Chapter 10

Relationships Among Variables

CHAPTER OUTLINE

In this chapter, we are concerned with perhaps the most important application of statistical inference: the problem of analyzing whether or not a relationship exists among variables and what form the relationship takes. As a particular instance of this, recall the example and discussion in Section 5.1.

Many of the most important problems in science and society are concerned with relationships among variables. For example, what is the relationship between the amount of carbon dioxide placed into the atmosphere and global temperatures? What is the relationship between class size and scholastic achievement by students? What is the relationship between weight and carbohydrate intake in humans? What is the relationship between lifelength and the dosage of a certain drug for cancer patients? These are all examples of questions whose answers involve relationships among variables. We will see that statistics plays a key role in answering such questions.

In Section 10.1, we provide a precise definition of what it means for variables to be related, and we distinguish between two broad categories of relationship, namely, association and cause–effect. Also, we discuss some of the key ideas involved in collecting data when we want to determine whether a cause–effect relationship exists. In the remaining sections, we examine the various statistical methodologies that are used to analyze data when we are concerned with relationships.

We emphasize the use of frequentist methodologies in this chapter. We give some examples of the Bayesian approach, but there are some complexities involved with the distributional problems associated with Bayesian methods that are best avoided at this

stage. Sampling algorithms for the Bayesian approach have been developed, along the lines of those discussed in Chapter 7 (see also Chapter 11), but their full discussion would take us beyond the scope of this text. It is worth noting, however, that Bayesian analyses with diffuse priors will often yield results very similar to those obtained via the frequentist approach.

As discussed in Chapter 9, model checking is an important feature of any statistical analysis. For the models used in this chapter, a full discussion of the more rigorous P-value approach to model checking requires more development than we can accomplish in this text. As such, we emphasize the informal approach to model checking, via residual and probability plots. This should not be interpreted as a recommendation that these are the preferred methods for such models.

10.1 Related Variables

Consider a population Π with two variables $X, Y : \Pi \to R^1$ defined on it. What does it mean to say that the variables X and Y are related? Perhaps our first inclination is to say that there must be a formula relating the two variables, such as $Y = a + bX^2$ for some choice of constants a and b, or $Y = \exp(X)$, etc. But consider a population Π of humans and suppose $X(\pi)$ is the weight of π in kilograms and $Y(\pi)$ is the height of individual $\pi \in \Pi$ in centimeters. From our experience, we know that taller people tend to be heavier, so we believe that there is some kind of relationship between height and weight. We know, too, that there cannot be an exact formula that describes this relationship, because people with the same weight will often have different heights, and people with the same height will often have different weights.

10.1.1 The Definition of Relationship

If we think of all the people with a given weight x, then there will be a distribution of heights for all those individuals π that have weight x. We call this distribution the conditional distribution of Y, given that $X = x$.

We can now express what we mean by our intuitive idea that X and Y are related, for, as we change the value of the weight that we condition on, we expect the conditional distribution to change. In particular, as x increases, we expect that the location of the conditional distribution will increase, although other features of the distribution may change as well. For example, in Figure 10.1.1 we provide a possible plot of two approximating densities for the conditional distributions of Y given $X = 70$ kg and the conditional distribution of Y given $X = 90$ kg.

We see that the conditional distribution has shifted up when X goes from 70 to 90 kg but also that the shape of the distribution has changed somewhat as well. So we can say that a relationship definitely exists between X and Y, at least in this population. Notice that, as defined so far, X and Y are not random variables, but they become so when we randomly select π from the population. In that case, the conditional distributions referred to become the conditional probability distributions of the random variable Y, given that we observe $X = 70$ and $X = 90$, respectively.

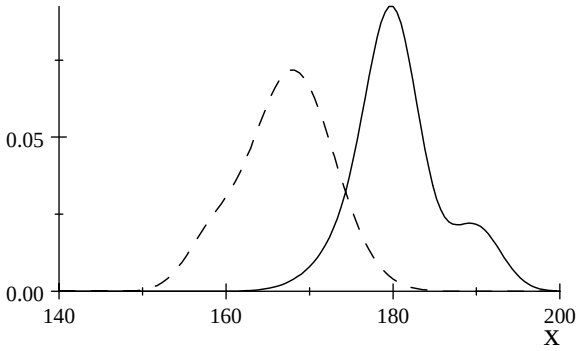

Figure 10.1.1: Plot of two approximating densities for the conditional distribution of Y given $X = 70$ kg (dashed line) and the conditional distribution of Y given $X = 90$ kg (solid line).

We will adopt the following definition to precisely specify what we mean when we say that variables are related.

Definition 10.1.1 Variables X and Y are *related variables* if there is any change in the conditional distribution of Y, given $X = x$, as x changes.

We could instead define what it means for variables to be *unrelated*. We say that variables X and Y are unrelated if they are independent. This is equivalent to Definition 10.1.1, because two variables are independent if and only if the conditional distribution of one given the other does not depend on the condition (Exercise 10.1.1).

There is an apparent asymmetry in Definition 10.1.1, because the definition considers only the conditional distribution of Y given X and not the conditional distribution of X given Y. But, if there is a change in the conditional distribution of Y given $X = x$, as we change x, then by the above comment, X and Y are not independent; thus there must be a change in the conditional distribution of X given $Y = y$, as we change y (also see Problem 10.1.22).

Notice that the definition is applicable no matter what kind of variables we are dealing with. So both could be quantitative variables, or both categorical variables, or one could be a quantitative variable while the other is a categorical variable.

Definition 10.1.1 says that X and Y are related if *any* change is observed in the conditional distribution. In reality, this would mean that there is practically always a relationship between variables X and Y. It seems likely that we will always detect some difference if we carry out a census and calculate all the relevant conditional distributions. This is where the idea of the *strength of a relationship among variables* becomes relevant, for if we see large changes in the conditional distributions, then we can say a strong relationship exists. If we see only very small changes, then we can say a very weak relationship exists that is perhaps of no practical importance.

The Role of Statistical Models

If a relationship exists between two variables, then its form is completely described by the set of conditional distributions of Y given X. Sometimes it may be necessary to describe the relationship using all these conditional distributions. In many problems, however, we look for a simpler presentation. In fact, we often assume a statistical model that prescribes a simple form for how the conditional distributions change as we change X.

Consider the following example.

EXAMPLE 10.1.1 *Simple Normal Linear Regression Model*
In Section 10.3.2, we will discuss the simple normal linear regression model, where the conditional distribution of quantitative variable Y, given the quantitative variable $X = x$, is assumed to be distributed

$$N(\beta_1 + \beta_2 x, \sigma^2),$$

where β_1, β_2, and σ^2 are unknown. For example, Y could be the blood pressure of an individual and X the amount of salt the person consumed each day.

In this case, the conditional distributions have constant shape and change, as x changes, only through the conditional mean. The mean moves along the line given by $\beta_1 + \beta_2 x$ for some intercept β_1 and slope β_2. If this model is correct, then the variables are unrelated if and only if $\beta_2 = 0$, as this is the only situation in which the conditional distributions can remain constant as we change x. ∎

Statistical models, like that described in Example 10.1.1, can be wrong. There is nothing requiring that two quantitative variables *must* be related in that way. For example, the conditional variance of Y can vary with x, and the very shape of the conditional distribution can vary with x, too. The model of Example 10.1.1 is an instance of a simplifying assumption that is appropriate in many practical contexts. However, methods such as those discussed in Chapter 9 must be employed to check model assumptions before accepting statistical inferences based on such a model. We will always consider model checking as part of our discussion of the various models used to examine the relationship among variables.

Response and Predictor Variables

Often, we think of Y as a dependent variable (depending on X) and of X as an independent variable (free to vary). Our goal, then, is to predict the value of Y given the value of X. In this situation, we call Y the *response variable* and X the *predictor variable*.

Sometimes, though, there is really nothing to distinguish the roles of X and Y. For example, suppose that X is the weight of an individual in kilograms and Y is the height in centimeters. We could then think of predicting weight from height or conversely. It is then immaterial which we choose to condition on.

In many applications, there is more than one response variable and more than one predictor variable X. We will not consider the situation in which we have more than one response variable, but we will consider the case in which $X = (X_1, \ldots, X_k)$ is

k-dimensional. Here, the various predictors that make up X could be all categorical, all quantitative, or some mixture of categorical and quantitative variables.

The definition of a relationship existing between response variable Y and the set of predictors $(X_1, \ldots, X_k)$ is exactly as in Definition 10.1.1. In particular, a relationship exists between Y and $(X_1, \ldots, X_k)$ if there is any change in the conditional distribution of Y given $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ when $(x_1, \ldots, x_k)$ is varied. If such a relationship exists, then the form of the relationship is specified by the full set of conditional distributions. Again, statistical models are often used where simplifying assumptions are made about the form of the relationship. Consider the following example.

EXAMPLE 10.1.2 *The Normal Linear Model with k Predictors*
In Section 10.3.4, we will discuss the normal multiple linear regression model. For this, the conditional distribution of quantitative variable Y, given that the quantitative predictors $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$, is assumed to be the

$$N(\beta_1 + \beta_2 x_1 + \cdots + \beta_{k+1} x_k, \sigma^2)$$

distribution, where $\beta_1, \ldots, \beta_{k+1}$ and σ^2 are unknown. For example, Y could be blood pressure, X_1 the amount of daily salt intake, X_2 the age of the individual, X_3 the weight of the individual, etc.

In this case, the conditional distributions have constant shape and change, as the values of the predictors $(x_1, \ldots, x_k)$ change only through the conditional mean, which changes according to the function $\beta_1 + \beta_2 x_1 + \cdots + \beta_{k+1} x_k$. Notice that, if this model is correct, then the variables are unrelated if and only if $\beta_2 = \cdots = \beta_{k+1} = 0$, as this is the only situation in which the conditional distributions can remain constant as we change $(x_1, \ldots, x_k)$. ∎

When we split a set of variables $Y, X_1, \ldots, X_k$ into response Y and predictors $(X_1, \ldots, X_k)$, we are implicitly saying that we are directly interested only in the conditional distributions of Y given $(X_1, \ldots, X_k)$. There may be relationships among the predictors $X_1, \ldots, X_k$, however, and these can be of interest.

For example, suppose we have two predictors X_1 and X_2, and the conditional distribution of X_1 given X_2 is virtually degenerate at a value $a + cX_2$ for some constants a and c. Then it is not a good idea to include both X_1 and X_2 in a model, such as that discussed in Example 10.1.2, as this can make the analysis very sensitive to small changes in the data. This is known as the problem of *multicollinearity*. The effect of multicollinearity, and how to avoid it, will not be discussed any further in this text. This is, however, a topic of considerable practical importance.

Regression Models

Suppose that the response Y is quantitative and we have k predictors $(X_1, \ldots, X_k)$. One of the most important simplifying assumptions used in practice is the *regression assumption*, namely, we assume that, as we change $(X_1, \ldots, X_k)$, the only thing that can possibly change about the conditional distribution of Y given $(X_1, \ldots, X_k)$, is the conditional mean $E(Y \mid X_1, \ldots, X_k)$. The importance of this assumption is that, to analyze the relationship between Y and $(X_1, \ldots, X_k)$, we now need only consider how

$E(Y \mid X_1, \ldots, X_k)$ changes as $(X_1, \ldots, X_k)$ changes. Indeed, if $E(Y \mid X_1, \ldots, X_k)$ does not change as $(X_1, \ldots, X_k)$ changes, then there is no relationship between Y and the predictors. Of course, this kind of an analysis is dependent on the regression assumption holding, and the methods of Section 9.1 must be used to check this. *Regression models* — namely, statistical models where we make the regression assumption — are among the most important statistical models used in practice. Sections 10.3 and 10.4 discuss several instances of regression models.

Regression models are often presented in the form

$$Y = E(Y \mid X_1, \ldots, X_k) + Z, \tag{10.1.1}$$

where $Z = Y - E(Y \mid X_1, \ldots, X_k)$ is known as the *error term.* We see immediately that, if the regression assumption applies, then the conditional distribution of Z given $(X_1, \ldots, X_k)$ is fixed as we change $(X_1, \ldots, X_k)$ and, conversely, if the conditional distribution of Z given $(X_1, \ldots, X_k)$ is fixed as we change $(X_1, \ldots, X_k)$, then the regression assumption holds. So when the regression assumption applies, (10.1.1) provides a decomposition of Y into two parts: (1) a part possibly dependent on $(X_1, \ldots, X_k)$, namely, $E(Y \mid X_1, \ldots, X_k)$, and (2) a part that is always independent of $(X_1, \ldots, X_k)$, namely, the error Z. Note that Examples 10.1.1 and 10.1.2 can be written in the form (10.1.1), where $Z \sim N(0, \sigma^2)$.

10.1.2 Cause–Effect Relationships and Experiments

Suppose now that we have variables X and Y defined on a population Π and have concluded that a relationship exists according to Definition 10.1.1. This may be based on having conducted a full census of Π, or, more typically, we will have drawn a simple random sample from Π and then used the methods of the remaining sections of this chapter to conclude that such a relationship exists. If Y is playing the role of the response and if X is the predictor, then we often want to be able to assert that changes in X are *causing* the observed changes in the conditional distributions of Y. Of course, if there are no changes in the conditional distributions, then there is no relationship between X and Y and hence no *cause–effect relationship,* either.

For example, suppose that the amount of carbon dioxide gas being released in the atmosphere is increasing, and we observe that mean global temperatures are rising. If we have reason to believe that the amount of carbon dioxide released can have an effect on temperature, then perhaps it is sensible to believe that the increase in carbon dioxide emissions is causing the observed increase in mean global temperatures. As another example, for many years it has been observed that smokers suffer from respiratory diseases much more frequently than do nonsmokers. It seems reasonable, then, to conclude that smoking causes an increased risk for respiratory disease. On the other hand, suppose we consider the relationship between weight and height. It seems clear that a relationship exists, but it does not make any sense to say that changes in one of the variables is causing the changes in the conditional distributions of the other.

Confounding Variables

When can we say that an observed relationship between X and Y is a cause–effect relationship? If a relationship exists between X and Y, then we know that there are at least two values x_1 and x_2 such that $f_Y(\cdot \mid X = x_1) \neq f_Y(\cdot \mid X = x_2)$, i.e., these two conditional distributions are not equal. If we wish to say that this difference is caused by the change in X, then we have to know categorically that there is no other variable Z defined on Π that *confounds* with X. The following example illustrates the idea of two variables confounding.

EXAMPLE 10.1.3
Suppose that Π is a population of students such that most females hold a part-time job and most males do not. A researcher is interested in the distribution of grades, as measured by grade point average (GPA), and is looking to see if there is a relationship between GPA and gender. On the basis of the data collected, the researcher observes a difference in the conditional distribution of GPA given gender and concludes that a relationship exists between these variables. It seems clear, however, that an assertion of a cause–effect relationship existing between GPA and gender is not warranted, as the difference in the conditional distributions could also be attributed to the difference in part-time work status rather than gender. In this example, part-time work status and gender are confounded. ∎

A more careful analysis might rescue the situation described in Example 10.1.3, for if X and Z denote the confounding variables, then we could collect data on Z as well and examine the conditional distributions $f_Y(\cdot \mid X = x, Z = z)$. In Example 10.1.3, these will be the conditional distributions of GPA, given gender and part-time work status. If these conditional distributions change as we change x, for some fixed value of z, then we could assert that a cause–effect relationship exists between X and Y *provided* there are no further confounding variables. Of course, there are probably still more confounding variables, and we really should be conditioning on all of them. This brings up the point that, in any practical application, we almost certainly will never even know all the potential confounding variables.

Controlling Predictor Variable Assignments

Fortunately, there is sometimes a way around the difficulties raised by confounding variables. Suppose we can *control* the value of the variable X for any $\pi \in \Pi$, i.e., we can *assign* the value x to π so that $X(\pi) = x$ for any of the possible values of x. In Example 10.1.3, this would mean that we could assign a part-time work status to any student in the population. Now consider the following idealized situation. Imagine assigning every element $\pi \in \Pi$ the value $X(\pi) = x_1$ and then carrying out a census to obtain the conditional distribution $f_Y(\cdot \mid X = x_1)$. Now imagine assigning every $\pi \in \Pi$ the value $X(\pi) = x_2$ and then carrying out a census to obtain the conditional distribution $f_Y(\cdot \mid X = x_2)$. If there is any difference in $f_Y(\cdot \mid X = x_1)$ and $f_Y(\cdot \mid X = x_2)$, then the only possible reason is that the value of X differs. Therefore, if $f_Y(\cdot \mid X = x_1) \neq f_Y(\cdot \mid X = x_2)$, we can assert that a cause–effect relationship exists.

A difficulty with the above argument is that typically we can never exactly determine $f_Y(\cdot \mid X = x_1)$ and $f_Y(\cdot \mid X = x_2)$. But in fact, we may be able to sample from them, then the methods of statistical inference become available to us to infer whether or not there is any difference. Suppose we take a random sample $\pi_1, \ldots, \pi_{n_1+n_2}$ from Π and randomly assign n_1 of these the value $X = x_1$, with the remaining π's assigned the value x_2. We obtain the Y values $y_{11}, \ldots, y_{1n_1}$ for those π's assigned the value x_1 and obtain the Y values $y_{21}, \ldots, y_{2n_2}$ for those π's assigned the value x_2. Then it is apparent that $y_{11}, \ldots, y_{1n_1}$ is a sample from $f_Y(\cdot \mid X = x_1)$ and $y_{21}, \ldots, y_{2n_2}$ is a sample from $f_Y(\cdot \mid X = x_2)$. In fact, provided that $n_1 + n_2$ is small relative to the population size, then we can consider these as i.i.d. samples from these conditional distributions.

So we see that in certain circumstances, it is possible to collect data in such a way that we can make inferences about whether or not a cause–effect relationship exists. We now specify the characteristics of the relevant data collection technique.

Conditions for Cause–Effect Relationships

First, if our inferences are to apply to a population Π, then we must have a random sample from that population. This is just the characteristic of what we called a sampling study in Section 5.4, and we must do this to avoid any selection effects. So if the purpose of a study is to examine the relationship between the duration of migraine headaches and the dosage of a certain drug, the investigator must have a random sample from the population of migraine headache sufferers.

Second, we must be able to *assign* any possible value of the predictor variable X to any selected π. If we cannot do this, or do not do this, then there may be hidden confounding variables (sometimes called *lurking variables*) that are influencing the conditional distributions of Y. So in a study of the effects of the dosage of a drug on migraine headaches, the investigator must be able to impose the dosage on each participant in the study.

Third, after deciding what values of X we will use in our study, we must randomly allocate these values to members of the sample. This is done to avoid the possibility of selection effects. So, after deciding what dosages to use in the study of the effects of the dosage of a drug on migraine headaches, and how many participants will receive each dosage, the investigator must randomly select the individuals who will receive each dosage. This will (hopefully) avoid selection effects, such as only the healthiest individuals getting the lowest dosage, etc.

When these requirements are met, we refer to the data collection process as an *experiment*. Statistical inference based on data collected via an experiment has the capability of inferring that cause–effect relationships exist, so this represents an important and powerful scientific tool.

A Hierarchy of Studies

Combining this discussion with Section 5.4, we see a hierarchy of data collection methods. Observational studies reside at the bottom of the hierarchy. Inferences drawn from observational studies must be taken with a degree of caution, for selection effects could mean that the results do not apply to the population intended, and the existence

of confounding variables means that we cannot make inferences about cause–effect relationships. For sampling studies, we know that any inferences drawn will be about the appropriate population; but the existence of confounding variables again causes difficulties for any statements about the existence of cause–effect relationships, e.g., just taking random samples of males and females from the population Π of Example 10.1.3 will not avoid the confounding variables. At the top of the hierarchy reside experiments.

It is probably apparent that it is often impossible to conduct an experiment. In Example 10.1.3, we cannot assign the value of gender, so nothing can be said about the existence of a cause–effect relationship between GPA and gender.

There are many notorious examples in which assertions are made about the existence of cause–effect relationships but for which no experiment is possible. For example, there have been a number of studies conducted where differences have been noted among the IQ distributions of various racial groups. It is impossible, however, to control the variable racial origin, so it is impossible to assert that the observed differences in the conditional distributions of IQ, given race, are caused by changes in race.

Another example concerns smoking and lung cancer in humans. It has been pointed out that it is impossible to conduct an experiment, as we cannot assign values of the predictor variable (perhaps different amounts of smoking) to humans at birth and then observe the response, namely, whether someone contracts lung cancer or not. This raises an important point. We do not simply reject the results of analyses based on observational studies or sampling studies because the data did not arise from an experiment. Rather, we treat these as evidence — potentially flawed evidence, but still evidence.

Think of eyewitness evidence in a court of law suggesting that a crime was committed by a certain individual. Eyewitness evidence may be unreliable, but if two or three unconnected eyewitnesses give similar reports, then our confidence grows in the reliability of the evidence. Similarly, if many observational and sampling studies seem to indicate that smoking leads to an increased risk for contracting lung cancer, then our confidence grows that a cause–effect relationship does indeed exist. Furthermore, if we can identify potentially confounding variables, then observational or sampling studies can be conducted taking these into account, increasing our confidence still more. Ultimately, we may not be able to definitively settle the issue via an experiment, but it is still possible to build overwhelming evidence that smoking and lung cancer do have a cause–effect relationship.

10.1.3 Design of Experiments

Suppose we have a response Y and a predictor X (sometimes called a *factor* in experimental contexts) defined on a population Π, and we want to collect data to determine whether a cause–effect relationship exists between them. Following the discussion in Section 10.1.1, we will conduct an experiment. There are now a number of decisions to be made, and our choices constitute what we call the *design* of the experiment.

For example, we are going to assign values of X to the sampled elements, now called *experimental units*, $\pi_1, \ldots, \pi_n$ from Π. Which of the possible values of X

should we use? When X can take only a small finite number of values, then it is natural to use these values. On the other hand, when the number of possible values of X is very large or even infinite, as with quantitative predictors, then we have to choose values of X to use in the experiment.

Suppose we have chosen the values $x_1, \ldots, x_k$ for X. We refer to $x_1, \ldots, x_k$ as the *levels* of X; any particular assignment x_i to a π_j in the sample will be called a *treatment*. Typically, we will choose the levels so that they span the possible range of X fairly uniformly. For example, if X is temperature in degrees Celsius, and we want to examine the relationship between Y and X for X in the range $[0, 100]$, then, using $k = 5$ levels, we might take $x_1 = 0$, $x_2 = 25$, $x_3 = 50$, $x_4 = 75$, and $x_5 = 100$.

Having chosen the levels of X, we then have to choose how many treatments of each level we are going to use in the experiment, i.e., decide how many response values n_i we are going to observe at level x_i for $i = 1, \ldots, k$.

In any experiment, we will have a finite amount of resources (money, time, etc.) at our disposal, which determines the sample size n from Π. The question then is how should we choose the n_i so that $n_1 + \cdots + n_k = n$? If we know nothing about the conditional distributions $f_Y(\cdot \mid X = x_i)$, then it makes sense to use *balance*, namely, choose $n_1 = \cdots = n_k$.

On the other hand, suppose we know that some of the $f_Y(\cdot \mid X = x_i)$ will exhibit greater variability than others. For example, we might measure variability by the variance of $f_Y(\cdot \mid X = x_i)$. Then it makes sense to allocate more treatments to the levels of X where the response is more variable. This is because it will take more observations to make accurate inferences about characteristics of such an $f_Y(\cdot \mid X = x_i)$ than for the less variable conditional distributions.

As discussed in Sections 6.3.4 and 6.3.5, we also want to choose the n_i so that any inferences we make have desired accuracy. Methods for choosing the sample sizes n_i, similar to those discussed in Chapter 7, have been developed for these more complicated designs, but we will not discuss these any further here.

Suppose, then, that we have determined $\{(x_1, n_1), \ldots, (x_k, n_k)\}$. We refer to this set of ordered pairs as the *experimental design*.

Consider some examples.

EXAMPLE 10.1.4

Suppose that Π is a population of students at a given university. The administration is concerned with determining the value of each student being assigned an academic advisor. The response variable Y will be a rating that a student assigns on a scale of 1 to 10 (completely dissatisfied to completely satisfied with their university experience) at the end of a given semester. We treat Y as a quantitative variable. A random sample of $n = 100$ students is selected from Π, and 50 of these are randomly selected to receive advisers while the remaining 50 are not assigned advisers.

Here, the predictor X is a categorical variable that indicates whether or not the student has an advisor. There are only $k = 2$ levels, and both are used in the experiment. If $x_1 = 0$ denotes no advisor and $x_2 = 1$ denotes having an advisor, then $n_1 = n_2 = 50$ and we have a balanced experiment. The experimental design is given by

$$\{(0, 50), (1, 50)\}.$$

At the end of the experiment, we want to use the data to make inferences about the conditional distributions $f_Y(\cdot \mid X = 0)$ and $f_Y(\cdot \mid X = 1)$ to determine whether a cause–effect relationship exists. The methods of Section 10.4 will be relevant for this. ∎

EXAMPLE 10.1.5
Suppose that Π is a population of dairy cows. A feed company is concerned with the relationship between weight gain, measured in kilograms, over a specific time period and the amount of a supplement, measured in grams/liter, of an additive put into the cows' feed. Here, the response Y is the weight gain — a quantitative variable. The predictor X is the concentration of the additive. Suppose X can plausibly range between 0 and 2, so it is also a quantitative variable.

The experimenter decides to use $k = 4$ levels with $x_1 = 0.00, x_2 = 0.66, x_3 = 1.32$, and $x_4 = 2.00$. Further, the sample sizes $n_1 = n_2 = n_3 = n_4 = 10$ were determined to be appropriate. So the balanced experimental design is given by

$$\{(0.00, 10)\,, (0.66, 10)\,, (1.32, 10)\,, (2.00, 10)\}.$$

At the end of the experiment, we want to make inferences about the conditional distributions $f_Y(\cdot \mid X = 0.00)$, $f_Y(\cdot \mid X = 0.66)$, $f_Y(\cdot \mid X = 1.32)$, and $f_Y(\cdot \mid X = 2.00)$. The methods of Section 10.3 are relevant for this. ∎

Control Treatment, the Placebo Effect, and Blinding

Notice that in Example 10.1.5, we included the level $X = 0$, which corresponds to no application of the additive. This is called a *control treatment*, as it gives a baseline against which we can assess the effect of the predictor. In many experiments, it is important to include a control treatment.

