) \end{pmatrix},$$

and the variances of the MLEs $\hat{\alpha}$ and $\hat{\beta}$ are usually approximated using this matrix. Note that the elements of $I(\alpha, \beta)$ do not depend on $Y_1^*, \ldots, Y_J^*$. Thus, the observed information is the same as the information in this case.

In Section 10.1.3, we used the approximation (10.1.7), namely, $\mathrm{Var}(h(\hat{\theta})|\theta) \approx [h'(\hat{\theta})]^2 / I(\hat{\theta})$, where $I(\cdot)$ was the information number. Here, we cannot do exactly the same with the information matrix, but rather we need to get the inverse of the

matrix and use the inverse elements to approximate the variance. Recall that the inverse of a 2×2 matrix is given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

To obtain the approximate variances, the MLEs are used to estimate the parameters in the matrix (12.3.10), and the estimates of the variances, $[\text{se}(\hat{\alpha})]^2$ and $[\text{se}(\hat{\beta})]^2$, are the diagonal elements of the inverse of $I(\hat{\alpha}, \hat{\beta})$. (The notation $\text{se}(\hat{\alpha})$ stands for *standard error of* $(\hat{\alpha})$.)

Example 12.3.2 (Challenger data continued) The estimated information matrix of the estimates from the Challenger data is given by

$$I(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} \sum_{j=1}^{J} \hat{F}_j(1 - \hat{F}_j) & \sum_{j=1}^{J} x_j \hat{F}_j(1 - \hat{F}_j) \\ \sum_{j=1}^{J} x_j \hat{F}_j(1 - \hat{F}_j) & \sum_{j=1}^{J} x_j^2 \hat{F}_j(1 - \hat{F}_j j) \end{pmatrix} = \begin{pmatrix} 3.15 & 214.75 \\ 214.75 & 14728.5 \end{pmatrix},$$

where $\hat{F}_j = e^{\hat{\alpha}+\hat{\beta}x_j}/(1 + e^{\hat{\alpha}+\hat{\beta}x_j})$ and has inverse

$$I(\hat{\alpha}, \hat{\beta})^{-1} = \begin{pmatrix} 54.44 & -.80 \\ -.80 & .012 \end{pmatrix}.$$

The likelihood asymptotics tell use that, for example, $\hat{\beta} \pm z_{\alpha/2}\text{se}(\hat{\beta})$ is, for large samples, an approximate $100(1-\alpha)\%$ confidence interval for β. So for the Challenger data we have a 95% confidence interval of

$$\beta \in -.232 \pm 1.96 \times \sqrt{.012} \quad \Rightarrow \quad -.447 \leq \beta \leq -.017,$$

supporting the conclusion that $\beta < 0$. ||

It is, perhaps, most common in this model to test the hypothesis $H_0 : \beta = 0$, because, as in simple linear regression, this hypothesis states there is no relationship between the predictor and response variables. The Wald test statistic, $Z = \hat{\beta}/\text{se}(\hat{\beta})$, has approximately a standard normal distribution if H_0 is true and the sample size is large. Thus, H_0 can be rejected if $|Z| \geq z_{\alpha/2}$. Alternatively, H_0 can be tested with the log LRT statistic

$$-2 \log \lambda(\mathbf{y}^*) = 2[\log L(\hat{\alpha}, \hat{\beta}|\mathbf{y}^*) - L(\hat{\alpha}_0, 0|\mathbf{y}^*)],$$

where $\hat{\alpha}_0$ is the MLE of α assuming $\beta = 0$. With standard binomial arguments (see Exercise 12.20), it can be shown that $\hat{\alpha}_0 = \sum_{i=1}^{n} y_i/n = \sum_{j=1}^{J} y_j^* / \sum_{j=1}^{J} n_j$. Therefore, under H_0, $-2 \log \lambda$ has an approximate χ_1^2 distribution, and we can reject H_0 at level α if $-2 \log \lambda \geq \chi_{1,\alpha}^2$.

We have introduced only the simplest logistic regression and generalized linear models. Much more can be found in standard texts such as Agresti (1990).

12.4 Robust Regression

As in Section 10.2, we now want to take a look at the performance of our procedures if the underlying model is not the correct one, and we will take a look at some robust alternatives to least squares estimation, starting with a comparison analogous to the mean/median comparison.

Recall that, when observing $x_1, x_2, \ldots, x_n$, we can define the mean and the median as minimizers of the following quantities:

$$\text{mean: } \min_m \left\{ \sum_{i=1}^n (x_i - m)^2 \right\}, \qquad \text{median: } \min_m \left\{ \sum_{i=1}^n |x_i - m| \right\}.$$

For simple linear regression, observing $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, we know that least squares regression estimates satisfy

$$\text{least squares: } \min_{a,b} \left\{ \sum_{i=1}^n [y_i - (a + bx_i)]^2 \right\},$$

and we analogously define *least absolute deviation (LAD)* regression estimates by

$$\text{least absolute deviation: } \min_{a,b} \left\{ \sum_{i=1}^n |y_i - (a + bx_i)| \right\}.$$

(The LAD estimates may not be unique. See Exercise 12.25.)

Thus, we see that the least squares estimators are the regression analogues of the sample mean. This should make us wonder about their robustness performance (as listed in items (1)–(3) of Section 10.2).

Example 12.4.1 (Robustness of least squares estimates) If we observe (y_1, x_1), (y_2, x_2), $\ldots$, (y_n, x_n), where

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

and the ε_i are uncorrelated with $\mathrm{E}\,\varepsilon_i = 0$ and $\mathrm{Var}\,\varepsilon_i = \sigma^2$, we know that the least squares estimator b with variance $\sigma^2 / \sum (x_i - \bar{x})^2$ is the BLUE of β, satisfying (1) of Section 10.2.