In medical experiments, there is often a *placebo effect* — that is, a disease sufferer given any treatment will often record an improvement in symptoms. The placebo effect is believed to be due to the fact that a sufferer will start to feel better simply because someone is paying attention to the condition. Accordingly, in any experiment to determine the efficacy of a drug in alleviating disease symptoms, it is important that a control treatment be used as well. For example, if we want to investigate whether or not a given drug alleviates migraine headaches, then among the dosages we select for the experiment, we should make sure that we include a pill containing none of the drug (the so-called *sugar pill*); that way we can assess the extent of the placebo effect. Of course, the recipients should not know whether they are receiving the sugar pill or the drug. This is called a *blind* experiment. If we also conceal the identity of the treatment from the experimenters, so as to avoid any biasing of the results on their part, then this is known as a *double-blind* experiment.

In Example 10.1.5, we assumed that it is possible to take a sample from the population of all dairy cows. Strictly speaking, this is necessary if we want to avoid selection effects and make sure that our inferences apply to the population of interest. In practice, however, taking a sample of experimental units from the full population of interest is often not feasible. For example, many medical experiments are conducted on ani-

mals, and these are definitely not random samples from the population of the particular animal in question, e.g., rats.

In such cases, however, we simply recognize the possibility that selection effects or lurking variables could render invalid the conclusions drawn from such analyses when they are to be applied to the population of interest. But we still regard the results as evidence concerning the phenomenon under study. It is the job of the experimenter to come as close as possible to the idealized situation specified by a valid experiment; for example, randomization is still employed when assigning treatments to experimental units so that selection effects are avoided as much as possible.

Interactions

In the experiments we have discussed so far, there has been one predictor. In many practical contexts, there is more than one predictor. Suppose, then, that there are two predictors X and W and that we have decided on the levels $x_1, \ldots, x_k$ for X and the levels $w_1, \ldots, w_l$ for W. One possibility is to look at the conditional distributions $f_Y(\cdot \mid X = x_i)$ for $i = 1, \ldots, k$ and $f_Y(\cdot \mid W = w_j)$ for $j = 1, \ldots, l$ to determine whether X and W individually have a relationship with the response Y. Such an approach, however, ignores the effect of the two predictors together. In particular, the way the conditional distributions $f_Y(\cdot \mid X = x, W = w)$ change as we change x may depend on w; when this is the case, we say that there is an *interaction* between the predictors.

To investigate the possibility of an interaction existing between X and W, we must sample from each of the kl distributions $f_Y(\cdot \mid X = x_i, W = w_j)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. The experimental design then takes the form

$$\{(x_1, w_1, n_{11}), (x_2, w_1, n_{21}), \ldots, (x_k, w_l, n_{kl})\},$$

where n_{ij} gives the number of applications of the treatment (x_i, w_j). We say that the two predictors X and W are *completely crossed* in such a design because each value of X used in the experiment occurs with each value of W used in the experiment. Of course, we can extend this discussion to the case where there are more than two predictors. We will discuss in Section 10.4.3 how to analyze data to determine whether there are any interactions between predictors.

EXAMPLE 10.1.6

Suppose we have a population Π of students at a particular university and are investigating the relationship between the response Y, given by a student's grade in calculus, and the predictors W and X. The predictor W is the number of hours of academic advising given monthly to a student; it can take the values 0, 1, or 2. The predictor X indicates class size, where $X = 0$ indicates small class size and $X = 1$ indicates large class size. So we have a quantitative response Y, a quantitative predictor W taking three values, and a categorical predictor X taking two values. The crossed values of the predictors (W, X) are given by the set

$$\{(0, 0), (1, 0), (2, 0), (0, 1), (1, 1), (2, 1)\},$$

so there are six treatments. To conduct the experiment, the university then takes a random sample of $6n$ students and randomly assigns n students to each treatment. ∎

Sometimes we include additional predictors in an experimental design even when we are not primarily interested in their effects on the response Y. We do this because we *know* that such a variable has a relationship with Y. Including such predictors allows us to condition on their values and so investigate more precisely the relationship Y has with the remaining predictors. We refer to such a variable as a *blocking variable*.

EXAMPLE 10.1.7
Suppose the response variable Y is yield of wheat in bushels per acre, and the predictor variable X is an indicator variable for which of three types of wheat is being planted in an agricultural study. Each type of wheat is going to be planted on a plot of land, where all the plots are of the same size, but it is known that the plots used in the experiment will vary considerably with respect to their fertility. Note that such an experiment is another example of a situation in which it is impossible to randomly sample the experimental units (the plots) from the full population of experimental units.

Suppose the experimenter can group the available experimental units into plots of low fertility and high fertility. We call these two classes of fields *blocks*. Let W indicate the type of plot. So W is a categorical variable taking two values. It then seems clear that the conditional distributions $f_Y(\cdot \mid X = x, W = w)$ will be much less variable than the conditional distributions $f_Y(\cdot \mid X = x)$.

In this case, W is serving as a blocking variable. The experimental units in a particular block, the one of low fertility or the one of high fertility, are more homogeneous than the full set of plots, so variability will be reduced and inferences will be more accurate. ∎

Summary of Section 10.1

- We say two variables are related if the conditional distribution of one given the other changes at all, as we change the value of the conditioning variable.
- To conclude that a relationship between two variables is a cause–effect relationship, we must make sure that (through conditioning) we have taken account of all confounding variables.
- Statistics provides a practical way of avoiding the effects of confounding variables via conducting an experiment. For this, we must be able to assign the values of the predictor variable to experimental units sampled from the population of interest.
- The design of experiments is concerned with determining methods of collecting the data so that the analysis of the data will lead to accurate inferences concerning questions of interest.

EXERCISES

10.1.1 Prove that discrete random variables X and Y are unrelated if and only if X and Y are independent.

10.1.2 Suppose that two variables X and Y defined on a finite population Π are functionally related as $Y = g(X)$ for some unknown nonconstant function g. Explain how

this situation is covered by Definition 10.1.1, i.e., the definition will lead us to conclude that X and Y are related. What about the situation in which $g(x) = c$ for some value c for every x? (Hint: Use the relative frequency functions of the variables.)

10.1.3 Suppose that a census is conducted on a population and the joint distribution of (X, Y) is obtained as in the following table.

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.15	0.18	0.40
$X = 2$	0.12	0.09	0.06

Determine whether or not a relationship exists between Y and X.

10.1.4 Suppose that a census is conducted on a population and the joint distribution of (X, Y) is obtained as in the following table.

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	$1/6$	$1/6$	$1/3$
$X = 2$	$1/12$	$1/12$	$1/6$

Determine whether or not a relationship exists between Y and X.

10.1.5 Suppose that X is a random variable and $Y = X^2$. Determine whether or not X and Y are related. What happens when X has a degenerate distribution?

10.1.6 Suppose a researcher wants to investigate the relationship between birth weight and performance on a standardized test administered to children at two years of age. If a relationship is found, can this be claimed to be a cause–effect relationship? Explain why or why not?

10.1.7 Suppose a large study of all doctors in Canada was undertaken to determine the relationship between various lifestyle choices and lifelength. If the conditional distribution of lifelength given various smoking habits changes, then discuss what can be concluded from this study.

10.1.8 Suppose a teacher wanted to determine whether an open- or closed-book exam was a more appropriate way to test students on a particular topic. The response variable is the grade obtained on the exam out of 100. Discuss how the teacher could go about answering this question.

10.1.9 Suppose a researcher wanted to determine whether or not there is a cause–effect relationship between the type of political ad (negative or positive) seen by a voter from a particular population and the way the voter votes. Discuss your advice to the researcher about how best to conduct the study.

10.1.10 If two random variables have a nonzero correlation, are they necessarily related? Explain why or why not.

10.1.11 An experimenter wants to determine the relationship between weight change Y over a specified period and the use of a specially designed diet. The predictor variable X is a categorical variable indicating whether or not a person is on the diet. A total of 200 volunteers signed on for the study; a random selection of 100 of these were given the diet and the remaining 100 continued their usual diet.

(a) Record the experimental design.

(b) If the results of the study are to be applied to the population of all humans, what concerns do you have about how the study was conducted?

(c) It is felt that the amount of weight lost or gained also is dependent on the initial weight W of a participant. How would you propose that the experiment be altered to take this into account?

10.1.12 A study will be conducted, involving the population of people aged 15 to 19 in a particular country, to determine whether a relationship exists between the response Y (amount spent in dollars in a week on music downloads) and the predictors W (gender) and X (age in years).

(a) If observations are to be taken from every possible conditional distribution of Y given the two factors, then how many such conditional distributions are there?

(b) Identify the types of each variable involved in the study.

(c) Suppose there are enough funds available to monitor 2000 members of the population. How would you recommend that these resources be allocated among the various combinations of factors?

(d) If a relationship is found between the response and the predictors, can this be claimed to be a cause–effect relationship? Explain why or why not.

(e) Suppose that in addition, it was believed that family income would likely have an effect on Y and that families could be classified into low and high income. Indicate how you would modify the study to take this into account.

10.1.13 A random sample of 100 households, from the set of all households containing two or more members in a given geographical area, is selected and their television viewing habits are monitored for six months. A random selection of 50 of the households is sent a brochure each week advertising a certain program. The purpose of the study is to determine whether there is any relationship between exposure to the brochure and whether or not this program is watched.

(a) Identify suitable response and predictor variables.

(b) If a relationship is found, can this be claimed to be a cause–effect relationship? Explain why or why not.

10.1.14 Suppose we have a quantitative response variable Y and two categorical predictor variables W and X, both taking values in $\{0, 1\}$. Suppose the conditional distributions of Y are given by

$$
\begin{aligned}
Y \mid W = 0, X = 0 &\sim N(3, 5)\\
Y \mid W = 1, X = 0 &\sim N(3, 5)\\
Y \mid W = 0, X = 1 &\sim N(4, 5)\\
Y \mid W = 1, X = 1 &\sim N(4, 5).
\end{aligned}
$$

Does W have a relationship with Y? Does X have a relationship with Y? Explain your answers.

10.1.15 Suppose we have a quantitative response variable Y and two categorical predictor variables W and X, both taking values in $\{0, 1\}$. Suppose the conditional distri-

butions of Y are given by

$$Y \mid W = 0, X = 0 \sim N(2, 5)$$
$$Y \mid W = 1, X = 0 \sim N(3, 5)$$
$$Y \mid W = 0, X = 1 \sim N(4, 5)$$
$$Y \mid W = 1, X = 1 \sim N(4, 5).$$

Does W have a relationship with Y? Does X have a relationship with Y? Explain your answers.

10.1.16 Do the predictors interact in Exercise 10.1.14? Do the predictors interact in Exercise 10.1.15? Explain your answers.

10.1.17 Suppose we have variables X and Y defined on the population $\Pi = \{1, 2, \ldots, 10\}$, where $X(i) = 1$ when i is odd and $X(i) = 0$ when i is even, $Y(i) = 1$ when i is divisible by 3 and $Y(i) = 0$ otherwise.

(a) Determine the relative frequency function of X.

(b) Determine the relative frequency function of Y.

(c) Determine the joint relative frequency function of (X, Y).

(d) Determine all the conditional distributions of Y given X.

(e) Are X and Y related? Justify your answer.

10.1.18 A mathematical approach to examining the relationship between variables X and Y is to see whether there is a function g such that $Y = g(X)$. Explain why this approach does not work for many practical applications where we are examining the relationship between variables. Explain how statistics treats this problem.

10.1.19 Suppose a variable X takes the values 1 and 2 on a population and the conditional distributions of Y given X are $N(0, 5)$ when $X = 1$, and $N(0, 7)$ when $X = 2$. Determine whether X and Y are related and if so, describe their relationship.

10.1.20 A variable Y has conditional distribution given X specified by $N(1 + 2x, |x|)$ when $X = x$. Determine if X and Y are related and if so, describe what their relationship is.

10.1.21 Suppose that $X \sim \text{Uniform}[-1, 1]$ and $Y = X^2$. Determine the correlation between Y and X. Are X and Y related?

PROBLEMS

10.1.22 If there is more than one predictor involved in an experiment, do you think it is preferable for the predictors to interact or not? Explain your answer. Can the experimenter control whether or not predictors interact?

10.1.23 Prove directly, using Definition 10.1.1, that when X and Y are related variables defined on a finite population Π, then Y and X are also related.

10.1.24 Suppose that X, Y, Z are independent $N(0, 1)$ random variables and that $U = X + Z$, $V = Y + Z$. Determine whether or not the variables U and V are related. (Hint: Calculate $\text{Cov}(U, V)$.)

10.1.25 Suppose that $(X, Y, Z) \sim \text{Multinomial}(n, 1/3, 1/3, 1/3)$. Are X and Y related?

10.1.26 Suppose that $(X, Y) \sim$ Bivariate-Normal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Show that X and Y are unrelated if and only if $\text{Corr}(X, Y) = 0$.

10.1.27 Suppose that (X, Y, Z) have probability function $p_{X,Y,Z}$. If Y is related to X but not to Z, then prove that $p_{X,Y,Z}(x, y, z) = p_{Y|X}(y \mid x) p_{X|Z}(x \mid z) p_Z(z)$.

10.2 Categorical Response and Predictors

There are two possible situations when we have a single categorical response Y and a single categorical predictor X. The categorical predictor is either random or deterministic, depending on how we sample. We examine these two situations separately.

10.2.1 Random Predictor

We consider the situation in which X is categorical, taking values in $\{1, \ldots, a\}$, and Y is categorical, taking values in $\{1, \ldots, b\}$. If we take a sample $\pi_1, \ldots, \pi_n$ from the population, then the values $X(\pi_i) = x_i$ are random, as are the values $Y(\pi_i) = y_j$.

Suppose the sample size n is very small relative to the population size (so we can assume that i.i.d. sampling is applicable). Then, letting $\theta_{ij} = P(X = i, Y = j)$, we obtain the likelihood function (see Problem 10.2.15)

$$L(\theta_{11}, \ldots, \theta_{ab} \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{a} \prod_{j=1}^{b} \theta_{ij}^{f_{ij}}, \tag{10.2.1}$$

where f_{ij} is the number of sample values with $(X, Y) = (i, j)$. An easy computation (see Problem 10.2.16) shows that the MLE of $(\theta_{11}, \ldots, \theta_{kl})$ is given by $\hat{\theta}_{ij} = f_{ij}/n$ and that the standard error of this estimate (because the incidence of a sample member falling in the (i, j)-th cell is distributed Bernoulli(θ_{ij}) and using Example 6.3.2) is given by

$$\sqrt{\frac{\hat{\theta}_{ij}(1 - \hat{\theta}_{ij})}{n}}.$$

We are interested in whether or not there is a relationship between X and Y. To answer this, we look at the conditional distributions of Y given X. The conditional distributions of Y given X, using $\theta_{i\cdot} = \theta_{i1} + \cdots + \theta_{ib} = P(X = i)$, are given in the following table.

	$Y = 1$	$\cdots$	$Y = b$
$X = 1$	$\theta_{11}/\theta_{1\cdot}$	$\cdots$	$\theta_{1b}/\theta_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$
$X = a$	$\theta_{a1}/\theta_{a\cdot}$	$\cdots$	$\theta_{ab}/\theta_{a\cdot}$

Then estimating $\theta_{ij}/\theta_{i\cdot}$ by $\hat{\theta}_{ij}/\hat{\theta}_{i\cdot} = f_{ij}/f_{i\cdot}$, where $f_{i\cdot} = f_{i1}+\cdots+f_{ib}$, the estimated conditional distributions are as in the following table.

	$Y=1$	$\cdots$	$Y=b$
$X=1$	$f_{11}/f_{1\cdot}$	$\cdots$	$f_{1b}/f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$
$X=a$	$f_{a1}/f_{a\cdot}$	$\cdots$	$f_{ab}/f_{a\cdot}$

If we conclude that there is a relationship between X and Y, then we look at the table of estimated conditional distributions to determine the form of the relationship, i.e., how the conditional distributions change as we change the value of X we are conditioning on.

How, then, do we infer whether or not a relationship exists between X and Y? No relationship exists between Y and X if and only if the conditional distributions of Y given $X = x$ do not change with x. This is the case if and only if X and Y are independent, and this is true if and only if

$$\theta_{ij} = P(X=i, Y=j) = P(X=i)P(Y=j) = \theta_{i\cdot}\theta_{\cdot j},$$

for every i and j where $\theta_{\cdot j} = \theta_{1j} + \cdots + \theta_{aj} = P(Y=j)$. Therefore, to assess whether or not there is a relationship between X and Y, it is equivalent to assess the null hypothesis $H_0 : \theta_{ij} = \theta_{i\cdot}\theta_{\cdot j}$ for every i and j.

How should we assess whether or not the observed data are surprising when H_0 holds? The methods of Section 9.1.2, and in particular Theorem 9.1.2, can be applied here, as we have that

$$(F_{11}, F_{12}, \ldots, F_{ab}) \sim \text{ Multinomial}(n, \theta_{1\cdot}\theta_{\cdot 1}, \theta_{1\cdot}\theta_{\cdot 2}, \ldots, \theta_{a\cdot}\theta_{\cdot b})$$

when H_0 holds, where F_{ij} is the count in the (i, j)-th cell.

To apply Theorem 9.1.2, we need the MLE of the parameters of the model under H_0. The likelihood, when H_0 holds, is

$$L(\theta_{1\cdot}, \ldots, \theta_{a\cdot}, \theta_{\cdot 1}, \ldots, \theta_{\cdot b} \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{a}\prod_{j=1}^{b}\left(\theta_{i\cdot}\theta_{\cdot j}\right)^{f_{ij}}. \quad (10.2.2)$$

From this, we deduce (see Problem 10.2.17) that the MLE's of the $\theta_{i\cdot}$ and $\theta_{\cdot j}$ are given by $\hat{\theta}_{i\cdot} = f_{i\cdot}/n$ and $\hat{\theta}_{\cdot j} = f_{\cdot j}/n$. Therefore, the relevant chi-squared statistic is

$$X^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(f_{ij} - n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}\right)^2}{n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}}.$$

Under H_0, the parameter space has dimension $(a-1) + (b-1) = a + b - 2$, so we compare the observed value of X^2 with the $\chi^2((a-1)(b-1))$ distribution because $ab - 1 - a - b + 2 = (a-1)(b-1)$.

Consider an example.

EXAMPLE 10.2.1 *Piston Ring Data*
The following table gives the counts of piston ring failures, where variable Y is the compressor number and variable X is the leg position based on a sample of $n = 166$. These data were taken from *Statistical Methods in Research and Production*, by O. L. Davies (Hafner Publishers, New York, 1961).

Here, Y takes four values and X takes three values (N = North, C = Central, and S = South).

	$Y=1$	$Y=2$	$Y=3$	$Y=4$
$X=$ N	17	11	11	14
$X=$ C	17	9	8	7
$X=$ S	12	13	19	28

The question of interest is whether or not there is any relation between compressor and leg position. Because $f_{1\cdot} = 53$, $f_{2\cdot} = 41$, and $f_{3\cdot} = 72$, the conditional distributions of Y given X are estimated as in the rows of the following table.

	$Y=1$	$Y=2$	$Y=3$	$Y=4$
$X=$ N	$17/53 = 0.321$	$11/53 = 0.208$	$11/53 = 0.208$	$14/53 = 0.264$
$X=$ C	$17/41 = 0.415$	$9/41 = 0.222$	$8/41 = 0.195$	$7/41 = 0.171$
$X=$ S	$12/72 = 0.167$	$13/72 = 0.181$	$19/72 = 0.264$	$28/72 = 0.389$

Comparing the rows, it certainly looks like there is a difference in the conditional distributions, but we must assess whether or not the observed differences can be explained as due to sampling error. To see if the observed differences are real, we carry out the chi-squared test.

Under the null hypothesis of independence, the MLE's are given by

$$\hat{\theta}_{\cdot 1} = \tfrac{46}{166}, \quad \hat{\theta}_{\cdot 2} = \tfrac{33}{166}, \quad \hat{\theta}_{\cdot 3} = \tfrac{38}{166}, \quad \hat{\theta}_{\cdot 4} = \tfrac{49}{166}$$

for the Y probabilities, and by

$$\hat{\theta}_{1\cdot} = \tfrac{53}{166}, \quad \hat{\theta}_{2\cdot} = \tfrac{41}{166}, \quad \hat{\theta}_{3\cdot} = \tfrac{72}{166}$$

for the X probabilities. Then the estimated expected counts $n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}$ are given by the following table.

	$Y=1$	$Y=2$	$Y=3$	$Y=4$
$X=$ N	14.6867	10.5361	12.1325	15.6446
$X=$ C	11.3614	8.1506	9.3855	12.1024
$X=$ S	19.9518	14.3133	16.4819	21.2530

The standardized residuals (using (9.1.6))

$$\frac{f_{ij} - n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}}{\sqrt{n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}(1 - \hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j})}}$$

are as in the following table.

	$Y=1$	$Y=2$	$Y=3$	$Y=4$
$X=\text{N}$	0.6322	0.1477	−0.3377	−0.4369
$X=\text{C}$	1.7332	0.3051	−0.4656	−1.5233
$X=\text{S}$	−1.8979	−0.3631	0.6536	1.5673

All of the standardized residuals seem reasonable, and we have that $X^2 = 11.7223$ with $P(\chi^2(6) > 11.7223) = 0.0685$, which is not unreasonably small.

So, while there may be some indication that the null hypothesis of no relationship is false, this evidence is not overwhelming. Accordingly, in this case, we may assume that Y and X are independent and use the estimates of cell probabilities obtained under this assumption. ∎

We must also be concerned with model checking, i.e., is the model that we have assumed for the data $(x_1, y_1), \ldots, (x_n, y_n)$ correct? If these observations are i.i.d., then indeed the model is correct, as that is all that is being effectively assumed. So we need to check that the observations are a plausible i.i.d. sample. Because the minimal sufficient statistic is given by $(f_{11}, \ldots, f_{ab})$, such a test could be based on the conditional distribution of the sample $(x_1, y_1), \ldots, (x_n, y_n)$ given $(f_{11}, \ldots, f_{ab})$. The distribution theory for such tests is computationally difficult to implement, however, and we do not pursue this topic further in this text.

10.2.2 Deterministic Predictor

Consider again the situation in which X is categorical, taking values in $\{1, \ldots, a\}$, and Y is categorical, taking values in $\{1, \ldots, b\}$. But now suppose that we take a sample $\pi_1, \ldots, \pi_n$ from the population, where we have specified that n_i sample members have the value $X = i$, etc. This could be by assignment, when we are trying to determine whether a cause–effect relationship exists; or we might have a populations $\Pi_1, \ldots, \Pi_a$ and want to see whether there is any difference in the distribution of Y between populations. Note that $n_1 + \cdots + n_a = n$.

In both cases, we again want to make inferences about the conditional distributions of Y given X as represented by the following table.

	$Y=1$	$\cdots$	$Y=b$
$X=1$	$\theta_{1\mid X=1}$	$\cdots$	$\theta_{b\mid X=1}$
$\vdots$	$\vdots$		$\vdots$
$X=a$	$\theta_{1\mid X=a}$	$\cdots$	$\theta_{b\mid X=a}$

A difference in the conditional distributions means there is a relationship between Y and X. If we denote the number of observations in the ith sample that have $Y = j$ by f_{ij}, then assuming the sample sizes are small relative to the population sizes, the likelihood function is given by

$$L(\theta_{1\mid X=1}, \ldots, \theta_{b\mid X=a} \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{a} \prod_{j=1}^{b} (\theta_{j\mid X=i})^{f_{ij}}, \qquad (10.2.3)$$

and the MLE is given by $\hat{\theta}_{j|X=i} = f_{ij}/n_i$ (Problem 10.2.18).

There is no relationship between Y and X if and only if the conditional distributions do not vary as we vary X, or if and only if

$$H_0 : \theta_{j|X=1} = \cdots = \theta_{j|X=a} = \theta_j$$

for all $j = 1, \ldots, b$ for some probability distribution $\theta_1, \ldots, \theta_b$. Under H_0, the likelihood function is given by

$$L(\theta_1, \ldots, \theta_b \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{j=1}^{b} \theta_j^{f_{\cdot j}}, \tag{10.2.4}$$

and the MLE of θ_j is given by $\hat{\theta}_j = f_{\cdot j}/n$ (see Problem 10.2.19). Then, applying Theorem 9.1.2, we have that the statistic

$$X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(f_{ij} - n_i \hat{\theta}_j)^2}{n_i \hat{\theta}_j}$$

has an approximate $\chi^2((a-1)(b-1))$ distribution under H_0 because there are $a(b-1)$ free parameters in the full model, $(b-1)$ parameters in the independence model, and $a(b-1) - (b-1) = (a-1)(b-1)$.

Consider an example.

EXAMPLE 10.2.2

This example is taken from a famous applied statistics book, *Statistical Methods,* 6th ed., by G. Snedecor and W. Cochran (Iowa State University Press, Ames, 1967). Individuals were classified according to their blood type Y (O, A, B, and AB, although the AB individuals were eliminated, as they were small in number) and also classified according to X, their disease status (peptic ulcer = P, gastric cancer = G, or control = C). So we have three populations; namely, those suffering from a peptic ulcer, those suffering from gastric cancer, and those suffering from neither. We suppose further that the individuals involved in the study can be considered as random samples from the respective populations.

The data are given in the following table.

	$Y = \text{O}$	$Y = \text{A}$	$Y = \text{B}$	Total
$X = \text{P}$	983	679	134	1796
$X = \text{G}$	383	416	84	883
$X = \text{C}$	2892	2625	570	6087

The estimated conditional distributions of Y given X are then as follows.