To investigate how b performs for small deviations, we assume that

$$\mathrm{Var}(\varepsilon_i) = \begin{cases} \sigma^2 & \text{with probability } 1 - \delta \\ \tau^2 & \text{with probability } \delta. \end{cases}$$

Writing $b = \sum d_i Y_i$, where $d_i = (x_i - \bar{x}) / \sum (x_i - \bar{x})^2$, we now have

$$\mathrm{Var}(b) = \sum_{i=1}^n d_i^2 \,\mathrm{Var}(\varepsilon_i) = \frac{(1 - \delta)\sigma^2 + \delta\tau^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

This shows that, as with the sample mean, for small perturbations b performs pretty well. (But we could, of course, blow things up by contaminating with a Cauchy pdf,

Table 12.4.1. *Values of CO_2 and O_2 in the pouches of 23 potoroos (McPherson 1990)*

Animal	1	2	3	4	5	6	7	8
% O_2	20	19.6	19.6	19.4	18.4	19	19	18.3
% CO_2	1	1.2	1.1	1.4	2.3	1.7	1.7	2.4
Animal	9	10	11	12	13	14	15	16
% O_2	18.2	18.6	19.2	18.2	18.7	18.5	18	17.4
% CO_2	2.1	2.1	1.2	2.3	1.9	2.4	2.6	2.9
Animal	17	18	19	20	21	22	23	
% O_2	16.5	17.2	17.3	17.8	17.3	18.4	16.9	
% CO_2	4.0	3.3	3.0	3.4	2.9	1.9	3.9	

for example.) The behavior of the least squares intercept a is similar (see Exercise 12.22). See also Exercise 12.23 to see how bias contamination affects things. ||

We next look at the effect of a "catastrophic" observation and compare least squares to its median-analogous alternative, least absolute deviation regression.

Example 12.4.2 (Catastrophic observations) McPherson (1990) describes an experiment in which the levels of carbon dioxide (CO_2) and oxygen (O_2) were measured in the pouches of 24 *potoroos* (a marsupial). Interest is in the regression of CO_2 on O_2, where the experimenter expects a slope of -1. The data for 23 animals (one had missing values) are given in Table 12.4.1. For the original data, the least squares and LAD lines are quite close:

$$\text{least squares: } y = 18.67 - .89x,$$

$$\text{least absolute deviation: } y = 18.59 - .89x.$$

However, an aberrant observation can upset least squares. When entering the data the O_2 value of 18 on Animal 15 was incorrectly entered as 10 (we really did this). For this new (incorrect) data set we have

$$\text{least squares: } y = 6.41 - .23x,$$

$$\text{least absolute deviation: } y = 15.95 - .75x,$$

showing that the aberrant observation had much less of an effect on LAD. See Figure 12.4.1 for a display of the regression lines.

These calculations illustrate the resistance of LAD, as opposed to least squares. Since we have the mean/median analogy, we can surmise that this behavior is reflected in breakdown values, which are 0% for least squares and 50% for LAD. ||

However, the mean/median analogy continues. Although the LAD estimator is robust to catastrophic observations, it loses much in efficiency with respect to the least squares estimator (see also Exercise 12.25).

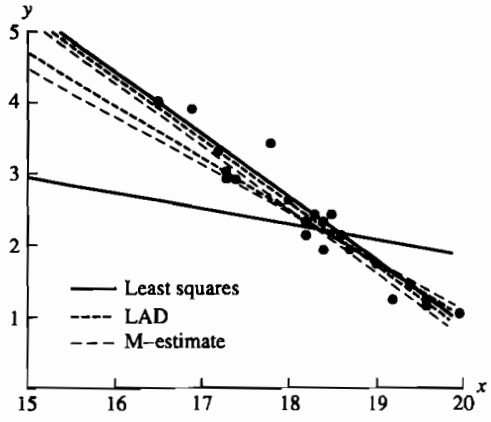

Figure 12.4.1. *Least squares, LAD, and M-estimate fits for the data of Table 12.4.1, for both the original data and the data with* $(18, 2.6)$ *mistyped as* $(10, 2.6)$. *The LAD and M-estimate lines are quite similar, while the least squares line reacts to the changed data.*

Example 12.4.3 (Asymptotic normality of the LAD estimator) We adapt the argument leading to (10.2.6) to derive the asymptotic distribution of the LAD estimator. Also, to simplify things, we consider only the model

$$Y_i = \beta x_i + \varepsilon_i,$$

that is, we set $\alpha = 0$. (This avoids having to deal with a bivariate limiting distribution.)

In the terminology of M-estimators, the LAD estimator is obtained by minimizing

$$\begin{aligned} \sum_{i=1}^{n} \rho(y_i - \beta x_i) &= \sum_{i=1}^{n} |y_i - \beta x_i| \\ &= \sum_{i=1}^{n} (y_i - \beta x_i) I(y_i > \beta x_i) - (y_i - \beta x_i) I(y_i < \beta x_i). \end{aligned} \tag{12.4.1}$$

We then calculate $\psi = \rho'$ and solve $\sum_i \psi(y_i - \beta x_i) = 0$ for β, where

$$\psi(y_i - \beta x_i) = x_i I(y_i > \beta x_i) - x_i I(y_i < \beta x_i).$$

If $\hat{\beta}_L$ is the solution, expand ψ in a Taylor series around β:

$$\sum_{i=1}^{n} \psi(y_i - \hat{\beta}_L x_i) = \sum_{i=1}^{n} \psi(y_i - \beta x_i) + (\hat{\beta}_L - \beta) \frac{d}{d\hat{\beta}_L} \sum_{i=1}^{n} \psi(y_i - \hat{\beta}_L x_i) \Bigg|_{\hat{\beta}_L = \beta} + \cdots.$$