	$Y = \text{O}$	$Y = \text{A}$	$Y = \text{B}$
$X = \text{P}$	$983/1796 = 0.547$	$679/1796 = 0.378$	$134/1796 = 0.075$
$X = \text{G}$	$383/883 = 0.434$	$416/883 = 0.471$	$84/883 = 0.095$
$X = \text{C}$	$2892/6087 = 0.475$	$2625/6087 = 0.431$	$570/6087 = 0.093$

We now want to assess whether or not there is any evidence for concluding that a difference exists among these conditional distributions. Under the null hypothesis that no difference exists, the MLE's of the probabilities $\theta_1 = P(Y = \text{O})$, $\theta_2 = P(Y = \text{A})$, and $\theta_3 = P(Y = \text{B})$ are given by

$$\begin{aligned}\hat{\theta}_1 &= \frac{983 + 383 + 2892}{1796 + 883 + 6087} = 0.4857,\\ \hat{\theta}_2 &= \frac{679 + 416 + 2625}{1796 + 883 + 6087} = 0.4244,\\ \hat{\theta}_3 &= \frac{134 + 84 + 570}{1796 + 883 + 6087} = 0.0899.\end{aligned}$$

Then the estimated expected counts $n_i\hat{\theta}_j$ are given by the following table.

	$Y = \text{O}$	$Y = \text{A}$	$Y = \text{B}$
$X = \text{P}$	872.3172	762.2224	161.4604
$X = \text{G}$	428.8731	374.7452	79.3817
$X = \text{C}$	2956.4559	2583.3228	547.2213

The standardized residuals (using (9.1.6)) $(f_{ij} - n_i\hat{\theta}_j)/(n_i\hat{\theta}(1 - \hat{\theta}_j))^{1/2}$ are given by the following table.

	$Y = \text{O}$	$Y = \text{A}$	$Y = \text{B}$
$X = \text{P}$	5.2219	−3.9705	−2.2643
$X = \text{G}$	−3.0910	2.8111	0.5441
$X = \text{C}$	−1.659 2	1.0861	1.0227

We have that $X^2 = 40.5434$ and $P(\chi^2(4) > 40.5434) = 0.0000$, so we have strong evidence against the null hypothesis of no relationship existing between Y and X. Observe the large residuals when $X = \text{P}$ and $Y = \text{O}$, $Y = \text{A}$.

We are left with examining the conditional distributions to ascertain what form the relationship between Y and X takes. A useful tool in this regard is to plot the conditional distributions in bar charts, as we have done in Figure 10.2.1. From this, we see that the peptic ulcer population has greater proportion of blood type O than the other populations.

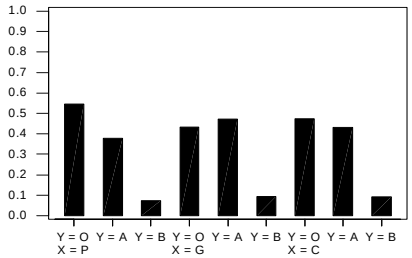

Figure 10.2.1: Plot of the conditional distributions of Y, given X, in Example 10.2.2.

∎

10.2.3 Bayesian Formulation

We now add a prior density π for the unknown values of the parameters of the models discussed in Sections 10.2.1 and 10.2.2. Depending on how we choose π, and depending on the particular computation we want to carry out, we could be faced with some difficult computational problems. Of course, we have the Monte Carlo methods available in such circumstances, which can often render a computation fairly straightforward.

The most common choice of prior in these circumstances is to choose a conjugate prior. Because the likelihoods discussed in this section are as in Example 7.1.3, we see immediately that Dirichlet priors will be conjugate for the full model in Section 10.2.1 and that products of independent Dirichlet priors will be conjugate for the full model in Section 10.2.2.

In Section 10.2.1, the general likelihood — i.e., no restrictions on the θ_{ij} — is of the form

$$L(\theta_{11}, \ldots, \theta_{ab} \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{a} \prod_{j=1}^{b} \theta_{ij}^{f_{ij}}.$$

If we place a Dirichlet($\alpha_{11}, \ldots, \alpha_{ab}$) prior on the parameter, then the posterior density is proportional to

$$\prod_{i=1}^{a} \prod_{j=1}^{b} \theta_{ij}^{f_{ij}+\alpha_{ij}-1},$$

so the posterior is a Dirichlet($f_{11} + \alpha_{11}, \ldots, f_{ab} + \alpha_{ab}$) distribution.

In Section 10.2.2, the general likelihood is of the form

$$L(\theta_{1|X=1}, \ldots, \theta_{b|X=a} \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{a} \prod_{j=1}^{b} (\theta_{j|X=i})^{f_{ij}}.$$

Because $\sum_{j=1}^{b} \theta_{j|X=i} = 1$ for each $i = 1, \ldots, a$, we must place a prior on each distribution $(\theta_{1|X=i}, \ldots, \theta_{b|X=i})$. If we choose the prior on the ith distribution to be Dirichlet($\alpha_{1|i}, \ldots, \alpha_{a|i}$), then the posterior density is proportional to

$$\prod_{i=1}^{a} \prod_{j=1}^{b} \theta_{j|i}^{f_{ij}+\alpha_{j|i}-1}.$$

We recognize this as the product of independent Dirichlet distributions, with the posterior distribution on $(\theta_{1|X=i}, \ldots, \theta_{b|X=i})$ equal to a

$$\text{Dirichlet}(f_{i1} + \alpha_{1|i}, \ldots, f_{ib} + \alpha_{b|i})$$

distribution.

A special and important case of the Dirichlet priors corresponds to the situation in which we feel that we have no information about the parameter. In such a situation, it

makes sense to choose all the parameters of the Dirichlet to be 1, so that the priors are all uniform.

There are many characteristics of a Dirichlet distribution that can be evaluated in closed form, e.g., the expectation of any polynomial (see Problem 10.2.20). But still there will be many quantities for which exact computations will not be available. It turns out that we can always easily generate samples from Dirichlet distributions, provided we have access to a generator for beta distributions. This is available with most statistical packages. We now discuss how to do this.

EXAMPLE 10.2.3 *Generating from a Dirichlet*$(\alpha_1, \ldots, \alpha_k)$ *Distribution*
The technique we discuss here is a commonly used method for generating from multivariate distributions. If we want to generate a value of the random vector $(X_1, \ldots, X_k)$, then we can proceed as follows. First, generate a value x_1 from the marginal distribution of X_1. Next, generate a value x_2 from the conditional distribution of X_2 given $X_1 = x_1$. Then generate a value x_3 from the conditional distribution of X_3, given that $X_1 = x_1$ and $X_2 = x_2$, etc.

If the distribution of X is discrete, then we have that the probability of a particular vector of values $(x_1, x_2, \ldots, x_k)$ arising via this scheme is

$$P(X_1 = x_1)P(X_2 = x_2 \mid X_1 = x_1) \cdots P(X_k = x_k \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1}).$$

Expanding each of these conditional probabilities, we obtain

$$P(X_1 = x_1)\frac{P(X_1=x_1, X_2=x_2)}{P(X_1=x_1)} \cdots \frac{P(X_1=x_1,\ldots,X_{k-1}=x_{k-1}, X_k=x_k)}{P(X_1=x_1,\ldots,X_{k-1}=x_{k-1})},$$

which equals $P(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_k = x_k)$, and so $(x_1, x_2, \ldots, x_k)$ is a value from the joint distribution of $(X_1, \ldots, X_k)$. This approach also works for absolutely continuous distributions, and the proof is the same but uses density functions instead.

In the case of $(X_1, \ldots, X_{k-1}) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, we have that (see Challenge 10.2.23) $X_1 \sim \text{Beta}(\alpha_1, \alpha_2 + \cdots + \alpha_k)$ and X_i given $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$ has the same distribution as $(1 - x_1 - \cdots - x_{i-1})U_i$, where

$$U_i \sim \text{Beta}(\alpha_i, \alpha_{i+1} + \cdots + \alpha_k)$$

and $U_2, \ldots, U_{k-1}$ are independent. Note that $X_k = 1 - X_1 - \cdots - X_{k-1}$ for any Dirichlet distribution. So we generate $X_1 \sim \text{Beta}(\alpha_1, \alpha_2 + \cdots + \alpha_k)$, generate $U_2 \sim \text{Beta}(\alpha_2, \alpha_3 + \cdots + \alpha_k)$ and put $X_2 = (1-X_1)U_2$, generate $U_3 \sim \text{Beta}(\alpha_3, \alpha_4 + \cdots + \alpha_k)$ and put $X_3 = (1 - X_1 - X_2)U_3$, etc.

Below, we present a table of a sample of $n = 5$ values from a Dirichlet(2, 3, 1, 1.5) distribution.

	X_1	X_2	X_3	X_4
1	0.116159	0.585788	0.229019	0.069034
2	0.166639	0.566369	0.056627	0.210366
3	0.411488	0.183686	0.326451	0.078375
4	0.483124	0.316647	0.115544	0.084684
5	0.117876	0.147869	0.418013	0.316242

Appendix B contains the code used for this. It can be modified to generate from any Dirichlet distribution. ∎

Summary of Section 10.2

- In this section, we have considered the situation in which we have a categorical response variable and a categorical predictor variable.
- We distinguished two situations. The first arises when the value of the predictor variable is not assigned, and the second arises when it is.
- In both cases, the test of the null hypothesis that no relationship exists involved the chi-squared test.

EXERCISES

10.2.1 The following table gives the counts of accidents for two successive years in a particular city.

	June	July	August
Year 1	60	100	80
Year 2	80	100	60

Is there any evidence of a difference in the distribution of accidents for these months between the two years?

10.2.2 The following data are from a study by Linus Pauling (1971) *(*"The significance of the evidence about ascorbic acid and the common cold,"*Proceedings of the National Academy of Sciences*, Vol. 68, p. 2678), concerned with examining the relationship between taking vitamin C and the incidence of colds. Of 279 participants in the study, 140 received a placebo (sugar pill) and 139 received vitamin C.

	No Cold	Cold
Placebo	31	109
Vitamin C	17	122

Assess the null hypothesis that there is no relationship between taking vitamin C and the incidence of the common cold.

10.2.3 A simulation experiment is carried out to see whether there is any relationship between the first and second digits of a random variable generated from a Uniform[0, 1] distribution. A total of 1000 uniforms were generated; if the first and second digits were in {0, 1, 2, 3, 4} they were recorded as a 0, and as a 1 otherwise. The cross-classified data are given in the following table.

	Second digit 0	Second digit 1
First digit 0	240	250
First digit 1	255	255

Assess the null hypothesis that there is no relationship between the digits.

10.2.4 Grades in a first-year calculus course were obtained for randomly selected students at two universities and classified as pass or fail. The following data were obtained.

	Fail	Pass
University 1	33	143
University 2	22	263

Is there any evidence of a relationship between calculus grades and university?

10.2.5 The following data are recorded in *Statistical Methods for Research Workers* , by R. A. Fisher (Hafner Press, New York, 1922), and show the classifications of 3883 Scottish children by gender (X) and hair color (Y).

	$Y =$ fair	$Y =$ red	$Y =$ medium	$Y =$ dark	$Y =$ jet black
$X =$ m	592	119	849	504	36
$X =$ f	544	97	677	451	14

(a) Is there any evidence for a relationship between hair color and gender?

(b) Plot the appropriate bar chart(s).

(c) Record the residuals and relate these to the results in parts (a) and (b). What do you conclude about the size of any deviation from independence?

10.2.6 Suppose we have a controllable predictor X that takes four different values, and we measure a binary-valued response Y. A random sample of 100 was taken from the population and the value of X was randomly assigned to each individual in such a way that there are 25 sample members taking each of the possible values of X. Suppose that the following data were obtained.

	$X = 1$	$X = 2$	$X = 3$	$X = 4$
$Y = 0$	12	10	16	14
$Y = 1$	13	15	9	11

(a) Assess whether or not there is any evidence against a cause–effect relationship existing between X and Y.

(b) Explain why it is possible in this example to assert that any evidence found that a relationship exists is evidence that a cause–effect relationship exists.

10.2.7 Write out in full how you would generate a value from a Dirichlet(1, 1, 1, 1) distribution.

10.2.8 Suppose we have two categorical variables defined on a population Π and we conduct a census. How would you decide whether or not a relationship exists between X and Y? If you decided that a relationship existed, how would you distinguish between a strong and a weak relationship?

10.2.9 Suppose you simultaneously roll two dice n times and record the outcomes. Based on these values, how would you assess the null hypothesis that the outcome on each die is independent of the outcome on the other?

10.2.10 Suppose a professor wants to assess whether or not there is any difference in the final grade distributions (A, B, C, D, and F) between males and females in a particular class. To assess the null hypothesis that there is no difference between these distributions, the professor carries out a chi-squared test.

(a) Discuss how the professor carried out this test.

(b) If the professor obtained evidence against the null hypothesis, discuss what concerns you have over the use of the chi-squared test.

10.2.11 Suppose that a chi-squared test is carried out, based on a random sample of n from a population, to assess whether or not two categorical variables X and Y are

independent. Suppose the P-value equals 0.001 and the investigator concludes that there is evidence against independence. Discuss how you would check to see if the deviation from independence was of practical significance.

PROBLEMS

10.2.12 In Example 10.2.1, place a uniform prior on the parameters (a Dirichlet distribution with all parameters equal to 1) and then determine the posterior distribution of the parameters.

10.2.13 In Example 10.2.2, place a uniform prior on the parameters of each population (a Dirichlet distribution with all parameters equal to 1) and such that the three priors are independent. Then determine the posterior distribution.

10.2.14 In a 2×2 table with probabilities θ_{ij}, prove that the row and column variables are independent if and only if

$$\frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} = 1,$$

namely, we have independence if and only if the *cross-ratio* equals 1.

10.2.15 Establish that the likelihood in (10.2.1) is correct when the population size is infinite (or when we are sampling with replacement from the population).

10.2.16 (MV) Prove that the MLE of $(\theta_{11}, \ldots, \theta_{ab})$ in (10.2.1) is given by $\hat{\theta}_{ij} = f_{ij}/n$. Assume that $f_{ij} > 0$ for every i, j. (Hint: Use the facts that a continuous function on this parameter space Ω must achieve its maximum at some point in Ω and that, if the function is continuously differentiable at such a point, then all its first-order partial derivatives are zero there. This will allow you to conclude that the unique solution to the score equations must be the point where the log-likelihood is maximized. Try the case where $a = 2, b = 2$ first.)

10.2.17 (MV) Prove that the MLE of $(\theta_{1\cdot}, \ldots, \theta_{a\cdot}, \theta_{\cdot 1}, \ldots, \theta_{\cdot b})$ in (10.2.2) is given by $\hat{\theta}_{i\cdot} = f_{i\cdot}/n$ and $\hat{\theta}_{\cdot j} = f_{\cdot j}/n$. Assume that $f_{i\cdot} > 0$, $f_{\cdot j} > 0$ for every i, j. (Hint: Use the hint in Problem 10.2.16.)

10.2.18 (MV) Prove that the MLE of $\left(\theta_{1|X=1}, \ldots, \theta_{b|X=a}\right)$ in (10.2.3) is given by $\hat{\theta}_{j|X=i} = f_{ij}/n_i$. Assume that $f_{ij} > 0$ for every i, j. (Hint: Use the hint in Problem 10.2.16.)

10.2.19 (MV) Prove that the MLE of $(\theta_1, \ldots, \theta_b)$ in (10.2.4) is given by $\hat{\theta}_j = f_{\cdot j}/n$. Assume that $f_{\cdot j} > 0$ for every i, j. (Hint: Use the hint in Problem 10.2.16.)

10.2.20 Suppose that $X = (X_1, \ldots, X_{k-1}) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$. Determine $E(X_1^{l_1} \cdots X_k^{l_k})$ in terms of the gamma function, when $l_i > 0$ for $i = 1, \ldots, k$.

COMPUTER PROBLEMS

10.2.21 Suppose that $(\theta_1, \theta_2, \theta_3, \theta_4) \sim \text{Dirichlet}(1, 1, 1, 1)$, as in Exercise 10.2.7. Generate a sample of size $N = 10^4$ from this distribution and use this to estimate the expectations of the θ_i. Compare these estimates with their exact values. (Hint: There is some relevant code in Appendix B for the generation; see Appendix C for formulas for the exact values of these expectations.)

10.2.22 For Problem 10.2.12, generate a sample of size $N = 10^4$ from the posterior distribution of the parameters and use this to estimate the posterior expectations of the cell probabilities. Compare these estimates with their exact values. (Hint: There is some relevant code in Appendix B for the generation; see Appendix C for formulas for the exact values of these expectations.)

CHALLENGES

10.2.23 (MV) Establish the validity of the method discussed in Example 10.2.3 for generating from a Dirichlet$(\alpha_1, \ldots, \alpha_k)$ distribution.

10.3 Quantitative Response and Predictors

When the response and predictor variables are all categorical, it can be difficult to formulate simple models that adequately describe the relationship between the variables. We are left with recording the conditional distributions and plotting these in bar charts. When the response variable is quantitative, however, useful models have been formulated that give a precise mathematical expression for the form of the relationship that may exist. We will study these kinds of models in the next three sections. This section concentrates on the situation in which all the variables are quantitative.

10.3.1 The Method of Least Squares

The *method of least squares* is a general method for obtaining an estimate of a distribution mean. It does not require specific distributional assumptions and so can be thought of as a distribution-free method (see Section 6.4).

Suppose we have a random variable Y, and we want to estimate $E(Y)$ based on a sample $(y_1, \ldots, y_n)$. The following principle is commonly used to generate estimates.

> The *least-squares principle* says that we select the point $t(y_1, \ldots, y_n)$, in the set of possible values for $E(Y)$, that minimizes the sum of squared deviations (hence, "least squares") given by $\sum_{i=1}^{n}(y_i - t(y_1, \ldots, y_n))^2$. Such an estimate is called a *least-squares estimate.*

Note that a least-squares estimate is defined for every sample size, even $n = 1$.

To implement least squares, we must find the minimizing point $t(y_1, \ldots, y_n)$. Perhaps a first guess at this value is the sample average $\bar{y}$. Because $\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - t(y_1, \ldots, y_n)) = (\bar{y} - t(y_1, \ldots, y_n))(\sum_{i=1}^{n} y_i - n\bar{y}) = 0$, we have

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - t(y_1, \ldots, y_n))^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - t(y_1, \ldots, y_n))^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - t(y_1, \ldots, y_n)) + \sum_{i=1}^{n}(\bar{y} - t(y_1, \ldots, y_n))^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - t(y_1, \ldots, y_n))^2 . \qquad (10.3.1)
\end{aligned}$$

Therefore, the smallest possible value of (10.3.1) is $\sum_{i=1}^{n}(y_i - \bar{y}^2)$, and this is assumed by taking $t(y_1, \ldots, y_n) = \bar{y}$. Note, however, that $\bar{y}$ might not be a possible value for $E(Y)$ and that, in such a case, it will not be the least-squares estimate. In general, (10.3.1) says that the least-squares estimate is the value $t(y_1, \ldots, y_n)$ that is closest to $\bar{y}$ and is a possible value for $E(Y)$.

Consider the following example.

EXAMPLE 10.3.1
Suppose that Y has one of the distributions on $S = \{0, 1\}$ given in the following table.

	$y = 0$	$y = 1$
$p_1(y)$	1/2	1/2
$p_2(y)$	1/3	2/3

Then the mean of Y is given by

$$E_1(Y) = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2} \text{ or } E_2(Y) = 0\left(\frac{1}{3}\right) + 1\left(\frac{2}{3}\right) = \frac{2}{3}.$$

Now suppose we observe the sample $(0, 0, 1, 1, 1)$ and so $\bar{y} = 3/5$. Because the possible values for $E(Y)$ are in $\{1/2, 2/3\}$, we see that $t(0, 0, 1, 1, 1) = 2/3$ because $(3/5 - 2/3)^2 = 0.004$ while $(3/5 - 1/2)^2 = 0.01$. ∎

Whenever the set of possible values for $E(Y)$ is an interval (a, b), however, and $P(Y \in (a, b)) = 1$, then $\bar{y} \in (a, b)$. This implies that $\bar{y}$ is the least-squares estimator of $E(Y)$. So we see that in quite general circumstances, $\bar{y}$ is the least-squares estimate.

There is an equivalence between least squares and the maximum likelihood method when we are dealing with normal distributions.

EXAMPLE 10.3.2 *Least Squares with Normal Distributions*
Suppose that $(y_1, \ldots, y_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where μ is unknown. Then the MLE of μ is obtained by finding the value of μ that maximizes

$$L(\mu \mid y_1, \ldots, y_n) = \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{y} - \mu)^2\right\}.$$

Equivalently, the MLE maximizes the log-likelihood

$$l(\mu \mid y_1, \ldots, y_n) = -\frac{n}{2\sigma_0^2}(\bar{y} - \mu)^2.$$

So we need to find the value of μ that minimizes $(\bar{y} - \mu)^2$ just as with least squares.

In the case of the normal location model, we see that the least-squares estimate and the MLE of θ agree. This equivalence is true in general for normal models (e.g., the location-scale normal model), at least when we are considering estimates of location parameters. ∎

Some of the most important applications of least squares arise when we have that the response is a random vector $Y = (Y_1, \ldots, Y_n)' \in R^n$ (the prime $'$ indicates that

we consider Y as a column), and we observe a single observation $y = (y_1, \ldots, y_n)' \in R^n$. The expected value of $Y \in R^n$ is defined to be the vector of expectations of its component random variables, namely,

$$E(Y) = \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{pmatrix} \in R^n.$$

The least-squares principle then says that, based on the single observation $y = (y_1, \ldots, y_n)$, we must find

$$t(y) = t(y_1, \ldots, y_n) = t_1(y_1, \ldots, y_n), \ldots, t_n(y_1, \ldots, y_n))',$$

in the set of possible values for $E(Y)$ (a subset of R^n), that minimizes

$$\sum_{i=1}^{n} (y_i - t_i(y_1, \ldots, y_n))^2. \tag{10.3.2}$$

So $t(y)$ is the possible value for $E(Y)$ that is closest to y, as the squared distance between two points $x, y \in R^n$ is given by $\sum_{i=1}^{n}(x_i - y_i)^2$.

As is common in statistical applications, suppose that there are predictor variables that may be related to Y and whose values are observed. In this case, we will replace $E(Y)$ by its conditional mean, given the observed values of the predictors. The least-squares estimate of the conditional mean is then the value $t(y_1, \ldots, y_n)$, in the set of possible values for the conditional mean of Y, that minimizes (10.3.2). We will use this definition in the following sections.

Finding the minimizing value of $t(y)$ in (10.3.2) can be a challenging optimization problem when the set of possible values for the mean is complicated. We will now apply least squares to some important problems where the least-squares solution can be found in closed form.

10.3.2 The Simple Linear Regression Model

Suppose we have a single quantitative response variable Y and a single quantitative predictor X, e.g., Y could be blood pressure measured in pounds per square inch and X could be age in years. To study the relationship between these variables, we examine the conditional distributions of Y, given $X = x$, to see how these change as we change x.

We might choose to examine a particular characteristic of these distributions to see how it varies with x. Perhaps the most commonly used characteristic is the conditional mean of Y given $X = x$, or $E(Y \mid X = x)$ (see Section 3.5).

In the *regression model* (see Section 10.1), we *assume* that the conditional distributions have constant shape and that they change, as we change x, at most through the conditional mean. In the *simple linear regression model*, we assume that the only way the conditional mean can change is via the relationship

$$E(Y \mid X = x) = \beta_1 + \beta_2 x,$$

for some unknown values of $\beta_1 \in R^1$ (the intercept term) and $\beta_2 \in R^1$ (the slope coefficient). We also refer to β_1 and β_2 as the *regression coefficients*.

Suppose we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y). Then, using the simple linear regression model, we have that

$$E\left(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \middle| X_1 = x_1, \ldots, X_n = x_n\right) = \begin{pmatrix} \beta_1 + \beta_2 x_1 \\ \vdots \\ \beta_1 + \beta_2 x_n \end{pmatrix}. \tag{10.3.3}$$

Equation (10.3.3) tells us that the conditional expected value of the response $(Y_1, \ldots, Y_n)'$ is in a particular subset of R^n. Furthermore, (10.3.2) becomes

$$\sum_{i=1}^{n} (y_i - t_i(y))^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2, \tag{10.3.4}$$

and we must find the values of β_1 and β_2 that minimize (10.3.4). These values are called the *least-squares estimates* of β_1 and β_2.

Before we show how to do this, consider an example.

EXAMPLE 10.3.3
Suppose we obtained the following $n = 10$ data points (x_i, y_i).

(3.9, 8.9)	(2.6, 7.1)	(2.4, 4.6)	(4.1, 10.7)	(−0.2, 1.0)
(5.4, 12.6)	(0.6, 3.3)	(−5.6, −10.4)	(−1.1, −2.3)	(−2.1, −1.6)

In Figure 10.3.1, we have plotted these points together with the line $y = 1 + x$.

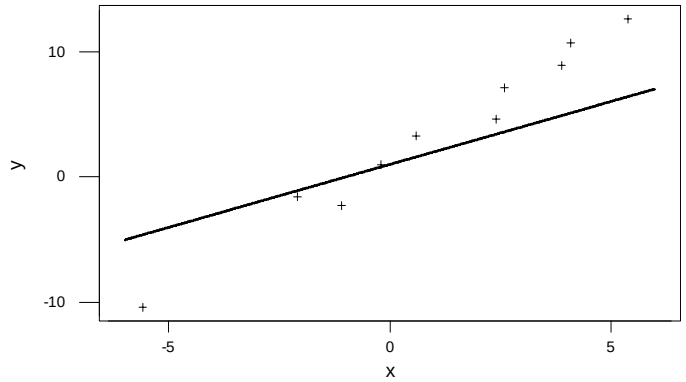

Figure 10.3.1: A plot of the data points (x_i, y_i) (+) and the line $y = 1 + x$ in Example 10.3.3.

Notice that with $\beta_1 = 1$ and $\beta_2 = 1$, then

$$(y_i - \beta_1 - \beta_2 x_i)^2 = (y_i - 1 - x_i)^2$$

is the squared vertical distance between the point (x_i, y_i) and the point on the line with the same x value. So (10.3.4) is the sum of these squared deviations and in this case equals

$$(8.9 - 1 - 3.9)^2 + (7.1 - 1 - 2.6)^2 + \cdots + (-1.6 - 1 + 2.1)^2 = 141.15.$$

If $\beta_1 = 1$ and $\beta_2 = 1$ were the least-squares estimates, then 141.15 would be equal to the smallest possible value of (10.3.4). In this case, it turns out (see Example 10.3.4) that the least-squares estimates are given by the values $\beta_1 = 1.33$, $\beta_2 = 2.06$, and the minimized value of (10.3.4) is given by 8.46, which is much smaller than 141.15.