Although the left-hand side of the equation is not equal to 0, we assume that it approaches 0 as $n \to \infty$ (see Exercise 12.27). Then rearranging we obtain

$$\sqrt{n}(\hat{\beta}_L - \beta) = \frac{\frac{-1}{\sqrt{n}} \sum_{i=1}^n \psi(y_i - \beta x_i)}{\frac{1}{n}\frac{d}{d\beta_L} \sum_{i=1}^n \psi(y_i - \hat{\beta}_L x_i)\Big|_{\hat{\beta}_L = \beta}}. \tag{12.4.2}$$

First look at the numerator. As $\mathrm{E}_\beta \psi(Y_i - \hat{\beta}_L x_i) = 0$ and $\mathrm{Var}\, \psi(Y_i - \hat{\beta}_L x_i) = x_i^2$, it follows that

$$\frac{-1}{\sqrt{n}} \sum_{i=1}^n \psi(Y_i - \hat{\beta} x_i) = \sqrt{n}\left[\frac{-1}{n} \sum_{i=1}^n \psi(Y_i - \hat{\beta} x_i)\right] \to \mathrm{n}\left(0, \sigma_x^2\right), \tag{12.4.3}$$

where $\sigma_x^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n x_i^2$. Turning to the denominator, we must be a bit careful as ψ has points of nondifferentiability. We therefore first apply the Law of Large Numbers before differentiating, and use the approximation

$$\begin{aligned} \frac{1}{n}\frac{d}{d\beta_0} \sum_{i=1}^n \psi(y_i - \beta_0 x_i) &\approx \frac{1}{n} \sum_{i=1}^n \frac{d}{d\beta_0} \mathrm{E}_\beta[\psi(Y_i - \beta_0 x_i)] \\ &= \frac{1}{n} \sum_{i=1}^n \frac{d}{d\beta_0} \left[x_i P_\beta(Y_i > \beta_0 x_i) - x_i P_\beta(Y_i < \beta_0 x_i)\right] \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 f(\beta_0 x_i - \beta x_i) + x_i^2 f(\beta_0 x_i - \beta x_i). \end{aligned} \tag{12.4.4}$$

If we now evaluate the derivative at $\beta_0 = \beta$, we have

$$\frac{1}{n}\frac{d}{d\beta_0} \sum_{i=1}^n \psi(y_i - \beta_0 x_i)\Big|_{\beta_0 = \beta} \approx 2f(0)\frac{1}{n} \sum_{i=1}^n x_i^2,$$

and putting this together with (12.4.2) and (12.4.3), we have

$$\sqrt{n}(\hat{\beta}_L - \beta) \to \mathrm{n}\left(0, \frac{1}{4f(0)^2\sigma_x^2}\right). \tag{12.4.5}$$

Finally, for the case of $\alpha = 0$, the least squares estimator is $\hat{\beta} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$ and satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{n}\left(0, \frac{1}{\sigma_x^2},\right)$$

so that the asymptotic relative efficiency of $\hat{\beta}_L$ to $\hat{\beta}$ is

$$\mathrm{ARE}(\hat{\beta}_L, \hat{\beta}) = \frac{1/\sigma_x^2}{1/(4f(0)^2\sigma_x^2)} = 4f(0)^2,$$

whose values are the same as those given in Table 10.2.1, comparing the median to the mean. Thus, for normal errors the ARE of the LAD estimator to least squares is only 64%, showing that the LAD estimator gives up a good bit of efficiency with respect to least squares. ‖

So we are in the same situation that we encountered in Section 10.2.1. The LAD alternative to least squares seems to lose too much in efficiency if the errors are truly normal. The compromise, once again, is an M-estimator. We can construct one by minimizing a function analogous to (10.2.2), which would be $\sum_i \rho_i(\alpha, \beta)$, where

$$\text{(12.4.6)} \qquad \rho_i(\alpha,\beta) = \begin{cases} \frac{1}{2}(y_i - \alpha - \beta x_i)^2 & \text{if } |y_i - \alpha - \beta x_i| \le k \\ k|y_i - \alpha - \beta x_i| - \frac{1}{2}k^2 & \text{if } |y_i - \alpha - \beta x_i| > k, \end{cases}$$

where k is a tuning parameter.

Example 12.4.4 (Regression M-estimator) Using the function (12.4.6) with $k = 1.5\sigma$, we fit the M-estimators of α and β for the data of Table 12.4.1. The results were

$$\text{M-estimate for original data: } y = 18.5 - .89x$$

$$\text{M-estimate for mistyped data: } y = 14.67 - .68x,$$

where we estimated σ by .23, the standard deviation of the residuals from the least squares fit.

Thus we see that the M-estimate is somewhat more resistant than the least squares line, behaving more like the LAD fit when there are outliers. ||

As in Section 10.2, we expect the ARE of the M-estimator to be better than that of the LAD. This is the case, however, the calculations become very involved (even more so than for the LAD) so we will not give the details here. Huber (1981, Chapter 7) gives a detailed treatment of M-estimator asymptotics; see also Portnoy (1987). We content ourselves with an evaluation of the M-estimator through a small simulation study, reproducing a table like Table 10.2.3.

Example 12.4.5 (Simulation of regression AREs) For the model $Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, 2, \ldots, 5$, we take the x_is to be $(-2, -1, 0, 1, 2)$, $\alpha = 0$, and $\beta = 1$. We generate ε_i from normals, double exponentials, and logistic distributions and calculate the variance of the least squares, LAD, and M-estimator. These are presented in the following table.

Regression M-estimator AREs, $k = 1.5$ *(based on* 10,000 *simulations)*

	Normal	Logistic	Double exponential
vs. least squares	0.98	1.03	1.07
vs. LAD	1.39	1.27	1.14

The M-estimator variance is similar to that of least squares for all three distributions and is a uniform improvement over LAD. The dominance of the M-estimator over LAD is more striking than that of the Huber estimator over the median (as given in Table 10.2.3). ||

12.5 Exercises

12.1 Verify the expressions in (12.2.7). (*Hint*: Use the Pythagorean Theorem.)

12.2 Show that the extrema of

$$f(b) = \frac{1}{1+b^2}\left[S_{yy} - 2bS_{xy} + b^2 S_{xx}\right]$$

are given by

$$b = \frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

Show that the "+" solution gives the minimum of $f(b)$.

12.3 In maximizing the likelihood (12.2.13), we first minimized, for each value of α, β, and σ_δ^2, the function

$$f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \left((x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2\right)$$

with respect to $\xi_1, \ldots, \xi_n$.

(a) Prove that this function is minimized at

$$\xi_i^* = \frac{x_i + \lambda\beta(y_i - \alpha)}{1 + \lambda\beta^2}.$$

(b) Show that the function

$$D_\lambda\left((x,y),(\xi,\alpha+\beta\xi)\right) = (x-\xi)^2 + \lambda(y - (\alpha+\beta\xi))^2$$

defines a *metric* between the points (x, y) and $(\xi, \alpha + \beta\xi)$. A *metric* is a distance measure, a function D that measures the distance between two points A and B. A metric satisfies the following four properties:

i. $D(A, A) = 0$.
ii. $D(A, B) > 0$ if $A \neq B$.
iii. $D(A, B) = D(B, A)$ (reflexive).
iv. $D(A, B) \leq D(A, C) + D(C, B)$ (triangle inequality).

12.4 Consider the MLE of the slope in the EIV model

$$\hat{\beta}(\lambda) = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}},$$

where $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$ is assumed known.

(a) Show that $\lim_{\lambda\to 0} \hat{\beta}(\lambda) = S_{xy}/S_{xx}$, the slope of the ordinary regression of y on x.

(b) Show that $\lim_{\lambda\to\infty} \hat{\beta}(\lambda) = S_{yy}/S_{xy}$, the reciprocal of the slope of the ordinary regression of x on y.

(c) Show that $\hat{\beta}(\lambda)$ is, in fact, monotone in λ and is increasing if $S_{xy} > 0$ and decreasing if $S_{xy} < 0$.

(d) Show that the orthogonal least squares line ($\lambda = 1$) is always between the lines given by the ordinary regressions of y on x and of x on y.

(e) The following data were collected in a study to examine the relationship between brain weight and body weight in a number of animal species.

Species	Body weight (kg) (x)	Brain weight (g) (y)
Arctic fox	3.385	44.50
Owl monkey	.480	15.50
Mountain beaver	1.350	8.10
Guinea pig	1.040	5.50
Chinchilla	.425	6.40
Ground squirrel	.101	4.00
Tree hyrax	2.000	12.30
Big brown bat	.023	.30

Calculate the MLE of the slope assuming the EIV model. Also, calculate the least squares slopes of the regressions of y on x and of x on y, and show how these quantities bound the MLE.

12.5 In the EIV functional relationship model, where $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$ is assumed known, show that the MLE of σ_δ^2 is given by (12.2.18).

12.6 Show that in the linear structural relationship model (12.2.6), if we integrate out ξ_i, the marginal distribution of (X_i, Y_i) is given by (12.2.19).

12.7 Consider a linear structural relationship model where we assume that ξ_i has an improper distribution, $\xi_i \sim \text{uniform}(-\infty, \infty)$.

(a) Show that for each i,

$$\int_{-\infty}^{\infty} \frac{1}{(2\pi)} \frac{1}{\sigma_\delta \sigma_\epsilon} \exp\left[-\left(\frac{(x_i - \xi_i)^2}{2\sigma_\delta^2}\right)\right] \exp\left[-\left(\frac{(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\epsilon^2}\right)\right] d\xi_i$$
$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\beta^2\sigma_\delta^2 + \sigma_\epsilon^2}} \exp\left[-\frac{1}{2}\frac{(y_i - (\alpha + \beta x_i))^2}{\beta^2\sigma_\delta^2 + \sigma_\epsilon^2}\right].$$

(Completing the square in the exponential makes the integration easy.)

(b) The result of the integration in part (a) looks like a pdf, and if we consider it a pdf of Y conditional on X, then we seem to have a linear relationship between X and Y. Thus, it is sometimes said that this "limiting case" of the structural relationship leads to simple linear regression and ordinary least squares. Explain why this interpretation of the above function is wrong.

12.8 Verify the nonidentifiability problems in the structural relationship model in the following ways.

(a) Produce two different sets of parameters that give the same marginal distribution to (X_i, Y_i).

(b) Show that there are at least two distinct parameter vectors that yield the same solution to the equations given in (12.2.20).

12.9 In the structural relationship model, the solution to the equations in (12.2.20) implies a restriction on $\hat{\beta}$, the same restriction seen in the functional relationship case (see Exercise 12.4).

(a) Show that in (12.2.20), the MLE of σ_δ^2 is nonnegative only if $S_{xx} \geq (1/\hat{\beta})S_{xy}$. Also, the MLE of σ_ϵ^2 is nonnegative only if $S_{yy} \geq \hat{\beta}S_{xy}$.

(b) Show that the restrictions in part (a), together with the rest of the equations in (12.2.20), imply that

$$\frac{|S_{xy}|}{S_{xx}} \le |\hat{\beta}| \le \frac{S_{yy}}{|S_{xy}|}.$$

12.10 (a) Derive the MLEs for $(\alpha, \beta, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$ in the structural relationship model by solving the equations (12.2.20) under the assumption that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$.