So we see that, in finding the least-squares estimates, we are in essence finding the line $\beta_1 + \beta_2 x$ that best fits the data, in the sense that the sum of squared vertical deviations of the observed points to the line is minimized. ∎

Scatter Plots

As part of Example 10.3.3, we plotted the points $(x_1, y_1), \ldots, (x_n, y_n)$ in a graph. This is called a *scatter plot,* and it is a recommended first step as part of any analysis of the relationship between quantitative variables X and Y. A scatter plot can give us a very general idea of whether or not a relationship exists and what form it might take.

It is important to remember, however, that the appearance of such a plot is highly dependent on the scales we choose for the axes. For example, we can make a scatter plot look virtually flat (and so indicate that no relationship exists) by choosing to place too wide a range of tick marks on the y-axis. So we must always augment a scatter plot with a statistical analysis based on numbers.

Least-Squares Estimates, Predictions, and Standard Errors

For the simple linear regression model, we can work out exact formulas for the least-squares estimates of $\beta_1$and β_2.

Theorem 10.3.1 Suppose that $E(Y \mid X = x) = \beta_1 + \beta_2 x$, and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y). Then the least-squares estimates of β_1 and β_2 are given by

$$b_1 = \bar{y} - b_2\bar{x} \quad \text{and} \quad b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

respectively, whenever $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$.

PROOF The proof of this result can be found in Section 10.6. ∎

We call the line $y = b_1 + b_2 x$ the *least-squares line,* or *best-fitting line,* and $b_1 + b_2 x$ is the least-squares estimate of $E(Y \mid X = x)$. Note that $\sum_{i=1}^n (x_i - \bar{x})^2 = 0$ if and only if $x_1 = \cdots = x_n$. In such a case we cannot use least squares to estimate β_1 and β_2, although we can still estimate $E(Y \mid X = x)$ (see Problem 10.3.19).

Now that we have estimates b_1, b_2 of the regression coefficients, we want to use these for inferences about β_1 and β_2. These estimates have the unbiasedness property.

Theorem 10.3.2 If $E(Y \mid X = x) = \beta_1 + \beta_2 x$, and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then
(i) $E(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \beta_1$,
(ii) $E(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \beta_2$.

PROOF The proof of this result can be found in Section 10.6. ∎

Note that Theorem 10.3.2 and the theorem of total expectation imply that $E(B_1) = \beta_1$ and $E(B_2) = \beta_2$ unconditionally as well.

Adding the assumption that the conditional variances exist, we have the following theorem.

Theorem 10.3.3 If $E(Y \mid X = x) = \beta_1 + \beta_2 x$, $\text{Var}(Y \mid X = x) = \sigma^2$ for every x, and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then
(i) $\text{Var}(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2)$,
(ii) $\text{Var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2/\sum_{i=1}^n (x_i - \bar{x})^2$,
(iii) $\text{Cov}(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = -\sigma^2\bar{x}/\sum_{i=1}^n (x_i - \bar{x})^2$.

PROOF See Section 10.6 for the proof of this result. ∎

For the least-squares estimate $b_1 + b_2 x$ of the mean $E(Y \mid X = x) = \beta_1 + \beta_2 x$, we have the following result.

Corollary 10.3.1

$$\text{Var}(B_1 + B_2 x \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \quad (10.3.5)$$

PROOF See Section 10.6 for the proof of this result. ∎

A natural predictor of a future value of Y, when $X = x$, is given by the conditional mean $E(Y \mid X = x) = \beta_1 + \beta_2 x$. Because we do not know the values of $\beta_1$and β_2, we use the estimated mean $b_1 + b_2 x$ as the predictor.

When we are predicting Y at an x value that lies within the range of the observed values of X, we refer to this as an *interpolation*. When we want to predict at an x value that lies outside this range, we refer to this as an *extrapolation*. Extrapolations are much less reliable than interpolations. The farther away x is from the observed range of X values, then, intuitively, the less reliable we feel such a prediction will be. Such considerations should always be borne in mind. From (10.3.5), we see that the variance of the prediction at the value $X = x$ increases as x moves away from $\bar{x}$. So to a certain extent, the standard error does reflect this increased uncertainty, but note that its form is based on the assumption that the simple linear regression model is correct.

Even if we accept the simple linear regression model based on the observed data (we will discuss model checking later in this section), this model may fail to apply for very different values of x, and so the predictions would be in error.

We want to use the results of Theorem 10.3.3 and Corollary 10.3.1 to calculate standard errors of the least-squares estimates. Because we do not know σ^2, however, we need an estimate of this quantity as well. The following result shows that

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2 \tag{10.3.6}$$

is an unbiased estimate of σ^2.

Theorem 10.3.4 If $E(Y \mid X = x) = \beta_1 + \beta_2 x$, $\text{Var}(Y \mid X = x) = \sigma^2$ for every x, and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then

$$E(S^2 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2.$$

PROOF See Section 10.6 for the proof of this result. ∎

Therefore, the standard error of b_1 is then given by

$$s\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2},$$

and the standard error of b_2 is then given by

$$s\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{-1/2}.$$

Under further assumptions, these standard errors can be interpreted just as we interpreted standard errors of estimates of the mean in the location and location-scale normal models.

EXAMPLE 10.3.4 (*Example 10.3.3 continued*)
Using the data in Example 10.3.3 and the formulas of Theorem 10.3.1, we obtain $b_1 = 1.33$, $b_2 = 2.06$ as the least-squares estimates of the intercept and slope, respectively. So the least-squares line is given by $1.33 + 2.06x$. Using (10.3.6), we obtain $s^2 = 1.06$ as the estimate of σ^2.

Using the formulas of Theorem 10.3.3, the standard error of b_1 is 0.3408, while the standard error of b_2 is 0.1023.

The prediction of Y at $X = 2.0$ is given by $1.33 + 2.06\,(2) = 5.45$. Using Corollary 10.3.1, this estimate has standard error 0.341. This prediction is an interpolation. ∎

The ANOVA Decomposition and the F-Statistic

The following result gives a decomposition of the *total sum of squares* $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Lemma 10.3.1 If $(x_1, y_1), \ldots, (x_n, y_n)$ are such that $\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$, then

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2.$$

PROOF The proof of this result can be found in Section 10.6. ∎

We refer to

$$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

as the *regression sum of squares* (RSS) and refer to

$$\sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

as the *error sum of squares* (ESS).

If we think of the total sum of squares as measuring the total observed variation in the response values y_i, then Lemma 10.3.1 provides a decomposition of this variation into the RSS, measuring changes in the response due to changes in the predictor, and the ESS, measuring changes in the response due to the contribution of random error.

It is common to write this decomposition in an *analysis of variance table* (ANOVA).

Source	Df	Sum of Squares	Mean Square
X	1	$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$	$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$
Error	$n-2$	$\sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$	s^2
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	

Here, Df stands for degrees of freedom (we will discuss how the Df entries are calculated in Section 10.3.4). The entries in the Mean Square column are calculated by dividing the corresponding sum of squares by the Df entry.

To see the significance of the ANOVA table, note that, from Theorem 10.3.3,

$$E\left(B_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,\middle|\, X_1 = x_1, \ldots, X_n = x_n \right) = \sigma^2 + \beta_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad (10.3.7)$$

which is equal to σ^2 if and only if $\beta_2 = 0$ (we are always assuming here that the x_i vary). Given that the simple linear regression model is correct, we have that $\beta_2 = 0$ if and only if there is no relationship between the response and the predictor. Therefore, $b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$ is an unbiased estimator of σ^2 if and only if $\beta_2 = 0$. Because s^2 is always an unbiased estimate of σ^2 (Theorem 10.3.4), a sensible statistic to use in assessing $H_0 : \beta_2 = 0$, is given by

$$F = \frac{\text{RSS}}{\text{ESS}/(n-2)} = \frac{b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s^2}, \quad (10.3.8)$$

as this is the ratio of two unbiased estimators of σ^2 when H_0 is true. We then conclude that we have evidence against H_0 when F is large, as (10.3.7) also shows that the numerator will tend to be larger than σ^2 when H_0 is false. We refer to (10.3.8) as the *F-statistic*. We will subsequently discuss the sampling distribution of F to see how to determine when the value F is so large as to be evidence against H_0.

EXAMPLE 10.3.5 (*Example 10.3.3 continued*)
Using the data of Example 10.3.3, we obtain

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \bar{y})^2 &= 437.01, \\ b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 &= 428.55, \\ \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2 &= 437.01 - 428.55 = 8.46, \end{aligned}$$

and so

$$F = \frac{b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s^2} = \frac{428.55}{1.06} = 404.29.$$

Note that F is much bigger than 1, and this seems to indicate a linear effect due to X. ∎

The Coefficient of Determination and Correlation

Lemma 10.3.1 implies that

$$R^2 = \frac{b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

satisfies $0 \leq R^2 \leq 1$. Therefore, the closer R^2 is to 1, the more of the observed total variation in the response is accounted for by changes in the predictor. In fact, we interpret R^2, called the *coefficient of determination*, as the proportion of the observed variation in the response explained by changes in the predictor via the simple linear regression.

The coefficient of determination is an important descriptive statistic, for, even if we conclude that a relationship does exist, it can happen that most of the observed variation is due to error. If we want to use the model to predict further values of the response, then the coefficient of determination tells us whether we can expect highly accurate predictions or not. A value of R^2 near 1 means highly accurate predictions, whereas a value near 0 means that predictions will not be very accurate.

EXAMPLE 10.3.6 (*Example 10.3.3 continued*)
Using the data of Example 10.3.3, we obtain $R^2 = 0.981$. Therefore, 98.1% of the observed variation in Y can be explained by the changes in X through the linear relation. This indicates that we can expect fairly accurate predictions when using this model, at least when we are predicting within the range of the observed X values. ∎

Recall that in Section 3.3, we defined the correlation coefficient between random variables X and Y to be

$$\rho_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{Sd}(X)\,\text{Sd}(Y)}.$$

In Corollary 3.6.1, we proved that $-1 \leq \rho_{XY} \leq 1$ with $\rho_{XY} = \pm 1$ if and only if $Y = a \pm cX$ for some constants $a \in R^1$ and $c > 0$. So ρ_{XY} can be taken as a measure of the extent to which a linear relationship exists between X and Y.

If we do not know the joint distribution of (X, Y), then we will have to estimate ρ_{XY}. Based on the observations $(x_1, y_1), \ldots, (x_n, y_n)$, the natural estimate to use is the *sample correlation coefficient*

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

where

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

is the *sample covariance* estimating $\text{Cov}(X, Y)$, and s_x, s_y are the sample standard deviations for the X and Y variables, respectively. Then $-1 \leq r_{xy} \leq 1$ with $r_{xy} = \pm 1$ if and only if $y_i = a \pm cx_i$ for some constants $a \in R^1$ and $c > 0$, for every i (the proof is the same as in Corollary 3.6.1 using the joint distribution that puts probability mass $1/n$ at each point (x_i, y_i) — see Problem 3.6.16).

The following result shows that the coefficient of determination is the square of the correlation between the observed X and Y values.

Theorem 10.3.5 If $(x_1, y_1), \ldots, (x_n, y_n)$ are such that $\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$, $\sum_{i=1}^{n} (y_i - \bar{y})^2 \neq 0$, then $R^2 = r_{xy}^2$.

PROOF We have

$$r_{xy}^2 = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2} = b_2^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = R^2,$$

where we have used the formula for b_2 given in Theorem 10.3.1. ∎

Confidence Intervals and Testing Hypotheses

We need to make some further assumptions in order to discuss the sampling distributions of the various statistics that we have introduced. We have the following results.

Theorem 10.3.6 If Y, given $X = x$, is distributed $N(\beta_1 + \beta_2 x, \sigma^2)$, and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then the conditional distributions of B_1, B_2, and S^2, given $X_1 = x_1, \ldots, X_n = x_n$, are as follows.
(i) $B_1 \sim N(\beta_1, \sigma^2(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2))$
(ii) $B_2 \sim N(\beta_2, \sigma^2/\sum_{i=1}^n (x_i - \bar{x})^2)$
(iii) $B_1 + B_2 x \sim N(\beta_1 + \beta_2 x, \sigma^2(1/n + (x - \bar{x})^2/\sum_{i=1}^n (x_i - \bar{x})^2))$
(iv) $(n-2)\,S^2/\sigma^2 \sim \chi^2(n-2)$ independent of (B_1, B_2)

PROOF The proof of this result can be found in Section 10.6. ∎

Corollary 10.3.2
(i) $(B_1 - \beta_1)/(S(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2)^{1/2}) \sim t(n-2)$
(ii) $(B_2 - \beta_2)(\sum_{i=1}^n (x_i - \bar{x})^2)^{1/2}/S \sim t(n-2)$
(iii)
$$\frac{B_1 + B_2 x - \beta_1 - \beta_2 x}{S((1/n + (x - \bar{x})^2)/\sum_{i=1}^n (x_i - \bar{x})^2)^{1/2}} \sim t(n-2)$$
(iv) If F is defined as in (10.3.8), then $H_0 : \beta_2 = 0$ is true if and only if $F \sim F(1, n-2)$.

PROOF The proof of this result can be found in Section 10.6. ∎

Using Corollary 10.3.2(i), we have that

$$b_1 \pm s\left(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2} t_{(1+\gamma)/2}\,(n-2)$$

is an exact γ-confidence interval for β_1. Also, from Corollary 10.3.2(ii),

$$b_2 \pm s\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{-1/2} t_{(1+\gamma)/2}\,(n-2)$$

is an exact γ-confidence interval for β_2.

From Corollary 10.3.2(iv), we can test $H_0 : \beta_2 = 0$ by computing the P-value

$$P\left(F \geq \frac{b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{s^2}\right), \tag{10.3.9}$$

where $F \sim F(1, n-2)$, to see whether or not the observed value (10.3.8) is surprising. This is sometimes called the *ANOVA test*. Note that Corollary 10.3.2(ii) implies that we can also test $H_0 : \beta_2 = 0$ by computing the P-value

$$P\left(|T| \geq \frac{b_2 \left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}}{s}\right), \tag{10.3.10}$$

where $T \sim t(n-2)$. The proof of Corollary 10.3.2(iv) reveals that (10.3.9) and (10.3.10) are equal.

EXAMPLE 10.3.7 (*Example 10.3.3 continued*)
Using software or Table D.4, we obtain $t_{0.975}(8) = 2.306$. Then, using the data of Example 10.3.3, we obtain a 0.95-confidence interval for β_1 as

$$b_1 \pm s\left(1/n + \bar{x}^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{1/2} t_{(1+\gamma)/2}(n-2)$$
$$= 1.33 \pm (0.3408)\,(2.306) = [0.544, 2.116]$$

and a 0.95-confidence interval for β_2 as

$$b_2 \pm s\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{-1/2} t_{(1+\gamma)/2}(n-2)$$
$$= 2.06 \pm (0.1023)\,(2.306) = [1.824, 2.296]\,.$$

The 0.95-confidence interval for β_2 does not include 0, so we have evidence against the null hypothesis $H_0 : \beta_2 = 0$ and conclude that there is evidence of a relationship between X and Y. This is confirmed by the F-test of this null hypothesis, as it gives the P-value $P(F \geq 404.29) = 0.000$ when $F \sim F(1, 8)$.

Analysis of Residuals

In an application of the simple regression model, we must check to make sure that the assumptions make sense in light of the data we have collected. Model checking is based on the residuals $y_i - b_1 - b_2 x_i$ (after standardization), as discussed in Section 9.1. Note that the ith residual is just the difference between the observed value y_i at x_i and the predicted value $b_1 + b_2 x_i$ at x_i.

From the proof of Theorem 10.3.4, we have the following result.

Corollary 10.3.3
(i) $E(Y_i - B_1 - B_2 x_i \mid X_1 = x_1, \ldots, X_n = x_n) = 0$
(ii) $\mathrm{Var}(Y_i - B_1 - B_2 x_i \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$

This leads to the definition of the ith standardized residual as

$$\frac{y_i - b_1 - b_2 x_i}{s\left(1 - \frac{1}{n} - (x_i - \bar{x})^2 / \sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2\right)^{1/2}}. \tag{10.3.11}$$

Corollary 10.3.3 says that (10.3.11), with σ replacing s, is a value from a distribution with conditional mean 0 and conditional variance 1. Furthermore, when the conditional distribution of the response given the predictors is normal, then the conditional distribution of this quantity is $N(0, 1)$ (see Problem 10.3.21). These results

are approximately true for (10.3.11) for large n. Furthermore, it can be shown (see Problem 10.3.20) that

$$\begin{aligned}&\text{Cov}(Y_i - B_1 - B_2 x_i, Y_j - B_1 - B_2 x_j \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= -\sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})\,(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \right).\end{aligned}$$

Therefore, under the normality assumption, the residuals are approximately independent when n is large and

$$\frac{x_i - \bar{x}}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2}} \to 0$$

as $n \to \infty$. This will be the case whenever $\text{Var}(X)$ is finite (see Challenge 10.3.27) or, in the design context, when the values of the predictor are chosen accordingly. So one approach to model checking here is to see whether the values given by (10.3.11) look at all like a sample from the $N(0, 1)$ distribution. For this, we can use the plots discussed in Chapter 9.

EXAMPLE 10.3.8 (*Example 10.3.3 continued*)
Using the data of Example 10.3.3, we obtain the following standardized residuals.

−0.49643	0.43212	−1.73371	1.00487	0.08358
0.17348	0.75281	−0.28430	−1.43570	1.51027

These are plotted against the predictor x in Figure 10.3.2.

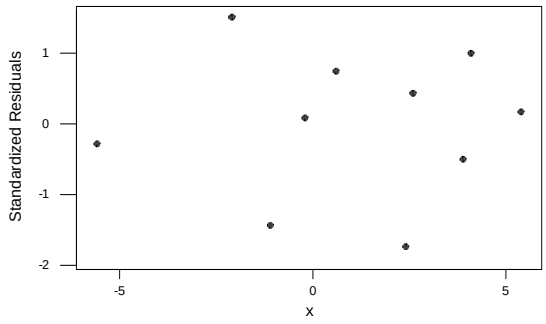

Figure 10.3.2: Plot of the standardized residuals in Example 10.3.8.

It is recommended that we plot the standardized residuals against the predictor, as this may reveal some underlying relationship that has not been captured by the model. This residual plot looks reasonable. In Figure 10.3.3, we have a normal probability plot of the standardized residuals. These points lie close to the line through the origin with slope equal to 1, so we conclude that we have no evidence against the model here. ∎

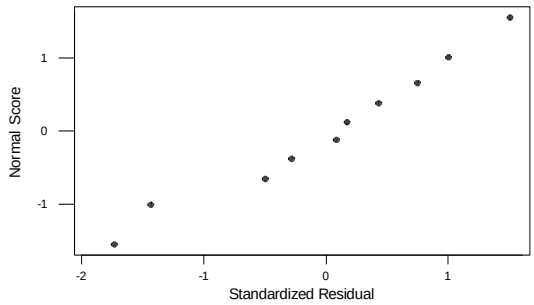

Figure 10.3.3: Normal probability plot of the standardized residuals in Example 10.3.8.

What do we do if model checking leads to a failure of the model? As discussed in Chapter 9, perhaps the most common approach is to consider making various transformations of the data to see whether there is a simple modification of the model that will pass. We can make transformations, not only to the response variable Y, but to the predictor variable X as well.

An Application of Simple Linear Regression Analysis

The following data set is taken from *Statistical Methods*, 6th ed., by G. Snedecor and W. Cochran (Iowa State University Press, Ames, 1967) and gives the record speed Y in miles per hour at the Indianapolis Memorial Day car races in the years 1911–1941, excepting the years 1917–1918. We have coded the year X starting at 0 in 1911 and incrementing by 1 for each year. There are $n = 29$ data points (x_i, y_i). The goal of the analysis is to obtain the least-squares line and, if warranted, make inferences about the regression coefficients. We take the normal simple linear regression model as our statistical model. Note that this is an observational study.

Year	Speed	Year	Speed	Year	Speed
0	74.6	12	91.0	22	104.2
1	78.7	13	98.2	23	104.9
2	75.9	14	101.1	24	106.2
3	82.5	15	95.9	25	109.1
4	89.8	16	97.5	26	113.6
5	83.3	17	99.5	27	117.2
8	88.1	18	97.6	28	115.0
9	88.6	19	100.4	29	114.3
10	89.6	20	96.6	30	115.1
11	94.5	21	104.1		

Using Theorem 10.3.1, we obtain the least-squares line as $y = 77.5681 + 1.27793x$. This line, together with a *scatter plot* of the values (x_i, y_i), is plotted in Figure 10.3.4.

The fit looks quite good, but this is no guarantee of model correctness, and we must carry out some form of model checking.

Figure 10.3.5 is a plot of the standardized residuals against the predictor. This plot looks reasonable, with no particularly unusual pattern apparent. Figure 10.3.6 is a normal probability plot of the standardized residuals. The curvature in the center might give rise to some doubt about the normality assumption. We generated a few samples of $n = 29$ from an $N(0, 1)$ distribution, however, and looking at the normal probability plots (always recommended) reveals that this is not much cause for concern. Of course, we should also carry out model checking procedures based upon the standardized residuals and using P-values, but we do not pursue this topic further here.

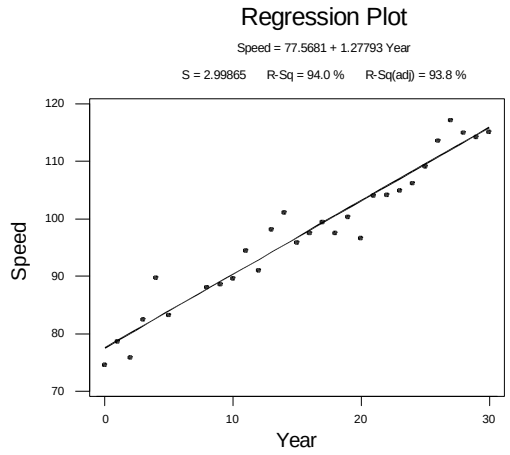

Figure 10.3.4: A scatter plot of the data together with a plot of the least-squares line.

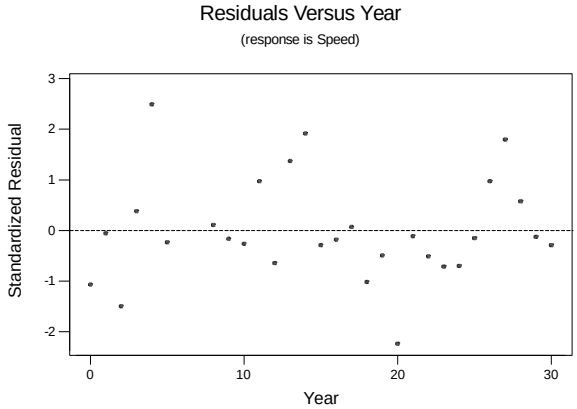

Figure 10.3.5: A plot of the standardized residuals against the predictor.

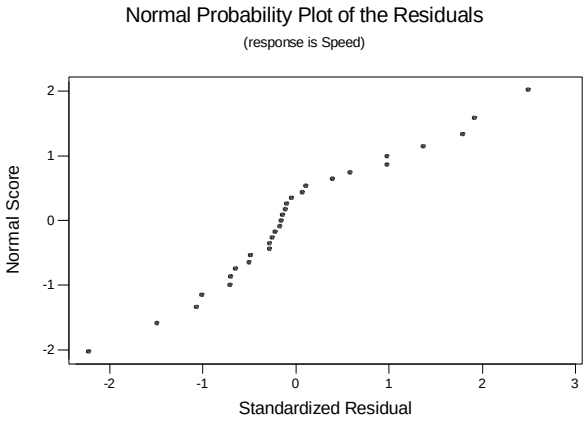

Figure 10.3.6: A normal probability plot of the standardized residuals.

Based on the results of our model checking, we decide to proceed to inferences about the regression coefficients. The estimates and their standard errors are given in the following table, where we have used the estimate of σ^2 given by $s^2 = (2.999)^2$, to compute the standard errors. We have also recorded the t-statistics appropriate for testing each of the hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$.

Coefficient	Estimate	Standard Error	t-statistic
β_1	77.568	1.118	69.39
β_2	1.278	0.062	20.55

From this, we see that the P-value for assessing $H_0 : \beta_2 = 0$ is given by

$$P(|T| \geq 20.55) = 0.000,$$

when $T \sim t\,(27)$, and so we have strong evidence against H_0. It seems clear that there is a strong positive relationship between Y and X. Since the 0.975 point of the $t\,(27)$ distribution equals 2.0518, a 0.95-confidence interval for β_2 is given by

$$1.278 \pm (0.062)\, 2.0518 = [1.1508, 1.4052]\,.$$

The ANOVA decomposition is given in the following table.

Source	Df	Sum of Squares	Mean Square
Regression	1	3797.0	3797.0
Error	27	242.8	9.0
Total	28	4039.8	

Accordingly, we have that $F = 3797.0/9.0 = 421.888$ and, as $F \sim F(1, 27)$ when $H_0 : \beta_2 = 0$ is true, $P(F > 421.888) = 0.000$, which simply confirms (as it must) what we got from the preceding t-test.

The coefficient of determination is given by $R^2 = 3797.0/4039.8 = 0.94$. Therefore, 94% of the observed variation in the response variable can be explained by the

changes in the predictor through the simple linear regression. The value of R^2 indicates that the fitted model will be an excellent predictor of future values, provided that the value of X that we want to predict at, is in the range (or close to it) of the values of X used to fit the model.

10.3.3 Bayesian Simple Linear Model (Advanced)

For the Bayesian formulation of the simple linear regression model with normal error, we need to add a prior distribution for the unknown parameters of the model, namely, β_1, β_2, and σ^2. There are many possible choices for this. A relevant prior is dependent on the application.