(b) Calculate the MLEs for $(\alpha, \beta, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$ for the data of Exercise 12.4 by assuming the structural relationship model holds and that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$.

(c) Verify the relationship between variance estimates in the functional and structural relationship models. In particular, show that

$$\widehat{\text{Var}}\, X_i(\text{structural}) = 2\widehat{\text{Var}}\, X_i(\text{functional}).$$

That is, verify

$$\left(S_{xx} - \frac{S_{xy}}{\hat{\beta}}\right) = \frac{\lambda}{1+\lambda\hat{\beta}^2}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2.$$

(d) Verify the following equality, which is implicit in the MLE variance estimates given in (12.2.21). Show that

$$S_{xx} - \frac{S_{xy}}{\hat{\beta}} = \lambda(S_{yy} - \hat{\beta}S_{xy}).$$

12.11 (a) Show that for random variables X and Y and constants a, b, c, d,

$$\text{Cov}(aY + bX, cY + dX) = ac\text{Var}\, Y + (bc + ad)\text{Cov}(X, Y) + bd\text{Var}\, X.$$

(b) Use the result in part (a) to verify that in the structural relationship model with $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$,

$$\text{Cov}(\beta\lambda Y_i + X_i, Y_i - \beta X_i) = 0,$$

the identity on which the Creasy–Williams confidence set is based.

(c) Use the results of part (b) to show that

$$\frac{\sqrt{(n-2)}\, r_\lambda(\beta)}{\sqrt{1 - r_\lambda^2(\beta)}} \sim t_{n-2}$$

for any value of β, where $r_\lambda(\beta)$ is given in (12.2.23). Also, show that the confidence set defined in (12.2.24) has constant coverage probability equal to $1 - \alpha$.

12.12 Verify the following facts about $\hat{\beta}$ (the MLE of β when we assume $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$), $r_\lambda(\beta)$ of (12.2.23), and $C_\lambda(\hat{\beta})$, the Creasy–Williams confidence set of (12.2.24).

(a) $\hat{\beta}$ and $-1/(\lambda\hat{\beta})$ are the two roots of the quadratic equation defining the zeros of the first derivative of the likelihood function (12.2.14).

(b) $r_\lambda(\beta) = -r_\lambda(-1/(\lambda\beta))$ for every β.

(c) If $\beta \in C_\lambda(\hat{\beta})$, then $-1/(\lambda\beta) \in C_\lambda(\hat{\beta})$.

12.13 There is an interesting connection between the Creasy–Williams confidence set of (12.2.24) and the interval $C_G(\hat{\beta})$ of (12.2.22).

(a) Show that

$$C_G(\hat{\beta}) = \left\{\beta : \frac{(\beta - \hat{\beta})^2}{\hat{\sigma}_{\beta}^2/(n-2)} \leq F_{1,n-2,\alpha}\right\},$$

where $\hat{\beta}$ is the MLE of β and $\hat{\sigma}_{\beta}^2$ is the previously defined consistent estimator of σ_{β}^2.

(b) Show that the Creasy–Williams set can be written in the form

$$\left\{\beta : \frac{(\beta - \hat{\beta})^2}{\hat{\sigma}_{\beta}^2/(n-2)} \left[\frac{(1+\lambda\beta\hat{\beta})^2}{(1+\lambda\beta^2)^2}\right] \leq F_{1,n-2,\alpha}\right\}.$$

Hence $C_G(\hat{\beta})$ can be derived by replacing the term in square brackets with 1, its probability limit. (In deriving this representation, the fact that $\hat{\beta}$ and $-1/(\lambda\hat{\beta})$ are roots of the numerator of $r_{\lambda}(\beta)$ is of great help. In particular, the fact that

$$\frac{r_{\lambda}^2(\beta)}{1 - r_{\lambda}^2(\beta)} = \frac{\lambda^2 S_{xy}^2 (\beta - \hat{\beta})^2 (\beta + (1/\lambda\hat{\beta}))^2}{(1+\lambda\beta^2)^2 (S_{xx}S_{yy} - S_{xy}^2)}$$

is straightforward to establish.)

12.14 Graph the logistic regression function $\pi(x)$ from (12.3.2) for these three cases: $\alpha = \beta = 1$, $\alpha = \beta = 2$, and $\alpha = \beta = 3$.

12.15 For the logistic regression function in (12.3.2), verify these relationships.

(a) $\pi(-\alpha/\beta) = 1/2$

(b) $\pi((-\alpha/\beta) + c) = 1 - \pi((-\alpha/\beta) - c)$ for any c

(c) (12.3.3) for $d\pi(x)/dx$

(d) (12.3.4) about the odds ratio

(e) (12.3.5) about the multiplicative change in odds

(f) (12.3.6) and (12.3.8) regarding the likelihood equations for a Bernoulli GLM

(g) For logistic regression, $f_i/(F_i(1 - F_i)) = 1$ in (12.3.7) and (12.3.8)

12.16 Consider this logistic regression data. Only two values, $x = 0$ and 1, are observed. For $x = 0$ there are 10 successes in 10 trials. For $x = 1$ there are 5 successes in 10 trials. Show that the logistic regression MLEs $\hat{\alpha}$ and $\hat{\beta}$ do not exist for these data by verifying the following.

(a) The MLEs for $\pi(0)$ and $\pi(1)$, not restricted by (12.3.2), are given by $\hat{\pi}(0) = 1$ and $\hat{\pi}(1) = .5$.