To help simplify the calculations, we reparameterize the model as follows. Let $\alpha_1 = \beta_1 + \beta_2\bar{x}$ and $\alpha_2 = \beta_2$. It is then easy to show (see Problem 10.3.24) that

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2 &= \sum_{i=1}^{n}(y_i - \alpha_1 - \alpha_2(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^{n}((y_i - \bar{y}) - (\alpha_1 - \bar{y}) - \alpha_2(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\alpha_1 - \bar{y})^2 + \alpha_2^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &\quad - 2\alpha_2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \qquad (10.3.12)\end{aligned}$$

The likelihood function, using this reparameterization, then equals

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \alpha_1 - \alpha_2\left(x_i - \bar{x}\right)\right)^2\right).$$

From (10.3.12), and setting

$$\begin{aligned}c_x^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2, \\ c_y^2 &= \sum_{i=1}^{n}(y_i - \bar{y})^2, \\ c_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),\end{aligned}$$

we can write this as

$$\begin{aligned}
&\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{c_y^2}{2\sigma^2}\right)\exp\left(-\frac{n}{2\sigma^2}\,(\alpha_1-\bar{y})^2\right)\\
&\qquad\times\exp\left(-\frac{1}{2\sigma^2}\left\{\alpha_2^2c_x^2-2\alpha_2c_{xy}\right\}\right)\\
&=\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{c_y^2-c_x^2a^2}{2\sigma^2}\right)\exp\left(-\frac{n}{2\sigma^2}\,(\alpha_1-\bar{y})^2\right)\\
&\qquad\times\exp\left(-\frac{c_x^2}{2\sigma^2}\,(\alpha_2-a)^2\right),
\end{aligned}$$

where the last equality follows from $\alpha_2^2c_x^2-2\alpha_2c_{xy}=c_x^2\,(\alpha_2-a)^2-c_x^2a^2$ with $a=c_{xy}/c_x^2$.

This implies that, whenever the prior distribution on (α_1,α_2) is such that α_1 and α_2 are independent given σ^2, then the posterior distributions of α_1 and α_2 are also independent given σ^2. Note also that $\bar{y}$ and a are the least-squares estimates (as well as the MLE's) of α_1 and α_2, respectively (see Problem 10.3.24).

Now suppose we take the prior to be

$$\begin{aligned}
\alpha_1\,|\,\alpha_2,\sigma^2 &\sim N(\mu_1,\tau_1^2\sigma^2),\\
\alpha_2\,|\,\sigma^2 &\sim N(\mu_2,\tau_2^2\sigma^2),\\
1/\sigma^2 &\sim \text{Gamma}(\kappa,\nu).
\end{aligned}$$

Note that α_1 and α_2 are independent given σ^2.

As it turns out, this prior is conjugate, so we can easily determine an exact form for the posterior distribution (see Problem 10.3.25). The joint posterior of $\left(\alpha_1,\alpha_2,1/\sigma^2\right)$ is given by

$$\begin{aligned}
\alpha_1\,|\,\alpha_2,\sigma^2 &\sim N\left(\left(n+\frac{1}{\tau_1^2}\right)^{-1}\left(n\bar{y}+\frac{\mu_1}{\tau_1^2}\right),\left(n+\frac{1}{\tau_1^2}\right)^{-1}\sigma^2\right),\\
\alpha_2\,|\,\sigma^2 &\sim N\left(\left(c_x^2+\frac{1}{\tau_2^2}\right)^{-1}\left(c_x^2a+\frac{\mu_2}{\tau_2^2}\right),\left(c_x^2+\frac{1}{\tau_2^2}\right)^{-1}\sigma^2\right),\\
\frac{1}{\sigma^2} &\sim \text{Gamma}\left(\kappa+\frac{n}{2},\nu_{xy}\right),
\end{aligned}$$

where

$$\nu_{xy}=\frac{1}{2}\left\{\begin{array}{c}c_y^2-c_x^2a^2+\left[n\bar{y}^2+\frac{\mu_1^2}{\tau_1^2}-\left(n+\frac{1}{\tau_1^2}\right)^{-1}\left(n\bar{y}+\frac{\mu_1}{\tau_1^2}\right)^2\right]\\ +\left[c_x^2a^2+\frac{\mu_2^2}{\tau_2^2}-\left(c_x^2+\frac{1}{\tau_2^2}\right)^{-1}\left(c_x^2a+\frac{\mu_2}{\tau_2^2}\right)^2\right]\end{array}\right\}+\nu.$$

Of course, we must select the values of the hyperparameters $\mu_1, \tau_1, \mu_2, \tau_2, \kappa$, and ν to fully specify the prior.

Now observe that for a diffuse analysis, i.e., when we have little or no prior information about the parameters, we let $\tau_1 \to \infty$, $\tau_2 \to \infty$, and $\nu \to 0$, and the posterior converges to

$$\begin{aligned} \alpha_1 \,|\, \alpha_2, \sigma^2 &\sim N(\bar{y}, \sigma^2/n), \\ \alpha_2 \,|\, \sigma^2 &\sim N(a, \sigma^2/c_x^2), \\ 1/\sigma^2 &\sim \text{Gamma}(\kappa + n/2, \nu_{xy}) \end{aligned}$$

where $\nu_{xy} = (1/2)\{c_y^2 - c_x^2 a^2\}$. But this still leaves us with the necessity of choosing the hyperparameter κ. We will see, however, that this choice has only a small effect on the analysis when n is not too small.

We can easily work out the marginal posterior distribution of the α_i. For example, in the diffuse case, the marginal posterior density of α_2 is proportional to

$$\begin{aligned} &\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{1/2} \exp\left\{-\frac{c_x^2}{2\sigma^2}(\alpha_2 - a)^2\right\} \left(\frac{1}{\sigma^2}\right)^{\kappa+(n/2)-1} \exp\left\{-\frac{\nu_{xy}}{\sigma^2}\right\} d\left(\frac{1}{\sigma^2}\right) \\ &= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\kappa+(n/2)-(1/2)} \exp\left\{-\left(\nu_{xy} + \frac{c_x^2}{2}(\alpha_2 - a)^2\right)\frac{1}{\sigma^2}\right\} d\left(\frac{1}{\sigma^2}\right). \end{aligned}$$

Making the change of variable $1/\sigma^2 \to w$, where

$$w = \left(\nu_{xy} + \frac{c_x^2}{2}(\alpha_2 - a)^2\right)\frac{1}{\sigma^2}$$

in the preceding integral, shows that the marginal posterior density of α_2 is proportional to

$$\left(1 + \frac{c_x^2}{2\nu_{xy}}(\alpha_2 - a)^2\right)^{-(\kappa+(n+1)/2)} \int_0^\infty w^{\kappa+(n/2)-(1/2)} \exp\{-w\}\, dw,$$

which is proportional to

$$\left(1 + \frac{c_x^2}{2\nu_{xy}}(\alpha_2 - a)^2\right)^{-(2\kappa+n+1)/2}.$$

This establishes (see Problem 4.6.17) that the posterior distribution of α_2 is specified by

$$\sqrt{2\kappa + n}\,\frac{\alpha_2 - a}{\sqrt{2\nu_{xy}/c_x^2}} \sim t(2\kappa + n).$$

So a γ-HPD (highest posterior density) interval for α_2 is given by

$$a \pm \frac{1}{\sqrt{2\kappa + n}} \sqrt{\frac{2\nu_{xy}}{c_x^2}}\, t_{(1+\gamma)/2}(2\kappa + n).$$

Note that these intervals will not change much as we change κ, provided that n is not too small.

We consider an application of a Bayesian analysis for such a model.

EXAMPLE 10.3.9 *Haavelmo's Data on Income and Investment*
The data for this example were taken from *An Introduction to Bayesian Inference in Econometrics*, by A. Zellner (Wiley Classics, New York, 1996). The response variable Y is income in U.S. dollars per capita (deflated), and the predictor variable X is investment in dollars per capita (deflated) for the United States for the years 1922–1941. The data are provided in the following table.

Year	Income	Investment	Year	Income	Investment
1922	433	39	1932	372	22
1923	483	60	1933	381	17
1924	479	42	1934	419	27
1925	486	52	1935	449	33
1926	494	47	1936	511	48
1927	498	51	1937	520	51
1928	511	45	1938	477	33
1929	534	60	1939	517	46
1930	478	39	1940	548	54
1931	440	41	1941	629	100

In Figure 10.3.7, we present a normal probability plot of the standardized residuals, obtained via a least-squares fit. In Figure 10.3.8, we present a plot of the standardized residuals against the predictor. Both plots indicate that the model assumptions are reasonable.

Suppose now that we analyze these data using the limiting diffuse prior with $\kappa = 2$. Here, we have that $\bar{y} = 483$, $c_y^2 = 64993$, $c_x^2 = 5710.55$, and $c_{xy} = 17408.3$, so that $a = 17408.3/5710.55 = 3.05$ and $v_{xy} = (64993 - 17408.3)/2 = 23792.35$. The posterior is then given by

$$\begin{aligned} \alpha_1 \,|\, \alpha_2, \sigma^2 &\sim N(483, \sigma^2/20), \\ \alpha_2 \,|\, \sigma^2 &\sim N(3.05, \sigma^2/5710.55), \\ 1/\sigma^2 &\sim \text{Gamma}(12, 23792.35). \end{aligned}$$

The primary interest here is in the investment multiplier α_2. By the above results, a 0.95-HPD interval for α_2, using $t_{0.975}(24) = 2.0639$, is given by

$$\begin{aligned} & a \pm \frac{1}{\sqrt{2\kappa + n}} \sqrt{\frac{2v_{xy}}{r_x^2}}\, t_{(1+\gamma)/2}(2\kappa + n - 1) \\ &= 3.05 \pm \frac{1}{\sqrt{24}} \sqrt{\frac{2 \cdot 23792.35}{5710.55}}\, t_{0.975}(24) = 3.05 \pm (0.589)\, 2.0639 \\ &= (1.834, 4.266). \blacksquare \end{aligned}$$

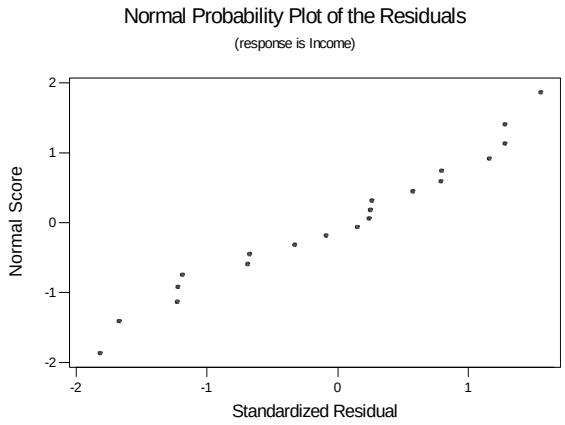

Figure 10.3.7: Normal probability plot of the standardized residuals in Example 10.3.9.

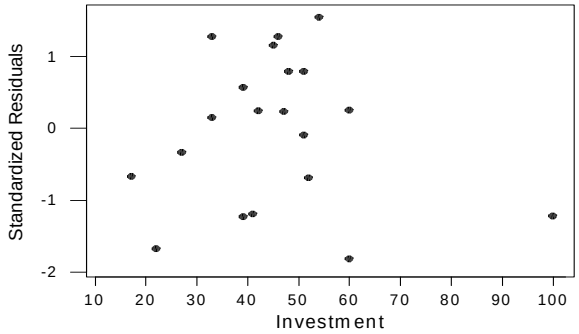

Figure 10.3.8: Plot of the standardized residuals against the predictor in Example 10.3.9.

10.3.4 The Multiple Linear Regression Model (Advanced)

We now consider the situation in which we have a quantitative response Y and quantitative predictors $X_1, \ldots, X_k$. For the regression model, we assume that the conditional distributions of Y, given the predictors, have constant shape and that they change, as the predictors change, at most through the conditional mean $E\,(Y \mid X_1 = x_1, \ldots, X_k = x_k)$. For the *linear regression model*, we assume that this conditional mean is of the form

$$E(Y \mid X_1 = x_1, \ldots, X_k = x_k) = \beta_1 x_1 + \cdots + \beta_k x_k. \tag{10.3.13}$$

This is linear in the unknown $\beta_i \in R^1$ for $i = 1, \ldots, k$.

We will develop only the broad outline of the analysis of the multiple linear regression model here. All results will be stated without proofs provided. The proofs can be found in more advanced texts. It is important to note, however, that all of these results are just analogs of the results we developed by elementary methods in Section 10.3.2, for the simple linear regression model.

Matrix Formulation of the Least-Squares Problem

For the analysis of the multiple linear regression model, we need some *matrix* concepts. We will briefly discuss some of these here, but also see Appendix A.4.

Let $A \in R^{m\times n}$ denote a rectangular array of numbers with m rows and n columns, and let a_{ij} denote the entry in the ith row and jth column (referred to as the (i, j)-th entry of A). For example,

$$A = \begin{pmatrix} 1.2 & 1.0 & 0.0 \\ 3.2 & 0.2 & 6.3 \end{pmatrix} \in R^{2\times 3}$$

denotes a 2×3 matrix and, for example, $a_{22} = 0.2$.

We can add two matrices of the same dimensions m and n by simply adding their elements componentwise. So if $A, B \in R^{m\times n}$ and $C = A + B$, then $c_{ij} = a_{ij} + b_{ij}$. Furthermore, we can multiply a matrix by a real number c by simply multiplying every entry in the matrix by c. So if $A \in R^{m\times n}$, then $B = cA \in R^{m\times n}$ and $b_{ij} = ca_{ij}$. We will sometimes write a matrix $A \in R^{m\times n}$ in terms of its columns as $A = (\ a_1 \quad \ldots \quad a_n\ )$ so that here $a_i \in R^m$. Finally, if $A \in R^{m\times n}$ and $b \in R^n$, then we define the product of A times b as $Ab = b_1a_1 + \cdots + b_na_n \in R^m$.

Suppose now that $Y \in R^n$ and that $E(Y)$ is constrained to lie in a set of the form

$$S = \{\beta_1 v_1 + \cdots + \beta_k v_k : \beta_i \in R^1, i = 1, \ldots, k\},$$

where $v_1, \ldots, v_k$ are fixed vectors in R^n. A set such as S is called a *linear subspace* of R^n. When $\{v_1, \ldots, v_k\}$ has the *linear independence property*, namely,

$$\beta_1 v_1 + \cdots + \beta_k v_k = 0$$

if and only if $\beta_1 = \cdots = \beta_k = 0$, then we say that S has dimension k and $\{v_1, \ldots, v_k\}$ is a *basis* for S.

If we set

$$V = (v_1 \cdots v_k) = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{pmatrix} \in R^{n\times k},$$

then we can write

$$E(Y) = \beta_1 v_1 + \cdots + \beta_k v_k = \begin{pmatrix} \beta_1 v_{11} + \beta_2 v_{12} + \cdots + \beta_k v_{1k} \\ \beta_1 v_{21} + \beta_2 v_{22} + \cdots + \beta_k v_{2k} \\ \vdots \\ \beta_1 v_{n1} + \beta_2 v_{n2} + \cdots + \beta_k v_{nk} \end{pmatrix} = V\beta$$

for some unknown point $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$. When we observe $y \in R^n$, then the least-squares estimate of $E(Y)$ is obtained by finding the value of β that minimizes

$$\sum_{i=1}^{n} \left(y_i - \beta_1 v_{i1} - \beta_2 v_{i2} - \cdots - \beta_k v_{ik}\right)^2.$$

It can be proved that a unique minimizing value for $\beta \in R^k$ exists whenever $\{v_1, \ldots, v_k\}$ is a basis. The minimizing value of β will be denoted by b and is called the *least-squares estimate* of β. The point $b_1 v_1 + \cdots + b_k v_k = Vb$ is the least-squares estimate of $E(Y)$ and is sometimes called the vector of *fitted values*. The point $y - Vb$ is called the vector of *residuals*.

We now consider how to calculate b. For this, we need to understand what it means to multiply the matrix $A \in R^{m \times k}$ on the right by the matrix $B \in R^{k \times n}$. The *matrix product* AB is defined to be the $m \times n$ matrix whose (i, j)-th entry is given by

$$\sum_{l=1}^{k} a_{il} b_{lj}.$$

Notice that the array A must have the same number of columns as the number of rows of B for this product to be defined. The *transpose* of a matrix $A \in R^{m \times k}$ is defined to be

$$A' = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1k} & \cdots & a_{mk} \end{pmatrix} \in R^{k \times m},$$

namely, the ith column of A becomes the ith row of A'. For a matrix $A \in R^{k \times k}$, the *matrix inverse* of A is defined to be the matrix A^{-1} such that

$$AA^{-1} = A^{-1}A = I,$$

where $I \in R^{k \times k}$ has 1's along its diagonal and 0's everywhere else; it is called the $k \times k$ *identity matrix*. It is not always the case that $A \in R^{k \times k}$ has an inverse, but when it does it can be shown that the inverse is unique. Note that there are many mathematical and statistical software packages that include the facility for computing matrix products, transposes, and inverses.

We have the following fundamental result.

Theorem 10.3.7 If $E(Y) \in S = \{\beta_1 v_1 + \cdots + \beta_k v_k : \beta_i \in R^1, i = 1, \ldots, k\}$ and the columns of $V = (v_1 \cdots v_k)$ have the linear independence property, then $(V'V)^{-1}$ exists, the least-squares estimate of β is unique, and it is given by

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = (V'V)^{-1} V'y. \tag{10.3.14}$$

Least-Squares Estimates, Predictions, and Standard Errors

For the linear regression model (10.3.13), we have that (writing X_{ij} for the jth value of X_i)

$$E\left(\left(\begin{array}{c} Y_1 \\ \vdots \\ Y_n \end{array}\right) \,\middle|\, X_{ij} = x_{ij} \text{ for all } i, j\right) = \left(\begin{array}{c} \beta_1 x_{11} + \cdots + \beta_k x_{1k} \\ \vdots \\ \beta_1 x_{n1} + \cdots + \beta_k x_{nk} \end{array}\right)$$
$$= \beta_1 v_1 + \cdots + \beta_k v_k = V\beta,$$

where $\beta = (\beta_1, \ldots, \beta_k)'$ and

$$V = \left(\begin{array}{cccc} v_1 & v_2 & \ldots & v_k \end{array}\right) = \left(\begin{array}{ccc} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{array}\right) \in R^{n \times k}.$$

We will assume, hereafter, that the columns $v_1, \ldots, v_k$ of V have the linear independence property. Then (replacing expectation by conditional expectation) it is immediate that the least-squares estimate of β is given by (10.3.14).

As with the simple linear regression model, we have a number of results concerning the least-squares estimates. We state these here without proof.

Theorem 10.3.8 If the $(x_{i1}, \ldots, x_{ik}, y_i)$ are independent observations for $i = 1, \ldots, n$, and the linear regression model applies, then

$$E(B_i \mid X_{ij} = x_{ij} \text{ for all } i, j) = \beta_i$$

for $i = 1, \ldots, k$.

So Theorem 10.3.8 states that the least-squares estimates are unbiased estimates of the linear regression coefficients.

If we want to assess the accuracy of these estimates, then we need to be able to compute their standard errors.

Theorem 10.3.9 If the $(x_{i1}, \ldots, x_{ik}, y_i)$ are independent observations for $i = 1, \ldots, n$, from the linear regression model, and if $\text{Var}(Y \mid X_1 = x_1, \ldots, X_k = x_k) = \sigma^2$ for every $x_1, \ldots, x_k$, then

$$\text{Cov}(B_i, B_j \mid X_{ij} = x_{ij} \text{ for all } i, j) = \sigma^2 c_{ij}, \tag{10.3.15}$$

where c_{ij} is the (i, j)-th entry in the matrix $(V'V)^{-1}$.

We have the following result concerning the estimation of the mean

$$E(Y \mid X_1 = x_1, \ldots, X_k) = x_k = \beta_1 x_1 + \cdots + \beta_k x_k$$

by the estimate $b_1 x_1 + \cdots + b_k x_k$.

Corollary 10.3.4

$$\begin{aligned}&\mathrm{Var}(B_1x_1 + \cdots + B_kx_k \mid X_{ij} = x_{ij} \text{ for all } i, j)\\ &= \sigma^2\Big(\sum_{i=1}^{k} x_i^2 c_{ii} + 2\sum_{i<j} x_i x_j c_{ij}\Big) = \sigma^2 x'(V'V)^{-1}x,\end{aligned} \tag{10.3.16}$$

where $x = (x_1, \ldots, x_k)$.

We also use $b_1x_1 + \cdots + b_kx_k = b'x$ as a prediction of a new response value when $X_1 = x_1, \ldots, X_k = x_k$.

We see, from Theorem 10.3.9 and Corollary 10.3.4, that we need an estimate of σ^2 to compute standard errors. The estimate is given by

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - b_1x_{i1} - \cdots - b_kx_{ik})^2 = \frac{1}{n-k}(y - Xb)'(y - Xb), \tag{10.3.17}$$

and we have the following result.

Theorem 10.3.10 If the $(x_{i1}, \ldots, x_{ik}, y_i)$ are independent observations for $i = 1, \ldots, n$, from the linear regression model, and if $\mathrm{Var}(Y \mid X_1 = x_1, \ldots, X_k = x_k) = \sigma^2$, then

$$E(S^2 \mid X_{ij} = x_{ij} \text{ for all } i, j) = \sigma^2.$$

Combining (10.3.15) and (10.3.17), we deduce that the standard error of b_i is $s\sqrt{c_{ii}}$. Combining (10.3.16) and (10.3.17), we deduce that the standard error of $b_1x_1 + \cdots + b_kx_k$ is

$$s\left(\sum_{i=1}^{k} x_i^2 c_{ii} + 2\sum_{i<j} x_i x_j c_{ij}\right)^{1/2} = s(x'(V'V)^{-1}x)^{1/2}.$$

The ANOVA Decomposition and F-Statistics

When one of the predictors $X_1, \ldots, X_k$ is constant, then we say that the model has an *intercept term*. By convention, we will always take this to be the first predictor. So when we want the model to have an intercept term, we take $X_1 \equiv 1$ and β_1 is the intercept, e.g., the simple linear regression model. Note that it is common to denote the intercept term by β_0 so that $X_0 \equiv 1$ and $X_1, \ldots, X_k$ denote the predictors that actually change. We will also adopt this convention when it seems appropriate.

Basically, inclusion of an intercept term is very common, as this says that, when the predictors that actually change have no relationship with the response Y, then the intercept is the unknown mean of the response. When we do not include an intercept, then this says we *know* that the mean response is 0 when there is no relationship between Y and the nonconstant predictors. Unless there is substantive, application-based evidence to support this, we will generally not want to make this assumption.

Denoting the intercept term by β_1, so that $X_1 \equiv 1$, we have the following ANOVA decomposition for this model that shows how to isolate the observed variation in Y that can be explained by changes in the nonconstant predictors.

Lemma 10.3.2 If, for $i = 1, \ldots, n$, the values $(x_{i1}, \ldots, x_{ik}, y_i)$ are such that the matrix V has linearly independent columns, with v_1 equal to a column of ones, then $b_1 = \bar{y} - b_2\bar{x}_2 - \cdots - b_k\bar{x}_k$ and

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(b_2(x_{i2} - \bar{x}_2) + \cdots + b_k(x_{ik} - \bar{x}_k))^2 \\ &\quad + \sum_{i=1}^{n}(y_i - b_1x_{i1} - \cdots - b_kx_{ik})^2.\end{aligned}$$

We call

$$\text{RSS}\,(X_2, \ldots, X_k) = \sum_{i=1}^{n}\left(b_2\,(x_{i2} - \bar{x}_2) + \cdots + b_k\,(x_{ik} - \bar{x}_k)\right)^2$$

the regression sum of squares and

$$\text{ESS} = \sum_{i=1}^{n}(y_i - b_1x_{i1} - \cdots - b_kx_{ik})^2$$

the error sum of squares. This leads to the following ANOVA table.

Source	Df	Sum of Squares	Mean Square
$X_2, \ldots, X_k$	$k-1$	$\text{RSS}(X_2, \ldots, X_k)$	$\text{RSS}(X_2, \ldots, X_k)/\,(k-1)$
Error	$n-k$	ESS	s^2
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	

When there is an intercept term, the null hypothesis of no relationship between the response and the predictors is equivalent to $H_0 : \beta_2 = \cdots = \beta_k = 0$. As with the simple linear regression model, the mean square for regression can be shown to be an unbiased estimator of σ^2 if and only if the null hypothesis is true. Therefore, a sensible statistic to use for assessing the null hypothesis is the F-statistic

$$F = \frac{\text{RSS}(X_2, \ldots, X_k)/\,(k-1)}{s^2},$$

with large values being evidence against the null.

Often, we want to assess the null hypothesis $H_0 : \beta_{l+1} = \cdots = \beta_k = 0$ or, equivalently, the hypothesis that the model is given by

$$E\,(Y \mid X_1 = x_1, \ldots, X_k = x_k) = \beta_1x_1 + \cdots + \beta_lx_l,$$

where $l < k$. This hypothesis says that the last $k - l$ predictors $X_{l+1}, \ldots, X_k$, have no relationship with the response.

If we denote the least-squares estimates of $\beta_1, \ldots, \beta_l$, obtained by fitting the smaller model, by $b_1^*, \ldots, b_l^*$, then we have the following result.

Lemma 10.3.3 If the $(x_{i1}, \ldots, x_{ik}, y_i)$ for $i = 1, \ldots, n$ are values for which the matrix V has linearly independent columns, with v_1 equal to a column of ones, then

$$\begin{aligned} \text{RSS}(X_2, \ldots, X_k) &= \sum_{i=1}^{n} (b_2(x_{i2} - \bar{x}_2) + \cdots + b_k(x_{ik} - \bar{x}_k))^2 \\ &\geq \sum_{i=1}^{n} (b_2^*(x_{i2} - \bar{x}_2) + \cdots + b_l^*(x_{il} - \bar{x}_l))^2 \\ &= \text{RSS}(X_2, \ldots, X_l). \end{aligned} \tag{10.3.18}$$

On the right of the inequality in (10.3.18), we have the regression sum of squares obtained by fitting the model based on the first l predictors. Therefore, we can interpret the difference of the left and right sides of (10.3.18), namely,

$$\text{RSS}(X_{l+1}, \ldots, X_k \mid X_2, \ldots, X_l) = \text{RSS}(X_2, \ldots, X_k) - \text{RSS}(X_2, \ldots, X_l)$$

as the contribution of the predictors $X_{l+1}, \ldots, X_k$ to the regression sum of squares when the predictors $X_1, \ldots, X_l$ are in the model. We get the following ANOVA table (actually only the first three columns of the ANOVA table) corresponding to this decomposition of the total sum of squares.