(b) The overall maximum of the likelihood function given by the estimates in part (a) can not be achieved at any finite values of the logistic regression parameters α and β, but can be achieved in the limit as $\beta \to -\infty$ and $\alpha = -\beta$.

12.17 In *probit regression*, the link function is the standard normal cdf $\Phi(x) = P(Z \leq x)$, where $Z \sim \text{n}(0,1)$. Thus, in this model we observe $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, where $Y_i \sim \text{Bernoulli}(\pi_i)$ and $\pi_i = \Phi(\alpha + \beta x_i)$.

(a) Write out the likelihood function and show how to solve for the MLEs of α and β.

(b) Fit the probit model to the data of Table 12.3.1. Comment on any differences from the logistic fit.

12.18 Brown and Rothery (1993, Chapter 4) discuss a generalization of the linear logistic model to the quadratic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i + \gamma x_i^2.$$

(a) Write out the likelihood function and show how to solve for the MLEs of α, β, and γ.

(b) Using the log LRT, show how to test the hypothesis $H_0 : \gamma = 0$, that is, that the model is really linear logistic.

(c) Fit the quadratic logistic model to the data in the table on survival of sparrowhawks of different ages.

Age	1	2	3	4	5	6	7	8	9
No. of birds	77	149	182	118	78	46	27	10	4
No. surviving	35	89	130	79	52	28	14	3	1

(d) Decide whether the better model for the sparrowhawks is linear or quadratic, that is, test $H_0 : \gamma = 0$.

12.19 For the logistic regression model:

(a) Show that $\left(\sum_{j=1}^J Y_j^*, \sum_{j=1}^J Y_j^* x_j\right)$ is a sufficient statistic for (α, β).

(b) Verify the formula for the logistic regression information matrix in (12.3.10).

12.20 Consider a logistic regression model and assume $\beta = 0$.

(a) If $0 < \sum_{i=1}^n y_i < n$, show that the MLE of $\pi(x)$ (which does not depend on x in this case) is $\hat{\pi} = \sum_{i=1}^n y_i/n$.

(b) If $0 < \sum_{i=1}^n y_i < n$, show that the MLE of α is $\hat{\alpha}_0 = \log\left((\sum_{i=1}^n y_i)/(n - \sum_{i=1}^n y_i)\right)$.

(c) Show that if $\sum_{i=1}^n y_i = 0$ or n, $\hat{\alpha}_0$ does not exist, but the LRT statistic for testing $H_0 : \beta = 0$ is still well defined.

12.21 Let $Y \sim$ binomial(n, π), and let $\hat{\pi} = Y/n$ denote the MLE of π. Let $W = \log(\hat{\pi}/(1-\hat{\pi}))$ denote the sample logit, the MLE of $\log(\pi/(1-\pi))$. Use the Delta Method to show that $1/(n\hat{\pi}(1-\hat{\pi}))$ is a reasonable estimate of Var W.

12.22 In Example 12.4.1 we examined how small perturbations affected the least squares estimate of slope. Perform the analogous calculation and assess the robustness (to small perturbations) of the least squares estimate of intercept.

12.23 In Example 12.4.1, in contrast to Example 10.2.1, when we introduced the contaminated distribution for ε_i, we did not introduce a bias. Show that if we had, it would not have mattered. That is, if we assume

$$(\mathrm{E}\,\varepsilon_i, \mathrm{Var}\,\varepsilon_i) = \begin{cases} (0, \sigma^2) & \text{with probability } 1-\delta \\ (\mu, \tau^2) & \text{with probability } \delta, \end{cases}$$

then:

(a) the least squares estimator b would still be an unbiased estimator of β.

(b) the least squares estimator a has expectation $\alpha + \delta\mu$, so the model may just as well be assumed to be $Y_i = \alpha + \delta\mu + \beta x_i + \varepsilon_i$.

12.24 For the model $Y_i = \beta x_i + \varepsilon_i$, show that the LAD estimator is given by $t_{(k^*+1)}$, where $t_i = y_i/x_i$, $t_{(1)} \le \cdots \le t_{(n)}$ and, if $x_{(i)}$ is the x value paired with $t_{(i)}$, k^* satisfies $\sum_{i=1}^{k^*} |x_{(i)}| \le \sum_{i=k^*+1}^n |x_{(i)}|$ and $\sum_{i=1}^{k^*+1} |x_{(i)}| > \sum_{i=k^*+2}^n |x_{(i)}|$.

12.25 A problem with the LAD regression line is that it is not always uniquely defined.

(a) Show that, for a data set with three observations, $(x_1, y_1), (x_1, y_2)$, and (x_3, y_3) (note the first two xs are the same), any line that goes through (x_3, y_3) and lies between (x_1, y_1) and (x_1, y_2) is a least absolute deviation line.

(b) For three individuals, measurements are taken on heart rate (x, in beats per minute) and oxygen consumption (y, in ml/kg). The (x, y) pairs are (127, 14.4), (127, 11.9), and (136, 17.9). Calculate the slope and intercept of the least squares line and the range of the least absolute deviation lines.

There seems to be some disagreement over the value of the least absolute deviation line. It is certainly more robust than least squares but can be very difficult to compute (but see Portnoy and Koenker 1997 for an efficient computing algorithm). It also seems that Ellis (1998) questions its robustness, and in a discussion Portnoy and Mizera (1998) question Ellis.

Exercises 12.26–12.28 will look at some of the details of Example 12.4.3.

12.26 (a) Throughout Example 12.4.3 we assumed that $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \to \sigma_x^2 < \infty$. Show that this condition is satisfied by (i) $x_i = 1$ (the case of the ordinary median) and (ii) $|x_i| \le 1$ (the case of bounded x_i).

(b) Show that, under the conditions on x_i in part (a), $\frac{1}{n}\sum_{i=1}^{n} \psi(y_i - \hat{\beta}_L x_i) \to 0$ in probability.