Source	Df	Sum of Squares
$X_2, \ldots, X_l$	$l-1$	$\text{RSS}(X_2, \ldots, X_l)$
$X_{l+1}, \ldots, X_k \mid X_2, \ldots, X_l$	$k-l$	$\text{RSS}(X_{l+1}, \ldots, X_k \mid X_2, \ldots, X_l)$
Error	$n-k$	ESS
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2$

It can be shown that the null hypothesis $H_0 : \beta_{l+1} = \cdots = \beta_k = 0$ holds if and only if

$$\text{RSS}(X_{l+1}, \ldots, X_k \mid X_2, \ldots, X_l) / (k-l)$$

is an unbiased estimator of σ^2. Therefore, a sensible statistic to use for assessing this null hypothesis is the F-statistic

$$F = \frac{\text{RSS}(X_{l+1}, \ldots, X_k \mid X_2, \ldots, X_l) / (k-l)}{s^2},$$

with large values being evidence against the null.

The Coefficient of Determination

The coefficient of determination for this model is given by

$$R^2 = \frac{\text{RSS}(X_2, \ldots, X_k)}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

which, by Lemma 10.3.2, is always between 0 and 1. The value of R^2 gives the proportion of the observed variation in Y that is explained by the inclusion of the nonconstant predictors in the model.

It can be shown that R^2 is the square of the *multiple correlation coefficient* between Y and $X_1, \ldots, X_k$. However, we do not discuss the multiple correlation coefficient in this text.

Confidence Intervals and Testing Hypotheses

For inference, we have the following result.

Theorem 10.3.11 If the conditional distribution of Y given $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ is $N(\beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$, and if we observe the independent values $(x_{i1}, \ldots, x_{ik}, y_i)$ for $i = 1, \ldots, n$, then the conditional distributions of the B_i and S^2, given $X_{ij} = x_{ij}$ for all i, j, are as follows.

(i) $B_i \sim N(\beta_i, \sigma^2 c_{ii})$

(ii) $B_1 x_1 + \cdots + B_k x_k$ is distributed

$$N\left(\beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2\left(\sum_{i=1}^{k} x_i^2 c_{ii} + 2\sum_{i<j} x_i x_j c_{ij}\right)\right)$$

(iii) $(n-k)\,S^2/\sigma^2 \sim \chi^2(n-k)$ independent of $(B_1, \ldots, B_k)$

Corollary 10.3.5

(i) $(B_i - \beta_i)/s c_{ii}^{1/2} \sim t(n-k)$

(ii)

$$\frac{B_1 x_1 + \cdots + B_k x_k - \beta_1 x_1 - \cdots - \beta_k x_k}{S\left(\sum_{i=1}^{k} x_i^2 c_{ii} + 2\sum_{i<j} x_i x_j c_{ij}\right)^{1/2}} \sim t(n-k)$$

(iii) $H_0 : \beta_{l+1} = \cdots = \beta_k = 0$ is true if and only if

$$F = \frac{(\mathrm{RSS}(X_2, \ldots, X_k) - \mathrm{RSS}(X_2, \ldots, X_l))/(k-l)}{S^2} \sim F(k-l, n-k)$$

Analysis of Residuals

In an application of the multiple regression model, we must check to make sure that the assumptions make sense. Model checking is based on the residuals $y_i - b_1 x_{i1} - \cdots - b_k x_{ik}$ (after standardization), just as discussed in Section 9.1. Note that the ith residual is simply the difference between the observed value y_i at $(x_{i1}, \ldots, x_{ik})$ and the predicted value $b_1 x_{i1} + \cdots + b_k x_{ik}$ at $(x_{i1}, \ldots, x_{ik})$.

We also have the following result (this can be proved as a Corollary of Theorem 10.3.10).

Corollary 10.3.6
(i) $E(Y_i - B_1 x_{i1} - \cdots - B_k x_{ik} \mid V) = 0$
(ii) $\text{Cov}(Y_i - B_1 x_{i1} - \cdots - B_k x_{ik}, Y_j - B_1 x_{j1} - \cdots - B_k x_{jk,} \mid V) = \sigma^2 d_{ij}$, where d_{ij} is the (i, j)-th entry of the matrix $I - V(V'V)^{-1}V'$.

Therefore, the standardized residuals are given by

$$\frac{y_j - b_1 x_{j1} - \cdots - b_k x_{jk}}{s d_{ii}^{1/2}}. \tag{10.3.19}$$

When s is replaced by σ in (10.3.19), Corollary 10.3.6 implies that this quantity has conditional mean 0 and conditional variance 1. Furthermore, when the conditional distribution of the response given the predictors is normal, then it can be shown that the conditional distribution of this quantity is $N(0, 1)$. These results are also approximately true for (10.3.19) for large n. Furthermore, it can be shown that the covariances between the standardized residuals go to 0 as $n \to \infty$, under certain reasonable conditions on distribution of the predictor variables. So one approach to model checking here is to see whether the values given by (10.3.19) look at all like a sample from the $N(0, 1)$ distribution.

What do we do if model checking leads to a failure of the model? As in Chapter 9, we can consider making various transformations of the data to see if there is a simple modification of the model that will pass. We can make transformations not only to the response variable Y, but to the predictor variables $X_1, \ldots, X_k$ as well.

An Application of Multiple Linear Regression Analysis

The computations needed to implement a multiple linear regression analysis cannot be carried out by hand. These are much too time-consuming and error-prone. It is therefore important that a statistician have a computer with suitable software available when doing a multiple linear regression analysis.

The data in Table 10.1 are taken from *Statistical Theory and Methodology in Science and Engineering*, 2nd ed., by K. A. Brownlee (John Wiley & Sons, New York, 1965). The response variable Y is stack loss (Loss), which represents 10 times the percentage of ammonia lost as unabsorbed nitric oxide. The predictor variables are $X_1 =$ air flow (Air), $X_2 =$ temperature of inlet water (Temp), and $X_3 =$ the concentration of nitric acid (Acid). Also recorded is the day (Day) on which the observation was taken.

We consider the model $Y \mid x_1, x_2, x_3 \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \sigma^2)$. Note that we have included an intercept term. Figure 10.3.9 is a normal probability plot of the standardized residuals. This looks reasonable, except for one residual, -2.63822, that diverges quite distinctively from the rest of the values, which lie close to the 45-degree line. Printing out the standardized residuals shows that this residual is associated with the observation on the twenty-first day. Possibly there was something unique about this day's operations, and so it is reasonable to discard this data value and refit the model. Figure 10.3.10 is a normal probability plot obtained by fitting the model to the first 20 observations. This looks somewhat better, but still we might be concerned about at least one of the residuals that deviates substantially from the 45-degree line.

Day	Air	Temp	Acid	Loss	Day	Air	Temp	Acid	Loss
1	80	27	89	42	12	58	17	88	13
2	80	27	88	37	13	58	18	82	11
3	75	25	90	37	14	58	19	93	12
4	62	24	87	28	15	50	18	89	8
5	62	22	87	18	16	50	18	86	7
6	62	23	87	18	17	50	19	72	8
7	62	24	93	19	18	50	19	79	8
8	62	24	93	20	19	50	20	80	9
9	58	23	87	15	20	56	20	82	15
10	58	18	80	14	21	70	20	91	15
11	58	18	89	14					

Table 10.1: Data for Application of Multiple Linear Regression Analysis

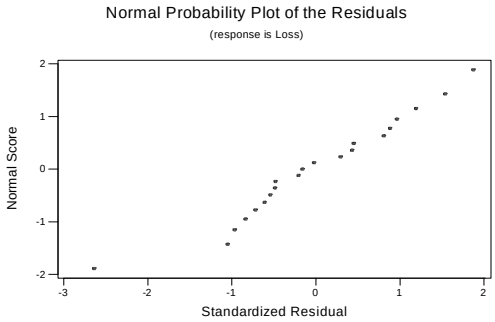

Figure 10.3.9: Normal probability plot of the standardized residuals based on all the data.

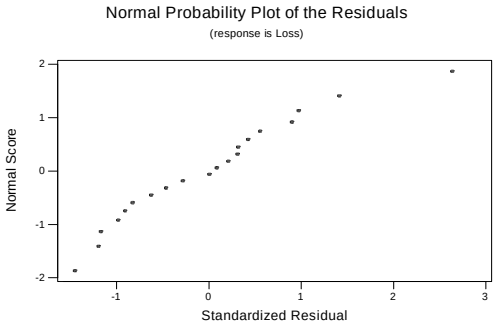

Figure 10.3.10: Normal probability plot of the standardized residuals based on the first 20 data values.

Following the analysis of these data in *Fitting Equations to Data,* by C. Daniel and F. S. Wood (Wiley-Interscience, New York, 1971), we consider instead the model

$$\ln Y \mid x_1, x_2, x_3 \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \sigma^2), \tag{10.3.20}$$

i.e., we transform the response variable by taking its logarithm and use all of the data. Often, when models do not fit, simple transformations like this can lead to major improvements. In this case, we see a much improved normal probability plot, as provided in Figure 10.3.11.

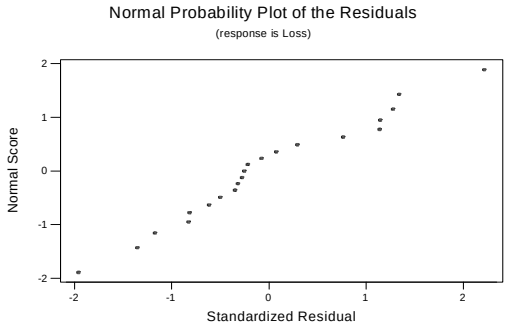

Figure 10.3.11: Normal probability plot of the standardized residuals for all the data using $\ln Y$ as the response.

We also looked at plots of the standardized residuals against the various predictors, and these looked reasonable. Figure 10.3.12 is a plot of the standardized residuals against the values of Air.

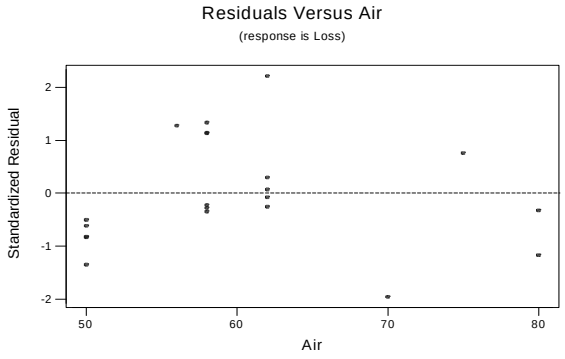

Figure 10.3.12: A plot of the standardized residuals for all the data, using $\ln Y$ as the response, against the values of the predictor Air.

Now that we have accepted the model (10.3.20), we can proceed to inferences about the unknowns of the model. The least-squares estimates of the β_i,their standard errors (Se), the corresponding t-statistics for testing the $\beta_i = 0$, and the P-values for this are given in the following table.

Coefficient	Estimate	Se	t-statistic	P-value
β_0	−0.948700	0.647700	−1.46	0.161
β_1	0.034565	0.007343	4.71	0.000
β_2	0.063460	0.020040	3.17	0.006
β_3	0.002864	0.008510	0.34	0.742

The estimate of σ^2 is given by $s^2 = 0.0312$.

To test the null hypothesis that there is no relationship between the response and the predictors, or that, equivalently, $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, we have the following ANOVA table.

Source	Df	Sum of Squares	Mean Square
X_1, X_2, X_3	3	4.9515	1.6505
Error	17	0.5302	0.0312
Total	20	5.4817	

The value of the F-statistic is given by $1.6505/0.0312 = 52.900$, and when $F \sim F(3, 17)$, we have that $P(F > 52.900) = 0.000$. So there is substantial evidence against the null hypothesis. To see how well the model explains the variation in the response, we computed the value of $R^2 = 86.9\%$. Therefore, approximately 87% of the observed variation in Y can be explained by changes in the predictors in the model.

While we have concluded that a relationship exists between the response and the predictors, it may be that some of the predictors have no relationship with the response. For example, the table of t-statistics above would seem to indicate that perhaps X_3 (acid) is not affecting Y. We can assess this via the following ANOVA table, obtained by fitting the model $\ln Y \mid x_1, x_2, x_3 \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$.

Source	Df	Sum of Squares	Mean Square
X_1, X_2	2	4.9480	2.4740
$X_3 \mid X_1, X_2$	1	0.0035	0.0035
Error	17	0.5302	0.0312
Total	20	5.4817	

Note that $\text{RSS}(X_3 \mid X_1, X_2) = 4.9515 - 4.9480 = 0.0035$. The value of the F-statistic for testing $H_0 : \beta_3 = 0$ is $0.0035/0.0312 = 0.112$, and when $F \sim F(1, 17)$, we have that $P(F > 0.112) = 0.742$. So we have no evidence against the null hypothesis and can drop X_3 from the model. Actually, this is the same P-value as obtained via the t-test of this null hypothesis, as, in general, the t-test that a single regression coefficient is 0 is equivalent to the F-test. Similar tests of the need to include X_1 and X_2 do not lead us to drop these variables from the model.

So based on the above results, we decide to drop X_3 from the model and use the equation

$$E(Y \mid X_1 = x_1, X_2 = x_2) = -0.7522 + 0.035402X_1 + 0.06346X_2 \qquad (10.3.21)$$

to describe the relationship between Y and the predictors. Note that the least-squares estimates of β_0, β_1, and β_2 in (10.3.21) are obtained by refitting the model without X_3.

Summary of Section 10.3

- In this section, we examined the situation in which the response variable and the predictor variables are quantitative.
- In this situation, the linear regression model provides a possible description of the form of any relationship that may exist between the response and the predictors.
- Least squares is a standard method for fitting linear regression models to data.
- The ANOVA is a decomposition of the total variation observed in the response variable into a part attributable to changes in the predictor variables and a part attributable to random error.
- If we assume a normal linear regression model, then we have inference methods available such as confidence intervals and tests of significance. In particular, we have available the F-test to assess whether or not a relationship exists between the response and the predictors.
- A normal linear regression model is checked by examining the standardized residuals.

EXERCISES

10.3.1 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown. What is the least-squares estimate of the mean of this distribution?

10.3.2 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Uniform$[0, \theta]$, where $\theta > 0$ is unknown. What is the least-squares estimate of the mean of this distribution?

10.3.3 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Exponential(θ), where $\theta > 0$ is unknown. What is the least-squares estimate of the mean of this distribution?

10.3.4 Consider the $n = 11$ data values in the following table.

Observation	X	Y	Observation	X	Y
1	−5.00	−10.00	7	1.00	3.52
2	−4.00	−8.83	8	2.00	5.64
3	−3.00	−9.15	9	3.00	7.28
4	−2.00	−4.26	10	4.00	7.62
5	−1.00	−0.30	11	5.00	8.51
6	0.00	−0.04			

Suppose we consider the simple normal linear regression to describe the relationship between the response Y and the predictor X.

(a) Plot the data in a scatter plot.

(b) Calculate the least-squares line and plot this on the scatter plot in part (a).

(c) Plot the standardized residuals against X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) If appropriate, calculate 0.95-confidence intervals for the intercept and slope.

(g) Construct the ANOVA table to test whether or not there is a relationship between the response and the predictors. What is your conclusion?

(h) If the model is correct, what proportion of the observed variation in the response is explained by changes in the predictor?

(i) Predict a future Y at $X = 0.0$. Is this prediction an extrapolation or an interpolation? Determine the standard error of this prediction.

(j) Predict a future Y at $X = 6.0$. Is this prediction an extrapolation or an interpolation? Determine the standard error of this prediction.

(k) Predict a future Y at $X = 20.0$. Is this prediction an extrapolation or an interpolation? Determine the standard error of this prediction. Compare this with the standard errors obtained in parts (i) and (j) and explain the differences.

10.3.5 Consider the $n = 11$ data values in the following table.

Observation	X	Y	Observation	X	Y
1	−5.00	65.00	7	1.00	6.52
2	−4.00	39.17	8	2.00	17.64
3	−3.00	17.85	9	3.00	34.28
4	−2.00	7.74	10	4.00	55.62
5	−1.00	2.70	11	5.00	83.51
6	0.00	−0.04			

Suppose we consider the simple normal linear regression to describe the relationship between the response Y and the predictor X.

(a) Plot the data in a scatter plot.

(b) Calculate the least-squares line and plot this on the scatter plot in part (a).

(c) Plot the standardized residuals against X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) If appropriate, calculate 0.95-confidence intervals for the intercept and slope.

(g) Do the results of your analysis allow you to conclude that there is a relationship between Y and X? Explain why or why not.

(h) If the model is correct, what proportion of the observed variation in the response is explained by changes in the predictor?

10.3.6 Suppose the following data record the densities of an organism in a containment vessel for 10 days. Suppose we consider the simple normal linear regression to describe the relationship between the response Y (density) and the predictor X (day).

Day	Number/Liter	Day	Number/Liter
1	1.6	6	1341.6
2	16.7	7	2042.9
3	65.2	8	7427.0
4	23.6	9	15571.8
5	345.3	10	33128.5

(a) Plot the data in a scatter plot.

(b) Calculate the least-squares line and plot this on the scatter plot in part (a).

(c) Plot the standardized residuals against X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) Can you think of a transformation of the response that might address any problems found? If so, repeat parts (a) through (e) after performing this transformation. (Hint: The scatter plot looks like exponential growth. What transformation is the inverse of exponentiation?)

(g) Calculate 0.95-confidence intervals for the appropriate intercept and slope.

(h) Construct the appropriate ANOVA table to test whether or not there is a relationship between the response and the predictors. What is your conclusion?

(i) Do the results of your analysis allow you to conclude that there is a relationship between Y and X? Explain why or why not.

(j) Compute the proportion of variation explained by the predictor for the two models you have considered. Compare the results.

(k) Predict a future Y at $X = 12$. Is this prediction an extrapolation or an interpolation?

10.3.7 A student takes weekly quizzes in a course and receives the following grades over 12 weeks.

Week	Grade	Week	Grade
1	65	7	74
2	55	8	76
3	62	9	48
4	73	10	80
5	68	11	85
6	76	12	90

(a) Plot the data in a scatter plot with $X =$ week and $Y =$ grade.

(b) Calculate the least-squares line and plot this on the scatter plot in part (a).

(c) Plot the standardized residuals against X.

(d) What are your conclusions based on the plot produced in (c)?

(e) Calculate 0.95-confidence intervals for the intercept and slope.

(f) Construct the ANOVA table to test whether or not there is a relationship between the response and the predictors. What is your conclusion?

(g) What proportion of the observed variation in the response is explained by changes in the predictor?

10.3.8 Suppose that $Y = E(Y \mid X) + Z$, where X, Y and Z are random variables.

(a) Show that $E(Z \mid X) = 0$.

(b) Show that $\text{Cov}(E(Y \mid X), Z) = 0$. (Hint: Write $Z = Y - E(Y \mid X)$ and use Theorems 3.5.2 and 3.5.4.)

(c) Suppose that Z is independent of X. Show that this implies that the conditional distribution of Y given X depends on X only through its conditional mean. (Hint: Evaluate the conditional distribution function of Y given $X = x$.)

10.3.9 Suppose that X and Y are random variables such that a regression model describes the relationship between Y and X. If $E(Y \mid X) = \exp\{\beta_1 + \beta_2 X\}$, then discuss whether or not this is a simple linear regression model (perhaps involving a predictor other than X).

10.3.10 Suppose that X and Y are random variables and $\text{Corr}(X, Y) = 1$. Does a simple linear regression model hold to describe the relationship between Y and X? If so, what is it?

10.3.11 Suppose that X and Y are random variables such that a regression model describes the relationship between Y and X. If $E(Y \mid X) = \beta_1 + \beta_2 X^2$, then discuss whether or not this is a simple linear regression model (perhaps involving a predictor other than X).

10.3.12 Suppose that $X \sim N(2, 3)$ independently of $Z \sim N(0, 1)$ and $Y = X + Z$. Does this structure imply that the relationship between Y and X can be summarized by a simple linear regression model? If so, what are β_1, β_2, and σ^2?

10.3.13 Suppose that a simple linear model is fit to data. An analysis of the residuals indicates that there is no reason to doubt that the model is correct; the ANOVA test indicates that there is substantial evidence against the null hypothesis of no relationship between the response and predictor. The value of R^2 is found to be 0.05. What is the interpretation of this number and what are the practical consequences?

COMPUTER EXERCISES

10.3.14 Suppose we consider the simple normal linear regression to describe the relationship between the response Y (income) and the predictor X (investment) for the data in Example 10.3.9.

(a) Plot the data in a scatter plot.

(b) Calculate the least-squares line and plot this on the scatter plot in part (a).

(c) Plot the standardized residuals against X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) If appropriate, calculate 0.95-confidence intervals for the intercept and slope.

(g) Do the results of your analysis allow you to conclude that there is a relationship between Y and X? Explain why or why not.

(h) If the model is correct, what proportion of the observed variation in the response is explained by changes in the predictor?

10.3.15 The following data are measurements of tensile strength (100 lb/in^2) and hardness (Rockwell E) on 20 pieces of die-cast aluminum.

Sample	Strength	Hardness	Sample	Strength	Hardness
1	293	53	11	298	60
2	349	70	12	292	51
3	340	78	13	380	95
4	340	55	14	345	88
5	340	64	15	257	51
6	354	71	16	265	54
7	322	82	17	246	52
8	334	67	18	286	64
9	247	56	19	324	83
10	348	86	20	282	56

Suppose we consider the simple normal linear regression to describe the relationship between the response Y (strength) and the predictor X (hardness).

(a) Plot the data in a scatter plot.

(b) Calculate the least-squares line and plot this on the scatter plot in part (a).

(c) Plot the standardized residuals against X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) If appropriate, calculate 0.95-confidence intervals for the intercept and slope.

(g) Do the results of your analysis allow you to conclude that there is a relationship between Y and X? Explain why or why not.

(h) If the model is correct, what proportion of the observed variation in the response is explained by changes in the predictor?

10.3.16 Tests were carried out to determine the effect of gas inlet temperature (degrees Fahrenheit) and rotor speed (rpm) on the tar content (grains/cu ft) of a gas stream, producing the following data.

Observation	Tar	Speed	Temperature
1	60.0	2400	54.5
2	65.0	2450	58.5
3	63.5	2500	58.0
4	44.0	2700	62.5
5	54.5	2700	68.0
6	26.0	2775	45.5
7	54.0	2800	63.0
8	53.5	2900	64.5
9	33.5	3075	57.0
10	44.0	3150	64.0

Suppose we consider the normal linear regression model

$$Y \mid W = w, X = x \sim N(\beta_1 + \beta_2 w + \beta_3 x, \sigma^2)$$

to describe the relationship between Y (tar content) and the predictors W (rotor speed) and X (temperature).

(a) Plot the response in scatter plots against each predictor.

(b) Calculate the least-squares equation.

(c) Plot the standardized residuals against W and X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) If appropriate, calculate 0.95-confidence intervals for the regression coefficients.

(g) Construct the ANOVA table to test whether or not there is a relationship between the response and the predictors. What is your conclusion?

(h) If the model is correct, what proportion of the observed variation in the response is explained by changes in the predictors?

(i) In an ANOVA table, assess the null hypothesis that there is no effect due to W, given that X is in the model.

(j) Estimate the mean of Y when $W = 2750$ and $X = 50.0$. If we consider this value as a prediction of a future Y at these settings, is this an extrapolation or interpolation?

10.3.17 Suppose we consider the normal linear regression model

$$Y \mid X = x \sim N(\beta_1 + \beta_2 x + \beta_3 x^2, \sigma^2)$$

for the data of Exercise 10.3.5.

(a) Plot the response Y in a scatter plot against X.

(b) Calculate the least-squares equation.

(c) Plot the standardized residuals against X.

(d) Produce a normal probability plot of the standardized residuals.

(e) What are your conclusions based on the plots produced in parts (c) and (d)?

(f) If appropriate, calculate 0.95-confidence intervals for the regression coefficients.

(g) Construct the ANOVA table to test whether or not there is a relationship between the response and the predictor. What is your conclusion?

(h) If the model is correct, what proportion of the observed variation in the response is explained by changes in the predictors?

(i) In an ANOVA table, assess the null hypothesis that there is no effect due to X^2, given that X is in the model.

(j) Compare the predictions of Y at $X = 6$ using the simple linear regression model and using the linear model with a linear and quadratic term.

PROBLEMS

10.3.18 Suppose that $(x_1, \ldots, x_n)$ is a sample from the mixture distribution

$$0.5\text{Uniform}[0, 1] + 0.5\text{Uniform}[2, \theta],$$

where $\theta > 2$ is unknown. What is the least-squares estimate of the mean of this distribution?

10.3.19 Consider the simple linear regression model and suppose that for the data collected, we have $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$. Explain how, and for which value of x, you would estimate $E(Y \mid X = x)$.

10.3.20 For the simple linear regression model, under the assumptions of Theorem 10.3.3, establish that

$$\begin{aligned}&\text{Cov}(Y_i - B_1 - B_2 x_i, Y_j - B_1 - B_2 x_j \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \sigma^2 \delta_{ij} - \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})\left(x_j - \bar{x}\right)}{\sum_{k=1}^{n}(x_k - \bar{x})^2} \right),\end{aligned}$$

where $\delta_{ij} = 1$ when $i = j$ and is 0 otherwise. (Hint: Use Theorems 3.3.2 and 10.3.3.)

10.3.21 Establish that (10.3.11) is distributed $N(0, 1)$ when S is replaced by σ in the denominator. (Hint: Use Theorem 4.6.1 and Problem 10.3.20.)

10.3.22 (*Prediction intervals*) Under the assumptions of Theorem 10.3.6, prove that the interval

$$b_1 + b_2 x \pm s \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} \right)^{1/2} t_{(1+\gamma)/2}(n-2),$$

based on independent $(x_1, y_1), \ldots, (x_n, y_n)$, will contain Y with probability equal to γ for a future independent (X, Y) with $X = x$. (Hint: Theorems 4.6.1 and 3.3.2 and Corollary 10.3.1.)

10.3.23 Consider the regression model with no intercept, given by $E(Y \mid X = x) = \beta x$,where $\beta \in R^1$ is unknown. Suppose we observe the independent values $x_1, y_1)$, $\ldots, (x_n, y_n)$.

(a) Determine the least-squares estimate of β.

(b) Prove that the least-squares estimate b of β is unbiased and, when $\text{Var}(Y \mid X = x) = \sigma^2$, prove that

$$\text{Var}(B \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

(c) Under the assumptions given in part (b), prove that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - b x_i)^2$$

is an unbiased estimator of σ^2.

(d) Record an appropriate ANOVA decomposition for this model and a formula for R^2, measuring the proportion of the variation observed in Y due to changes in X.

(e) When $Y \mid X = x \sim N(\beta x, \sigma^2)$, and we observe the independent values (x_1, y_1), $\ldots, (x_n, y_n)$, prove that $b \sim N(\beta, \sigma^2 / \sum_{i=1}^{n} x_i^2)$.

(f) Under the assumptions of part (e), and assuming that $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ independent of B (this can be proved), indicate how you would test the null hypothesis of no relationship between Y and X.

(g) How would you define standardized residuals for this model and use them to check model validity?

10.3.24 For data $(x_1, y_1), \ldots, (x_n, y_n)$, prove that if $\alpha_1 = \beta_1 + \beta_2 \bar{x}$ and $\alpha_2 = \beta_2$, then $\sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2$ equals

$$\sum_{i=1}^n (y_i - \bar{y})^2 + n(\alpha_1 - \bar{y})^2 + \alpha_2^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2\alpha_2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

From this, deduce that $\bar{y}$ and $a = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$ are the least squares of α_1 and α_2, respectively.

10.3.25 For the model discussed in Section 10.3.3, prove that the prior given by $\alpha_1 \mid \alpha_2$, $\sigma^2 \sim N(\mu_1, \tau_1^2 \sigma^2)$, $\alpha_2 \mid \sigma^2 \sim N(\mu_2, \tau_2^2 \sigma^2)$, and $1/\sigma^2 \sim \text{Gamma}(\kappa, \nu)$. Conclude that this prior is conjugate with the posterior distribution, as specified. (Hint: The development is similar to Example 7.1.4, as detailed in Section 7.5.)

10.3.26 For the model specified in Section 10.3.3, prove that when $\tau_1 \to \infty, \tau_2 \to \infty$, and $\nu \to 0$, the posterior distribution of α_1 is given by the distribution of $\bar{y} + \left(2\nu_{xy}/n\,(2\kappa + n)\right)^{1/2} Z$, where $Z \sim t(2\kappa + n)$ and $\nu_{xy} = (c_y^2 - a^2 c_x^2)/2$.

CHALLENGES

10.3.27 If $X_1, \ldots, X_n$ is a sample from a distribution with finite variance, then prove that

$$\frac{X_i - \bar{X}}{\sqrt{\sum_{k=1}^n \left(X_k - \bar{X}\right)^2}} \xrightarrow{a.s.} 0.$$

10.4 Quantitative Response and Categorical Predictors

In this section, we consider the situation in which the response is quantitative and the predictors are categorical. There can be many categorical predictors, but we restrict our discussion to at most two, as this gives the most important features of the general case. The general case is left to a further course.

10.4.1 One Categorical Predictor (One-Way ANOVA)

Suppose now that the response Y is quantitative and the predictor X is categorical, taking a values or levels denoted $1, \ldots, a$. With the regression model, we assume that the only aspect of the conditional distribution of Y, given $X = x$, that changes as x changes, is the mean. We let

$$\beta_i = E(Y \mid X = i)$$

denote the mean response when the predictor X is at level i. Note that this is immediately a linear regression model.

We introduce the *dummy variables*

$$X_i = \begin{cases} 1 & X = i \\ 0 & X \neq i \end{cases}$$

for $i = 1, \ldots, a$. Notice that, whatever the value is of the response Y, only one of the dummy variables takes the value 1, and the rest take the value 0. Accordingly, we can write

$$E(Y \mid X_1 = x_1, \ldots, X_a) = x_a = \beta_1 x_1 + \cdots + \beta_a x_a,$$

because one and only one of the $x_i = 1$, whereas the rest are 0. This has exactly the same form as the model discussed in Section 10.3.4, as the X_i are quantitative. As such, all the results of Section 10.3.4 immediately apply (we will restate relevant results here).

Inferences About Individual Means

Now suppose that we observe n_i values $\left(y_{i1}, \ldots, y_{in_i}\right)$ when $X = i$, and all the response values are independent. Note that we have a independent samples. The least-squares estimates of the β_i are obtained by minimizing

$$\sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \beta_i)^2.$$

The least-squares estimates are then equal to (see Problem 10.4.14)

$$b_i = \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

These can be shown to be unbiased estimators of the β_i.

Assuming that the conditional distributions of Y, given $X = x$, all have variance equal to σ^2, we have that the conditional variance of $\bar{Y}_i$ is given by σ^2/n_i, and the conditional covariance between $\bar{Y}_i$ and $\bar{Y}_j$, when $i \neq j$, is 0. Furthermore, under these conditions, an unbiased estimator of σ^2 is given by

$$s^2 = \frac{1}{N - a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,$$

where $N = n_1 + \cdots + n_k$.

If, in addition, we assume the normal linear regression model, namely,

$$Y \mid X = i \sim N(\beta_i, \sigma^2),$$

then $\bar{Y}_i \sim N(\beta_i, \sigma^2/n_i)$ independent of $(N - a)\, S^2/\sigma^2 \sim \chi^2(N - a)$. Therefore, by Definition 4.6.2,

$$T = \frac{\bar{Y}_i - \beta_i}{S/\sqrt{n_i}} \sim t(N - a),$$

which leads to a γ-confidence interval of the form

$$\bar{y}_i \pm \frac{s}{\sqrt{n_i}} t_{(1+\gamma)/2}(N-a)$$

for β_i. Also, we can test the null hypothesis $H_0 : \beta_i = \beta_{i0}$ by computing the P-value

$$P\left(|T| \geq \left|\frac{\bar{y}_i - \beta_{i0}}{s/\sqrt{n_i}}\right|\right) = 2\left(1 - G\left(\left|\frac{\bar{y}_i - \beta_{i0}}{s/\sqrt{n_i}}\right| \; ; \; N-a\right)\right),$$

where $G(\cdot \; ; \; N-a)$ is the cdf of the $t(N-a)$ distribution. Note that these inferences are just like those derived in Section 6.3 for the location-scale normal model, except we now use a different estimator of σ^2 (with more degrees of freedom).

Inferences about Differences of Means and Two Sample Inferences

Often we want to make inferences about a difference of means $\beta_i - \beta_j$. Note that $E(\bar{Y}_i - \bar{Y}_j) = \beta_i - \beta_j$ and

$$\text{Var}(\bar{Y}_i - \bar{Y}_j) = \text{Var}(\bar{Y}_i) + \text{Var}(\bar{Y}_j) = \sigma^2(1/n_i + 1/n_j)$$

because $\bar{Y}_i$ and $\bar{Y}_j$ are independent. By Theorem 4.6.1,

$$\bar{Y}_i - \bar{Y}_j \sim N(\beta_i - \beta_j, \sigma^2(1/n_i + 1/n_j)).$$

Furthermore,

$$\frac{(\bar{Y}_i - \bar{Y}_j) - (\beta_i - \beta_j)}{\sigma(1/n_i + 1/n_j)^{1/2}} \sim N(0,1)$$

independent of $(N-a)\,S^2/\sigma^2 \sim \chi^2(N-a)$. Therefore, by Definition 4.6.2,

$$\begin{aligned} T &= \left(\frac{(\bar{Y}_i - \bar{Y}_j) - (\beta_i - \beta_j)}{\sigma(1/n_i + 1/n_j)^{1/2}}\right) \Big/ \sqrt{\frac{(N-a)\,S^2}{(N-a)\,\sigma^2}} \\ &= \frac{(\bar{Y}_i - \bar{Y}_j) - (\beta_i - \beta_j)}{S(1/n_i + 1/n_j)^{1/2}} \sim t(N-a). \end{aligned} \tag{10.4.1}$$

This leads to the γ-confidence interval

$$\bar{y}_i - \bar{y}_j \pm s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\, t_{(1+\gamma)/2}(N-a)$$

for the difference of means $\beta_i - \beta_j$. We can test the null hypothesis $H_0 : \beta_i = \beta_j$, i.e., that the difference in the means equals 0, by computing the P-value

$$P\left(|T| \geq \left|\frac{\bar{y}_i - \bar{y}_j}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}\right|\right) = 2\left(1 - G\left(\left|\frac{\bar{y}_i - \bar{y}_j}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}\right| \; ; \; N-a\right)\right).$$

When $a = 2$, i.e., there are just two values for X, we refer to (10.4.1) as the *two-sample t-statistic*, and the corresponding inference procedures are called the *two-sample t-confidence interval* and the *two-sample t-test* for the difference of means. In this case, if we conclude that $\beta_1 \neq \beta_2$, then we are saying that a relationship exists between Y and X.

The ANOVA for Assessing a Relationship with the Predictor

Suppose, in the general case when $a \geq 2$, we are interested in assessing whether or not there is a relationship between the response and the predictor. There is no relationship if and only if all the conditional distributions are the same; this is true, under our assumptions, if and only if $\beta_1 = \cdots = \beta_a$, i.e., if and only if all the means are equal. So testing the null hypothesis that there is no relationship between the response and the predictor is equivalent to testing the null hypothesis $H_0 : \beta_1 = \cdots = \beta_a = \beta$ for some unknown β.

If the null hypothesis is true, the least-squares estimate of β is given by $\bar{y}$, the overall average response value. In this case, we have that the total variation decomposes as (see Problem 10.4.15)

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

and so the relevant ANOVA table for testing H_0 is given below.

Source	Df	Sum of Squares	Mean Square
X	$a-1$	$\sum_{i=1}^{a} n_i\left(\bar{y}_i-\bar{y}\right)^2$	$\sum_{i=1}^{a} n_i\left(\bar{y}_i-\bar{y}\right)^2/(a-1)$
Error	$N-a$	$\sum_{i=}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$	s^2
Total	$N-1$	$\sum_{i=}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2$	

To assess H_0, we use the F-statistic

$$F = \frac{\sum_{i=1}^{a} n_i\left(\bar{y}_i-\bar{y}\right)^2/(a-1)}{s^2}$$

because, under the null hypothesis, both the numerator and the denominator are unbiased estimators of σ^2. When the null hypothesis is false, the numerator tends to be larger than σ^2. When we add the normality assumption, we have that $F \sim F(a-1, N-a)$, and so we compute the P-value

$$P\left(F > \frac{\sum_{i=1}^{a} n_i\left(\bar{y}_i-\bar{y}\right)^2/(a-1)}{s^2}\right)$$

to assess whether the observed value of F is so large as to be surprising. Note that when $a = 2$, this P-value equals the P-value obtained via the two-sample t-test.

Multiple Comparisons

If we reject the null hypothesis of no differences among the means, then we want to see where the differences exist. For this, we use inference methods based on (10.4.1). Of course, we have to worry about the problem of *multiple comparisons*, as discussed in Section 9.3. Recall that this problem arises whenever we are testing many null hypotheses using a specific critical value, such as 5%, as a cutoff for a P-value, to decide whether or not a difference exists. The cutoff value for an individual P-value is referred to as the *individual error rate*. In effect, even if no differences exist, the probability of concluding that at least one difference exists, the *family error rate*, can be quite high.

There are a number of procedures designed to control the family error rate when making multiple comparisons. The simplest is to lower the individual error rate, as the family error rate is typically an increasing function of this quantity. This is the approach we adopt here, and we rely on statistical software to compute and report the family error rate for us. We refer to this procedure as *Fisher's multiple comparison test*.

Model Checking

To check the model, we look at the standardized residuals (see Problem 10.4.17) given by

$$\frac{y_{ij} - \bar{y}_i}{s\sqrt{1 - \frac{1}{n_i}}}. \tag{10.4.2}$$

We will restrict our attention to various plots of the standardized residuals for model checking.

We now consider an example.

EXAMPLE 10.4.1
A study was undertaken to determine whether or not eight different types of fat are absorbed in different amounts during the cooking of donuts. Results were collected based on cooking six different donuts and then measuring the amount of fat in grams absorbed. We take the variable X to be the type of fat and use the model of this section.

The collected data are presented in the following table.

Fat 1	164	177	168	156	172	195
Fat 2	172	197	167	161	180	190
Fat 3	177	184	187	169	179	197
Fat 4	178	196	177	181	184	191
Fat 5	163	177	144	165	166	178
Fat 6	163	193	176	172	176	178
Fat 7	150	179	146	141	169	183
Fat 8	164	169	155	149	170	167

A normal probability plot of the standardized residuals is provided in Figure 10.4.1. A plot of the standardized residuals against type of fat is provided in Figure 10.4.2.

Neither plot gives us great grounds for concern over the validity of the model, although there is some indication of a difference in the variability of the response as the type of fat changes. Another useful plot in this situation is a side-by-side boxplot, as it shows graphically where potential differences may lie. Such a plot is provided in Figure 10.4.3.

The following table gives the mean amounts of each fat absorbed.

Fat 1	Fat 2	Fat 3	Fat 4	Fat 5	Fat 6	Fat 7	Fat 8
172.00	177.83	182.17	184.50	165.50	176.33	161.33	162.33

The grand mean response is given by 172.8.

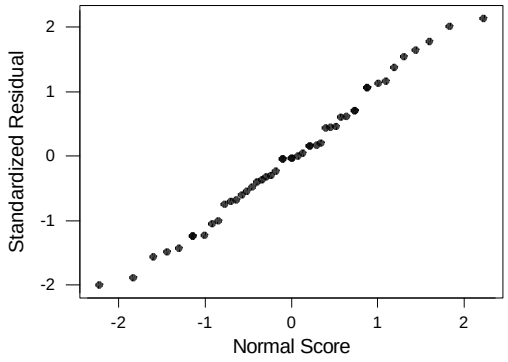

Figure 10.4.1: Normal probability plot of the standardized residuals in Example 10.4.1.

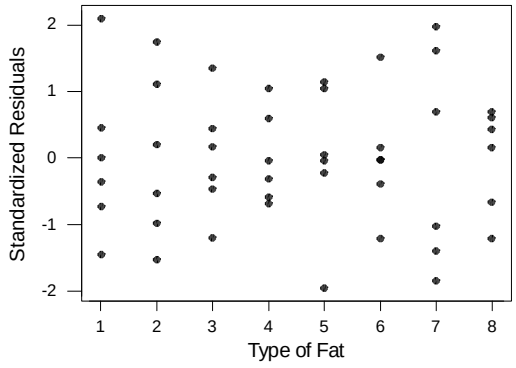

Figure 10.4.2: Standardized residuals versus type of fat in Example 10.4.1.

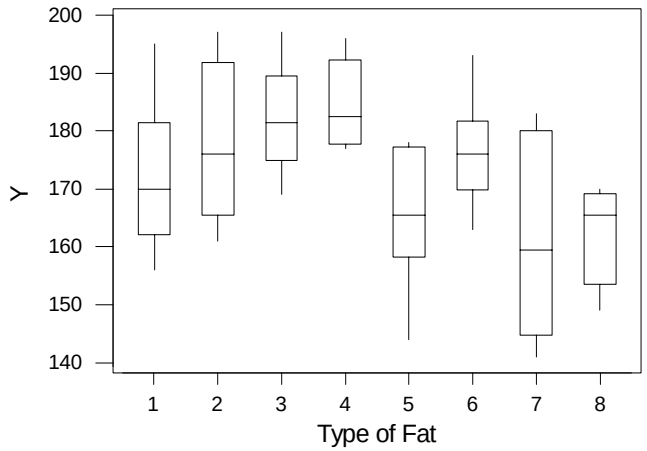

Figure 10.4.3: Side-by-side boxplots of the response versus type of fat in Example 10.4.1.

To assess the null hypothesis of no differences among the types of fat, we calculate the following ANOVA table.

Source	Df	Sum of Squares	Mean Square
X	7	3344	478
Error	40	5799	145
Total	47	9143	

Then we use the F-statistic given by $F = 478/145 = 3.3$. Because $F \sim F(7, 40)$ under H_0, we obtain the P-value $P(F > 3.3) = 0.007$. Therefore, we conclude that there is a difference among the fat types at the 0.05 level.

To ascertain where the differences exist, we look at all pairwise differences. There are $8 \cdot 7/2 = 28$ such comparisons. If we use the 0.05 level to determine whether or not a difference among means exists, then software computes the family error rate as 0.481, which seems uncomfortably high. When we use the 0.01 level, the family error rate falls to 0.151. With the individual error rate at 0.003, the family error rate is 0.0546. Using the individual error rate of 0.003, the only differences detected among the means are those between Fat 4 and Fat 7, and Fat 4 and Fat 8. Note that Fat 4 has the highest absorption whereas Fats 7 and 8 have the lowest absorptions.

Overall, the results are somewhat inconclusive, as we see some evidence of differences existing, but we are left with some anomalies as well. For example, Fats 4 and 5 are not different and neither are Fats 7 and 5, but Fats 4 and 7 are deemed to be different. To resolve such conflicts requires either larger sample sizes or a more refined experiment so that the comparisons are more accurate. ∎

10.4.2 Repeated Measures (Paired Comparisons)

Consider k quantitative variables $Y_1, \ldots, Y_k$ defined on a population Π. Suppose that our purpose is to compare the distributions of these variables. Typically, these will be similar variables, all measured in the same units.

EXAMPLE 10.4.2
Suppose that Π is a set of students enrolled in a first-year program requiring students to take both calculus and physics, and we want to compare the marks achieved in these subjects. If we let Y_1 denote the calculus grade and Y_2 denote the physics grade, then we want to compare the distributions of these variables. ∎

EXAMPLE 10.4.3
Suppose we want to compare the distributions of the duration of headaches for two treatments (A and B) in a population of migraine headache sufferers. We let Y_1 denote the duration of a headache after being administered treatment A, and let Y_2 denote the duration of a headache after being administered treatment B. ∎

The *repeated-measures* approach to the problem of comparing the distributions of $Y_1, \ldots, Y_k$, involves taking a random sample $\pi_1, \ldots, \pi_n$ from Π and, for each π_i, obtaining the k-dimensional value $(Y_1(\pi_i), \ldots, Y_k(\pi_i)) = (y_{i1}, \ldots, y_{ik})$. This gives a sample of n from a k-dimensional distribution. Obviously, this is called repeated measures because we are taking the measurements $Y_1(\pi_i), \ldots, Y_k(\pi_i)$ on the same π_i.

An alternative to repeated measures is to take k independent samples from Π and, for each of these samples, to obtain the values of one and only one of the variables Y_i. There is an important reason why the repeated-measures approach is preferred: We expect less variation in the values of differences, like $Y_i - Y_j$, under repeated-measures sampling, than we do under independent sampling because the values $Y_1(\pi), \ldots, Y_k(\pi)$ are being taken on the *same* member of the population in repeated measures.

To see this more clearly, suppose all of the variances and covariances exist for the joint distribution of $Y_1, \ldots, Y_k$. This implies that

$$\text{Var}(Y_i - Y_j) = \text{Var}(Y_i) + \text{Var}(Y_j) - 2\,\text{Cov}(Y_i, Y_j). \tag{10.4.3}$$

Because Y_i and Y_j are similar variables, being measured on the same individual, we expect them to be positively correlated. Now with independent sampling, we have that $\text{Var}(Y_i - Y_j) = \text{Var}(Y_i) + \text{Var}(Y_i)$, so the variances of differences should be smaller with repeated measures than with independent sampling.

When we assume that the distributions of the Y_i differ at most in their means, then it makes sense to make inferences about the differences of the population means $\mu_i - \mu_j$, using the differences of the sample means $\bar{y}_i - \bar{y}_j$. In the repeated-measures context, we can write

$$\bar{y}_i - \bar{y}_j = \frac{1}{n}\sum_{l=1}^{n}(y_{li} - y_{lj}).$$

Because the individual components of this sum are independent and so,

$$\text{Var}(\bar{Y}_i - \bar{Y}_j) = \frac{\text{Var}(Y_i) + \text{Var}(Y_j) - 2\,\text{Cov}(Y_i, Y_j)}{n}.$$

We can consider the differences $d_1 = y_{1i} - y_{1j}, \ldots, d_n = y_{ni} - y_{nj}$ to be a sample of n from a one-dimensional distribution with mean $\mu_i - \mu_j$ and variance σ^2 given by (10.4.3). Accordingly, we estimate $\mu_i - \mu_j$ by $\bar{d} = \bar{y}_i - \bar{y}_j$ and estimate σ^2 by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2. \tag{10.4.4}$$

If we assume that the joint distribution of $Y_1, \ldots, Y_k$ is multivariate normal (this means that any linear combination of these variables is normally distributed — see Problem 9.1.18), then this forces the distribution of $Y_i - Y_j$ to be $N(\mu_i - \mu_j, \sigma^2)$. Accordingly, we have all the univariate techniques discussed in Chapter 6 for inferences about $\mu_i - \mu_j$.

The discussion so far has been about whether the distributions of variables differed. Assuming these distributions differ at most in their means, this leads to a comparison of the means. We can, however, record an observation as (X, Y), where X takes values in $\{1, \ldots, k\}$ and $X = i$ means that $Y = Y_i$. Then the conditional distribution of Y given $X = i$ is the same as the distribution of Y_i. Therefore, if we conclude that the distributions of the Y_i are different, we can conclude that a relationship exists between Y and X. In Example 10.4.2, this means that a relationship exists between a student's grade and whether or not the grade was in calculus or physics. In Example 10.4.3, this means that a relationship exists between length of a headache and the treatment.

When can we assert that such a relationship is in fact a cause–effect relationship? Applying the discussion in Section 10.1.2, we know that we have to be able to assign the value of X to a randomly selected element of the population. In Example 10.4.2, we see this is impossible, so we cannot assert that such a relationship is a cause–effect relationship. In Example 10.4.3, however, we can indeed do this — namely, for a randomly selected individual, we randomly assign a treatment to the first headache experienced during the study period and then apply the other treatment to the second headache experienced during the study period.

A full discussion of repeated measures requires more advanced concepts in statistics. We restrict our attention now to the presentation of an example when $k = 2$, which is commonly referred to as *paired comparisons*.

EXAMPLE 10.4.4 *Blood Pressure Study*
The following table came from a study of the effect of the drug captopril on blood pressure, as reported in *Applied Statistics, Principles and Examples* by D. R. Cox and E. J. Snell (Chapman and Hall, London, 1981). Each measurement is the difference in the systolic blood pressure before and after having been administered the drug.

−9	−4	−21	−3	−20
−31	−17	−26	−26	−10
−23	−33	−19	−19	−23

Figure 10.4.4 is a normal probability plot for these data and, because this looks reasonable, we conclude that the inference methods based on the assumption of normality are acceptable. Note that here we have not standardized the variable first, so we are only looking to see if the plot is reasonably straight.

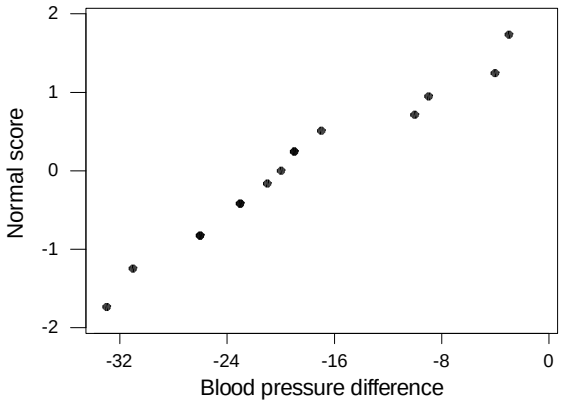

Figure 10.4.4: Normal probability plot for the data in Example 10.4.4.

The mean difference is given by $\bar{d} = -18.93$ with standard deviation $s = 9.03$. Accordingly, the standard error of the estimate of the difference in the means, using (10.4.4), is given by $s/\sqrt{15} = 2.33$. A 0.95-confidence interval for the difference in the mean systolic blood pressure, before and after being administered captopril, is then

$$\bar{d} \pm \frac{s}{\sqrt{n}}\, t_{0.975}(n-1) = -18.93 \pm 2.33\, t_{0.975}(14) = (-23.93, -13.93).$$

Because this does not include 0, we reject the null hypothesis of no difference in the means at the 0.05 level. The actual P-value for the two-sided test is given by

$$P(|T| > |-18.93/2.33|) = 0.000$$

because $T \sim t(14)$ under the null hypothesis H_0 that the means are equal. Therefore, we have strong evidence against H_0. It seems that we have strong evidence that the drug is leading to a drop in blood pressure. ∎

10.4.3 Two Categorical Predictors (Two-Way ANOVA)

Now suppose that we have a single quantitative response Y and two categorical predictors A and B, where A takes a levels and B takes b levels. One possibility is to consider running two one-factor studies. One study will examine the relationship between Y and A, and the second study will examine the relationship between Y and B. There are several disadvantages to such an approach, however.

First, and perhaps foremost, doing two separate analyses will not allow us to determine the joint relationship A and B have with Y. This relates directly to the concept

of *interaction* between predictors. We will soon define this concept more precisely, but basically, if A and B interact, then the conditional relationship between Y and A, given $B = j$, changes in some substantive way as we change j. If the predictors A and B do not interact, then indeed we will be able to examine the relationship between the response and each of the predictors separately. But we almost never know that this is the case beforehand and must assess whether or not an interaction exists based on collected data.