12.27 (a) Verify that $\frac{-1}{\sqrt{n}}\sum_{i=1}^{n} \psi(Y_i - \hat{\beta} x_i) \to \mathrm{n}\left(0, \sigma_x^2\right)$.

(b) Verify that $\frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\beta_0} \mathrm{E}_\beta[\psi(Y_i - \beta_0 x_i)]\big|_{\beta_0=\beta} = 2f(0)\frac{1}{n}\sum_{i=1}^{n} x_i^2$, and, with part (a), conclude that $\sqrt{n}(\hat{\beta}_L - \beta) \to \mathrm{n}\left(0, \frac{1}{4f(0)^2\sigma_x^2}\right)$.

12.28 Show that the least squares estimator is given by $\hat{\beta} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$, and $\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{n}(0, 1/\sigma_x^2)$.

12.29 Using a Taylor series argument as in Example 12.4.3, derive the asymptotic distribution of the median in iid sampling.

12.30 For the data of Table 12.4.1, use the parametric bootstrap to assess the standard error from the LAD and M-estimator fit. In particular:

(a) Fit the line $y = \alpha + \beta x$ to get estimates $\tilde{\alpha}$ and $\tilde{\beta}$.

(b) Calculate the residual mean squared error $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} [y_i - (\tilde{\alpha} + \tilde{\beta} x_i)]^2$.

(c) Generate new residuals from $\mathrm{n}(0, \hat{\sigma}^2)$ and re-estimate α and β.

(d) Do part (c) B times and calculate the standard deviation of $\tilde{\alpha}$ and $\tilde{\beta}$.

(e) Repeat parts (a)–(d) using both the double exponential and Laplace distributions for the errors. Compare your answers to the normal.

12.31 For the data of Table 12.4.1, we could also use the nonparametric bootstrap to assess the standard error from the LAD and M-estimator fit.

(a) Fit the line $y = \alpha + \beta x$ to get estimates $\tilde{\alpha}$ and $\tilde{\beta}$.

(b) Generate new residuals by resampling from the fitted residuals and re-estimate α and β.

(c) Do part (c) B times and calculate the standard deviation of $\tilde{\alpha}$ and $\tilde{\beta}$.

12.6 Miscellanea

12.6.1 The Meaning of Functional and Structural

The names *functional* and *structural* are, in themselves, a prime source of confusion in the EIV model. Kendall and Stuart (1979, Chapter 29) give a detailed discussion of these concepts, distinguishing among relationships between *mathematical* (nonrandom) variables and relationships between random variables. One way to see the relationship is to write the models in a hierarchy in which the structural relationship model is obtained by putting a distribution on the parameters of the functional model:

$$\text{Functional relationship model} \quad \left\{ \begin{array}{ll} \mathrm{E}(Y_i|\xi_i) = \alpha + \beta\xi_i + \epsilon_i & \epsilon_i \sim \mathrm{n}(0, \sigma_\epsilon^2) \\ \mathrm{E}(X_i|\xi_i) = \xi_i + \delta_i & \delta_i \sim \mathrm{n}(0, \sigma_\delta^2) \\ & \xi_i \sim \mathrm{n}(\xi, \sigma_\xi^2) \end{array} \right\} \quad \text{Structural relationship model}$$

The difference in the words may be understood through the following distinction, not a universally accepted one. In the subject of calculus, for example, we often see the equation $y = f(x)$, an equation that describes a *functional* relationship, that is, a relationship that is *assumed to exist between variables.* Thus, from the idea that a functional relationship is an assumed relationship between two variables, the equation $\eta_i = \alpha + \beta\xi_i$, where $\eta_i = \mathrm{E}(Y_i|\xi_i)$, is a functional (hypothesized) relationship in either the functional or structural relationship model.

On the other hand, a *structural* relationship is a relationship that *arises from the hypothesized structure of the problem.* Thus, in the structural relationship model, the relationship $\eta = \mathrm{E}Y_i = \alpha + \beta\xi = \alpha + \beta\mathrm{E}X_i$ can be deduced from the structure of the model; hence it is a structural relationship.

To make these ideas clearer, consider the case of simple linear regression where we assume that there is no error in the xs. The equation $\mathrm{E}(Y_i|x_i) = \alpha + \beta x_i$ is a functional relationship, a relationship that is hypothesized to exist between $\mathrm{E}(Y_i|x_i)$ and x_i. We can, however, also do simple linear regression under the assumption that the pair (X_i, Y_i) has a bivariate normal distribution and we operate conditional on the x_is. In this case, the relationship $\mathrm{E}(Y_i|x_i) = \alpha + \beta x_i$ follows from the structure of the hypothesized model and hence is a structural relationship.

Notice that, with these meanings, the distinction in terminology becomes a matter of taste. In any model we can deduce structural relationships from functional relationships, and vice versa. The important distinction is whether the nuisance parameters, the ξ_is, are integrated out before inference is done.

12.6.2 Consistency of Ordinary Least Squares in EIV Models

In general it is not a good idea to use the ordinary least squares estimator to estimate the slope in EIV regression. This is because the estimator is inconsistent.

Suppose that we assume a linear structural relationship (12.2.6). We have

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &\to \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var} X} \qquad \text{(as } n \to \infty\text{, using the WLLN)} \\ &= \frac{\beta\sigma_\xi^2}{\sigma_\delta^2 + \sigma_\xi^2}, \qquad \text{(from (12.2.19))}\end{aligned}$$

showing that $\hat{\beta}$ cannot be consistent. The same type of result can be obtained in the functional relationship case.

The behavior of $\hat{\beta}$ in EIV models is treated in Cochran (1968). Carroll, Gallo, and Gleser (1985) and Gleser, Carroll, and Gallo (1987) investigated conditions under which functions of the ordinary least squares estimator are consistent.

12.6.3 Instrumental Variables in EIV Models

The concept of instrumental variables goes back at least to Wald (1940), who constructed a consistent estimator of the slope with their help. To see what an instrumental variable is, write the EIV model in the form

$$\begin{aligned}Y_i &= \alpha + \beta\xi_i + \epsilon_i, \\ X_i &= \xi_i + \delta_i,\end{aligned}$$

and do some algebra to get

$$Y_i = \alpha + \beta X_i + [\epsilon_i - \beta\delta_i].$$

An *instrumental variable*, Z_i, is a random variable that predicts X_i well but is uncorrelated with $\nu_i = \epsilon_i - \beta\delta_i$. If such a variable can be identified, it can be used to improve predictions. In particular, it can be used to construct a consistent estimator of β.

Wald (1940) showed that, under fairly general conditions, the estimator

$$\hat{\beta}_{\mathrm{W}} = \frac{\bar{Y}_{(1)} - \bar{Y}_{(2)}}{\bar{X}_{(1)} - \bar{X}_{(2)}}$$

is a consistent estimator of β in identifiable models, where the subscripts refer to two groupings of the data. A variable Z_i, which takes on only two values to define

the grouping, is an instrumental variable. See Moran (1971) for a discussion of Wald's estimator.

Although instrumental variables can be of great help, there can be some problems associated with their use. For example, Feldstein (1974) showed instances where the use of instrumental variables can be detrimental. Moran (1971) discussed the difficulty of verifying the conditions needed to ensure consistency of a simple estimator like $\hat{\beta}_W$. Fuller (1987) provided an in-depth discussion of instrumental variables. A model proposed by Berkson (1950) exploited a correlation structure similar to that used with instrumental variables.

12.6.4 Logistic Likelihood Equations

In a logistic regression model, the likelihood equations are nonlinear in the parameters, and they must be solved numerically. The most commonly used method for solving these equations is the *Newton–Raphson method.* This method begins with an initial guess $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)})$ for the value of the MLEs. Then the log likelihood is approximated with a quadratic function, its second-order Taylor series about the point $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)})$. The next guess for the values of the MLEs, $(\hat{\alpha}^{(2)}, \hat{\beta}^{(2)})$, is the maximum of this quadratic function. Now another quadratic approximation is used, this one centered at $(\hat{\alpha}^{(2)}, \hat{\beta}^{(2)})$, and its maximum is the next guess for the values of the MLEs. The Taylor series approximations involve the first and second derivatives of the log likelihood. These are evaluated at the current guess $(\hat{\alpha}^{(t)}, \hat{\beta}^{(t)})$. These are the same second derivatives that appear in the information matrix in (12.3.10). Thus, a byproduct of this method of solving the likelihood equations is estimates of the variances and covariance of $\hat{\alpha}$ and $\hat{\beta}$. The convergence of the guesses $(\hat{\alpha}^{(t)}, \hat{\beta}^{(t)})$ to the MLEs $(\hat{\alpha}, \hat{\beta})$ is usually rapid for logistic regression models. It often takes only a few iterations to obtain satisfactory approximations.

The Newton–Raphson method is also called *iteratively reweighted least squares.* At each stage, the next guess for $(\hat{\alpha}, \hat{\beta})$ can be expressed as the solution of a least squares problem. But, this is a least squares problem in which the different terms in the sum of squares function are given different weights. In this case the weights are $n_j F_j^{(t)}(1 - F_j^{(t)})$, where $F_j^{(t)} = F(\hat{\alpha}^{(t)} + \hat{\beta}^{(t)} x_j)$ and F is the logistic cdf. This is the inverse of an approximation to the variance of the jth sample logit (see Exercise 12.21). The weights are recalculated at each stage, because the current guesses for the MLEs are used. That leads to the name "iteratively reweighted." Thus, the Newton–Raphson method is approximately the result of using the sample logits as the data and performing a weighted least squares to estimate the parameters.

12.6.5 More on Robust Regression

Robust alternatives to least squares have been an object of study for many years, and there is a vast body of literature addressing a variety of problems. In Section 12.4 we saw only a brief introduction to robust regression, but some of the advantages and difficulties should be apparent. There are many good books that treat robust regression in detail, including Hettmansperger and McKean (1996), Staudte and Sheather (1990), and Huber (1981). Some other topics that have received a lot of attention are discussed below.

Trimming and transforming

The work of Carroll, Ruppert, and co-authors has addressed many facets of robust regression. Ruppert and Carroll (1979) is a careful treatment of the asymptotics of *trimming* (the trimmed mean is discussed in Exercise 10.20), while Carroll and Ruppert (1985) examine alternatives to least squares when the errors are not identical. Later work looked at the advantages of transformations in regression (Carroll and Ruppert 1985, 1988).

Other robust alternatives

We looked at only the LAD estimator and one M-estimator. There are, of course, many other choices of robust estimators. One popular alternative is the least median of squares (LMS) estimator of Rousseeuw (1984); see also Rousseeuw and Leroy 1987). There are also *R-estimators*, rank-based regression estimators (see the review paper by Draper 1988). More recently, there has been work on *data depth* (Liu 1990, Liu and Singh 1992) with applications to regression in finding the *deepest line* (Rousseeuw and Hubert 1999).

Computing

From a practical view, computation of robust estimates can be quite challenging, as we are often faced with a difficult minimization problem. The review paper by Portnoy and Koenker (1997) is concerned with computation of LAD estimates. Hawkins (1993, 1994, 1995) has a number of algorithms for computing LMS and related estimates.