A second reason for including both predictors in the analysis is that this will often lead to a reduction in the contribution of random error to the results. By this, we mean that we will be able to explain some of the observed variation in Y by the inclusion of the second variable in the model. This depends, however, on the additional variable having a relationship with the response. Furthermore, for the inclusion of a second variable to be worthwhile, this relationship must be strong enough to justify the loss in degrees of freedom available for the estimation of the contribution of random error to the experimental results. As we will see, including the second variable in the analysis results in a reduction in the degrees of freedom in the Error row of the ANOVA table. Degrees of freedom are playing the role of sample size here. The fewer the degrees of freedom in the Error row, the less accurate our estimate of σ^2 will be.

When we include both predictors in our analysis, and we have the opportunity to determine the sampling process, it is important that we *cross* the predictors. By this, we mean that we observe Y at each combination

$$(A, B) = (i, j) \in \{1, \ldots, a\} \times \{1, \ldots, b\}.$$

Suppose, then, that we have n_{ij} response values at the $(A, B) = (i, j)$ setting of the predictors. Then, letting

$$E(Y \mid (A, B) = (i, j)) = \beta_{ij}$$

be the mean response when $A = i$ and $B = j$, and introducing the dummy variables

$$X_{ij} = \begin{cases} 1 & A = i, B = j \\ 0 & A \neq i \text{ or } B \neq j, \end{cases}$$

we can write

$$\begin{aligned} E(Y \mid X_{ij} &= x_{ij} \text{ for all } i, j) = \beta_{11}x_{11} + \beta_{21}x_{21} + \cdots + \beta_{ab}x_{ab} \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{ij}x_{ij}. \end{aligned}$$

The relationship between Y and the predictors is completely encompassed in the changes in the β_{ij} as i and j change. From this, we can see that a regression model for this situation is immediately a linear regression model.

Inferences About Individual Means and Differences of Means

Now let y_{ijk} denote the kth response value when $X_{ij} = 1$. Then, as in Section 10.4.1, the least-squares estimate of β_{ij} is given by

$$b_{ij} = \bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk},$$

the mean of the observations when $X_{ij} = 1$. If in addition we assume that the conditional distributions of Y, given the predictors all have variance equal to σ^2, then with $N = n_{11} + n_{21} + \cdots + n_{ab}$, we have that

$$s^2 = \frac{1}{N - ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \bar{y}_{ij}\right)^2 \tag{10.4.5}$$

is an unbiased estimator of σ^2. Therefore, using (10.4.5), the standard error of $\bar{y}_{ij}$ is given by $s/\sqrt{n_{ij}}$.

With the normality assumption, we have that $\bar{Y}_{ij} \sim N(\beta_{ij}, \sigma^2/n_{ij})$, independent of

$$\frac{(N - ab)\, S^2}{\sigma^2} \sim \chi^2(N - ab).$$

This leads to the γ-confidence intervals

$$\bar{y}_{ij} \pm \frac{s}{\sqrt{n_{ij}}} t_{(1+\gamma)/2}(N - ab)$$

for β_{ij} and

$$\bar{y}_{ij} - \bar{y}_{kl} \pm s\sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{kl}}}\, t_{(1+\gamma)/2}(N - ab)$$

for the difference of means $\beta_{ij} - \beta_{kl}$.

The ANOVA for Assessing Interaction and Relationships with the Predictors

We are interested in whether or not there is any relationship between Y and the predictors. There is no relationship between the response and the predictors if and only if all the β_{ij} are equal. Before testing this, however, it is customary to test the null hypothesis that there is no interaction between the predictors. The precise definition of no interaction here is that

$$\beta_{ij} = \mu_i + \upsilon_j$$

for all i and j for some constants μ_i and υ_j, i.e., the means can be expressed additively. Note that if we fix $B = j$ and let A vary, then these *response curves* (a response curve is a plot of the means of one variable while holding the value of the second variable fixed) are all parallel. This is an equivalent way of saying that there is no interaction between the predictors.

In Figure 10.4.5, we have depicted response curves in which the factors do not interact, and in Figure 10.4.6 we have depicted response curves in which they do. Note that the solid lines, for example, joining β_{11} and β_{21}, are there just to make it easier to display the parallelism (or lack thereof) and have no other significance.

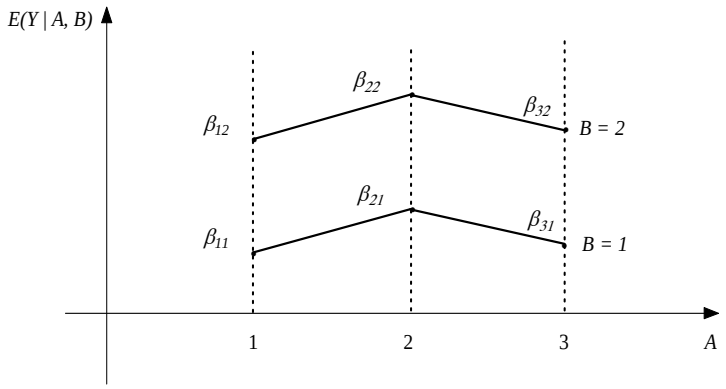

Figure 10.4.5: Response curves for expected response with two predictors, with A taking three levels and B taking two levels. Because they are parallel, the predictors do not interact.

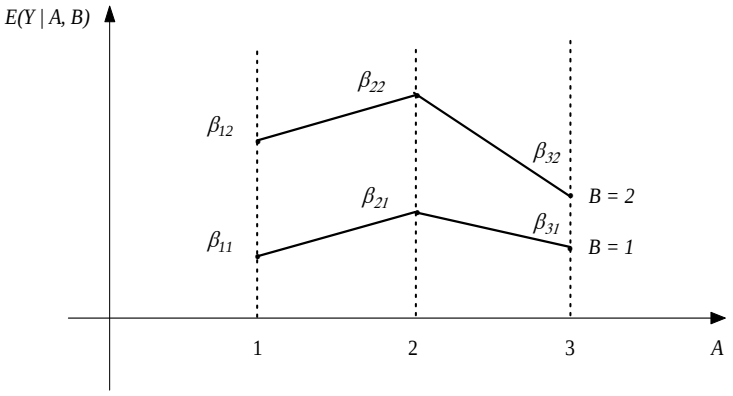

Figure 10.4.6: Response curves for expected response with two predictors, with A taking three levels and B taking two levels. They are not parallel, so the predictors interact.

To test the null hypothesis of no interaction, we must first fit the model where $\beta_{ij} = \mu_i + \upsilon_j$, i.e., find the least-squares estimates of the β_{ij} under these constraints. We will not pursue the mathematics of obtaining these estimates here, but rely on software to do this for us and to compute the sum of squares relevant for testing the null hypothesis of no interaction (from the results of Section 10.3.4, we know that this

is obtained by differencing the regression sum of squares obtained from the full model and the regression sums of squares obtained from the model with no interaction).

If we decide that an interaction exists, then it is immediate that both A and B have an effect on Y (if A does not have an effect, then A and B cannot interact — see Problem 10.4.16); we must look at differences among the $\bar{y}_{ij}$ to determine the form of the relationship. If we decide that no interaction exists, then A has an effect if and only if the μ_i vary, and B has an effect if and only if the υ_j vary. We can test the null hypothesis $H_0 : \mu_1 = \cdots = \mu_a$ of no effect due to A and the null hypothesis $H_0 : \upsilon_1 = \cdots = \upsilon_b$ of no effect due to V separately, once we have decided that no interaction exists.

The details for deriving the relevant sums of squares for all these hypotheses are not covered here, but many statistical packages will produce an ANOVA table, as given below.

Source	Df	Sum of Squares
A	$a-1$	$\text{RSS}(A)$
B	$b-1$	$\text{RSS}(B)$
$A \times B$	$(a-1)(b-1)$	$\text{RSS}(A \times B)$
Error	$N-ab$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(y_{ijk}-\bar{y}_{ij}\right)^2$
Total	$N-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(y_{ijk}-\bar{y}\right)^2$

Note that if we had included only A in the model, then there would be $N-a$ degrees of freedom for the estimation of σ^2. By including B, we lose $(N-a)-(N-ab) = a(b-1)$ degrees of freedom for the estimation of σ^2.

Using this table, we first assess the null hypothesis H_0 : no interaction between A and B, using $F \sim F((a-1)(b-1), N-ab)$ under H_0, via the P-value

$$P\left(F > \frac{\text{RSS}(A \times B)/(a-1)(b-1)}{s^2}\right),$$

where s^2 is given by (10.4.5). If we decide that no interaction exists, then we assess the null hypothesis H_0 : no effect due to A, using $F \sim F(a-1, N-ab)$ under H_0, via the P-value

$$P\left(F > \frac{\text{RSS}(A)/(a-1)}{s^2}\right),$$

and assess H_0 : no effect due to B, using $F \sim F(b-1, N-ab)$ under H_0, via the P-value

$$P\left(F > \frac{\text{RSS}(B)/(b-1)}{s^2}\right).$$

Model Checking

To check the model, we look at the standardized residuals given by (see Problem 10.4.18)

$$\frac{y_{ijk}-\bar{y}_{ij}}{s\sqrt{1-1/n_{ij}}}. \tag{10.4.6}$$

We will restrict our attention to various plots of the standardized residuals for model checking.

We consider an example of a two-factor analysis.

EXAMPLE 10.4.5

The data in the following table come from G. E. P. Box and D. R. Cox, "An analysis of transformations" (*Journal of the Royal Statistical Society*, 1964, Series B, p. 211) and represent survival times, in hours, of animals exposed to one of three different types of poisons and allocated four different types of treatments. We let A denote the treatments and B denote the type of poison, so we have $3 \times 4 = 12$ different (A, B) combinations. Each combination was administered to four different animals; i.e., $n_{ij} = 4$ for every i and j.

	$A1$	$A2$	$A3$	$A4$
$B1$	3.1, 4.5, 4.6, 4.3	8.2, 11.0, 8.8, 7.2	4.3, 4.5, 6.3, 7.5	4.5, 7.1, 6.6, 6.2
$B2$	3.6, 2.9, 4.0, 2.3	9.2, 6.1, 4.9, 12.4	4.4, 3.5, 3.1, 4.0	5.6, 10.2, 7.1, 3.8
$B3$	2.2, 2.1, 1.8, 2.3	3.0, 3.7, 3.8, 2.9	2.3, 2.5, 2.4, 2.2	3.0, 3.6, 3.1, 3.3

A normal probability plot for these data, using the standardized residuals after fitting the two-factor model, reveals a definite problem. In the above reference, a transformation of the response to the reciprocal $1/Y$ is suggested, based on a more sophisticated analysis, and this indeed leads to much more appropriate standardized residual plots. Figure 10.4.7 is a normal probability plot for the standardized residuals based on the reciprocal response. This normal probability plot looks reasonable.

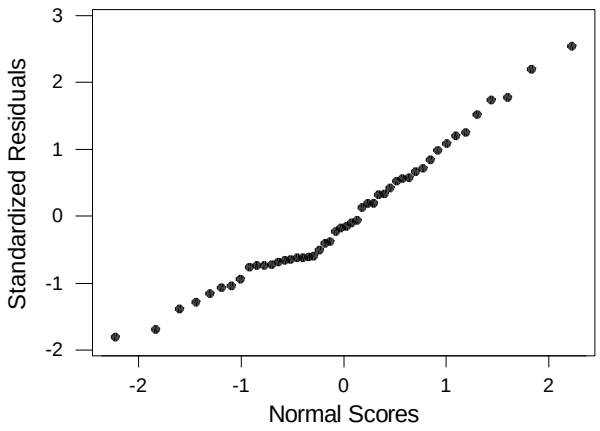

Figure 10.4.7: Normal probability plot of the standardized residuals in Example 10.4.5 using the reciprocal of the response.

Figure 10.4.8 is a plot of the standardized residuals against the various (A, B) combinations, where we have coded the combination (i, j) as $b(i-1)+j$ with $b = 3, i = 1, 2, 3, 4$, and $j = 1, 2, 3$. This coding assigns a unique integer to each combination (i, j) and is convenient when comparing scatter plots of the response for

each treatment. Again, this residual plot looks reasonable.

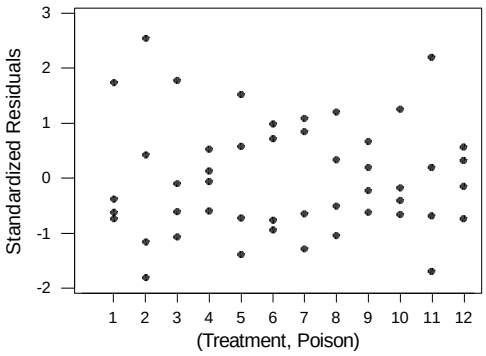

Figure 10.4.8: Scatter plot for the data in Example 10.4.5 of the standardized residuals against each value of (A, B) using the reciprocal of the response.

Below we provide the least-squares estimates of the β_{ij} for the transformed model.

	$A1$	$A2$	$A3$	$A4$
$B1$	0.24869	0.11635	0.18627	0.16897
$B2$	0.32685	0.13934	0.27139	0.17015
$B3$	0.48027	0.30290	0.42650	0.30918

The ANOVA table for the data, as obtained from a standard statistical package, is given below.

Source	Df	Sum of Squares	Mean Square
A	3	0.20414	0.06805
B	2	0.34877	0.17439
$A \times B$	6	0.01571	0.00262
Error	36	0.08643	0.00240
Total	47	0.65505	

From this, we determine that $s = \sqrt{0.00240} = 4.89898 \times 10^{-2}$, and so the standard errors of the least-squares estimates are all equal to $s/2 = 0.0244949$.

To test the null hypothesis of no interaction between A and B, we have, using $F \sim F(6, 36)$ under H_0, the P-value

$$P\left(F > \frac{0.00262}{0.00240}\right) = P\,(F > 1.09) = 0.387.$$

We have no evidence against the null hypothesis.

So we can go on to test the null hypothesis of no effect due to A and we have, using $F \sim F\,(2, 36)$ under H_0, the P-value

$$P\left(F > \frac{0.06805}{0.00240}\right) = P(F > 28.35) = 0.000.$$

We reject this null hypothesis.

Similarly, testing the null hypothesis of no effect due to B, we have, using $F \sim F(2, 36)$ under H_0, the P-value

$$P\left(F > \frac{0.17439}{0.00240}\right) = P(F > 72.66) = 0.000.$$

We reject this null hypothesis as well.

Accordingly, we have decided that the appropriate model is the *additive model* given by $E(1/Y \mid (A, B)) = (i, j) = \mu_i + \upsilon_j$ (we are still using the transformed response $1/Y$). We can also write this as $E(1/Y \mid (A, B)) = (i, j) = \left(\mu_i + \alpha\right) + \left(\upsilon_j - \alpha\right)$ for any choice of α. Therefore, there is no unique estimate of the additive effects due to A or B. However, we still have unique least-squares estimates of the means, which are obtained (using software) by fitting the model with constraints on the β_{ij} corresponding to no interaction existing. These are recorded in the following table.

	$A1$	$A2$	$A3$	$A4$
$B1$	0.26977	0.10403	0.21255	0.13393
$B2$	0.31663	0.15089	0.25942	0.18080
$B3$	0.46941	0.30367	0.41219	0.33357

As we have decided that there is no interaction between A and B, we can assess single-factor effects by examining the response means for each factor separately. For example, the means for investigating the effect of A are given in the following table.

$A1$	$A2$	$A3$	$A4$
0.352	0.186	0.295	0.216

We can compare these means using procedures based on the t-distribution. For example, a 0.95-confidence interval for the difference in the means at levels $A1$ and $A2$ is given by

$$\begin{aligned} \bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm \frac{s}{\sqrt{12}}\, t_{0.975}\,(36) &= (0.352 - 0.186) \pm \sqrt{\frac{0.00240}{12}}\, 2.0281 \\ &= (0.13732, 0.19468). \end{aligned} \tag{10.4.7}$$

This indicates that we would reject the null hypothesis of no difference between these means at the 0.05 level.

Notice that we have used the estimate of σ^2 based on the full model in (10.4.7). Logically, it would seem to make more sense to use the estimate based on fitting the additive model because we have decided that it is appropriate. When we do so, this is referred to as *pooling*, as it can be shown that the new error estimate is calculated by adding RSS($A \times B$) to the original ESS and dividing by the sum of the $A \times B$ degrees of freedom and the error degrees of freedom. Not to pool is regarded as a somewhat more conservative procedure. ∎

10.4.4 Randomized Blocks

With two-factor models, we generally want to investigate whether or not both of these factors have a relationship with the response Y. Suppose, however, that we know that a factor B has a relationship with Y, and we are interested in investigating whether or not another factor A has a relationship with Y. Should we run a single-factor experiment using the predictor A, or run a two-factor experiment including the factor B?

The answer is as we have stated at the start of Section 10.4.2. Including the factor B will allow us, if B accounts for a lot of the observed variation, to make more accurate comparisons. Notice, however, that if B does not have a substantial effect on Y, then its inclusion will be a waste, as we sacrificed $a(b-1)$ degrees of freedom that would otherwise go toward the estimation of σ^2.

So it is important that we do indeed *know* that B has a substantial effect. In such a case, we refer to B as a *blocking variable*. It is important again that the blocking variable B be crossed with A. Then we can test for any effect due to A by first testing for an interaction between A and B; if no such interaction is found, then we test for an effect due to A alone, just as we have discussed in Section 10.4.3.

A special case of using a blocking variable arises when we have $n_{ij} = 1$ for all i and j. In this case, $N = ab$, so there are no degrees of freedom available for the estimation of error. In fact, we have that (see Problem 10.4.19) $s^2 = 0$. Still, such a design has practical value, provided we are willing to *assume* that there is no interaction between A and B. This is called a *randomized block design.*

For a randomized block design, we have that

$$s^2 = \frac{\text{RSS}(A \times B)}{(a-1)\,(b-1)} \tag{10.4.8}$$

is an unbiased estimate of σ^2, and so we have $(a-1)\,(b-1)$ degrees of freedom for the estimation of error. Of course, this will not be correct if A and B do interact, but when they do not, this can be a highly efficient design, as we have removed the effect of the variation due to B and require only ab observations for this. When the randomized block design is appropriate, we test for an effect due to A, using $F \sim F(a-1, (a-1)\,(b-1))$ under H_0, via the P-value

$$P\left(F > \frac{\text{RSS}(A)/(a-1)}{s^2}\right).$$

10.4.5 One Categorical and One Quantitative Predictor

It is also possible that the response is quantitative while some of the predictors are categorical and some are quantitative. We now consider the situation where we have one categorical predictor A, taking a values, and one quantitative predictor W. We assume that the regression model applies. Furthermore, we restrict our attention to the situation where we suppose that, within each level of A, the mean response varies as

$$E(Y \mid (A, W)) = (i, w) = \beta_{i1} + \beta_{i2} w,$$

so that we have a simple linear regression model within each level of A.

If we introduce the dummy variables

$$X_{ij} = \begin{cases} W^{j-1} & A = i \\ 0 & A \neq i \end{cases}$$

for $i = 1, \ldots, a$ and $j = 1, 2$, then we can write the linear regression model as

$$E(Y \mid (X_{ij}) = (x_{ij})) = (\beta_{11}x_{11} + \beta_{12}x_{12}) + \cdots + (\beta_{a1}x_{a1} + \beta_{a2}x_{a2}).$$

Here, β_{i1} is the intercept and β_{i2} is the slope specifying the relationship between Y and W when $A = i$. The methods of Section 10.3.4 are then available for inference about this model.

We also have a notion of interaction in this context, as we say that the two predictors interact if the slopes of the lines vary across the levels of A. So saying that no interaction exists is the same as saying that the response curves are parallel when graphed for each level of A. If an interaction exists, then it is definite that both A and W have an effect on Y. Thus the null hypothesis that no interaction exists is equivalent to $H_0 : \beta_{12} = \cdots = \beta_{a2}$.

If we decide that no interaction exists, then we can test for no effect due to W by testing the null hypothesis that the common slope is equal to 0, or we can test the null hypothesis that there is no effect due to A by testing $H_0 : \beta_{11} = \cdots = \beta_{a1}$, i.e., that the intercept terms are the same across the levels of A.

We do not pursue the analysis of this model further here. Statistical software is available, however, that will calculate the relevant ANOVA table for assessing the various null hypotheses.

Analysis of Covariance

Suppose we are running an experimental design and for each experimental unit we can measure, but not control, a quantitative variable W that we believe has an effect on the response Y. If the effect of this variable is appreciable, then good statistical practice suggests we should include this variable in the model, as we will reduce the contribution of error to our experimental results and thus make more accurate comparisons. Of course, we pay a price when we do this, as we lose degrees of freedom that would otherwise be available for the estimation of error. So we must be sure that W does have a significant effect in such a case. Also, we do not test for an effect of such a variable, as we presumably know it has an effect. This technique is referred to as the *analysis of covariance* and is obviously similar in nature to the use of blocking variables.

Summary of Section 10.4

- We considered the situation involving a quantitative response and categorical predictor variables.
- By the introduction of dummy variables for the predictor variables, we can consider this situation as a particular application of the multiple regression model of Section 10.3.4.

- If we decide that a relationship exists, then we typically try to explain what form this relationship takes by comparing means. To prevent finding too many statistically significant differences, we lower the individual error rate to ensure a sensible family error rate.
- When we have two predictors, we first check to see if the factors interact. If the two predictors interact, then both have an effect on the response.
- A special case of a two-way analysis arises when one of the predictors serves as a blocking variable. It is generally important to know that the blocking variable has an effect on the response, so that we do not waste degrees of freedom by including it.
- Sometimes we can measure variables on individual experimental units that we know have an effect on the response. In such a case, we include these variables in our model, as they will reduce the contribution of random error to the analysis and make our inferences more accurate.

EXERCISES

10.4.1 The following values of a response Y were obtained for three settings of a categorical predictor A.

$A = 1$	2.9976	0.3606	4.7716	1.5652
$A = 2$	0.7468	1.3308	2.2167	−0.3184
$A = 3$	2.1192	2.3739	0.3335	3.3015

Suppose we assume the normal regression model for these data with one categorical predictor.

(a) Produce a side-by-side boxplot for the data.

(b) Plot the standardized residuals against A (if you are using a computer for your calculations, also produce a normal probability plot of the standardized residuals). Does this give you grounds for concern that the model assumptions are incorrect?

(c) Carry out a one-way ANOVA to test for any difference among the conditional means of Y given A.

(d) If warranted, construct 0.95-confidence intervals for the differences between the means and summarize your findings.

10.4.2 The following values of a response Y were obtained for three settings of a categorical predictor A.

$A = 1$	0.090	0.800	33.070	−1.890
$A = 2$	5.120	1.580	1.760	1.740
$A = 3$	5.080	−3.510	4.420	1.190

Suppose we assume the normal regression model for these data with one categorical predictor.

(a) Produce a side-by-side boxplot for the data.

(b) Plot the standardized residuals against A (if you are using a computer for your calculations, also produce a normal probability plot of the standardized residuals). Does this give you grounds for concern that the model assumptions are incorrect?

(c) If concerns arise about the validity of the model, can you "fix" the problem?

(d) If you have been able to fix any problems encountered with the model, carry out a one-way ANOVA to test for any differences among the conditional means of Y given A.

(e) If warranted, construct 0.95-confidence intervals for the differences between the means and summarize your findings.

10.4.3 The following table gives the percentage moisture content of two different types of cheeses determined by randomly sampling batches of cheese from the production process.

Cheese 1	39.02, 38.79, 35.74, 35.41, 37.02, 36.00
Cheese 2	38.96, 39.01, 35.58, 35.52, 35.70, 36.04

Suppose we assume the normal regression model for these data with one categorical predictor.

(a) Produce a side-by-side boxplot for the data.

(b) Plot the standardized residuals against Cheese (if you are using a computer for your calculations, also produce a normal probability plot of the standardized residuals). Does this give you grounds for concern that the model assumptions are incorrect?

(c) Carry out a one-way ANOVA to test for any differences among the conditional means of Y given Cheese. Note that this is the same as a t-test for the difference in the means.

10.4.4 In an experiment, rats were fed a stock ration for 100 days with various amounts of gossypol added. The following weight gains in grams were recorded.

0.00% Gossypol	228, 229, 218, 216, 224, 208, 235, 229, 233, 219, 224, 220, 232, 200, 208, 232
0.04% Gossypol	186, 229, 220, 208, 228, 198, 222, 273, 216, 198, 213
0.07% Gossypol	179, 193, 183, 180, 143, 204, 114, 188, 178, 134, 208, 196
0.10% Gossypol	130, 87, 135, 116, 118, 165, 151, 59, 126, 64, 78, 94, 150, 160, 122, 110, 178
0.13% Gossypol	154, 130, 118, 118, 118, 104, 112, 134, 98, 100, 104

Suppose we assume the normal regression model for these data and treat gossypol as a categorical predictor taking five levels.

(a) Create a side-by-side boxplot graph for the data. Does this give you any reason to be concerned about the assumptions that underlie an analysis based on the normal regression model?

(b) Produce a plot of the standardized residuals against the factor gossypol (if you are using a computer for your calculations, also produce a normal probability plot of the standardized residuals). What are your conclusions?

(c) Carry out a one-way ANOVA to test for any differences among the mean responses for the different amounts of gossypol.

(d) Compute 0.95-confidence intervals for all the pairwise differences of means and summarize your conclusions.

10.4.5 In an investigation into the effect of deficiencies of trace elements on a variable Y measured on sheep, the data in the following table were obtained.

Control	13.2, 13.6, 11.9, 13.0, 14.5, 13.4
Cobalt	11.9, 12.2, 13.9, 12.8, 12.7, 12.9
Copper	14.2, 14.0, 15.1, 14.9, 13.7, 15.8
Cobalt + Copper	15.0, 15.6, 14.5, 15.8, 13.9, 14.4

Suppose we assume the normal regression model for these data with one categorical predictor.

(a) Produce a side-by-side boxplot for the data.

(b) Plot the standardized residuals against the predictor (if you are using a computer for your calculations, also produce a normal probability plot of the standardized residuals). Does this give you grounds for concern that the model assumptions are incorrect?

(c) Carry out a one-way ANOVA to test for any differences among the conditional means of Y given the predictor.

(d) If warranted, construct 0.95-confidence intervals for all the pairwise differences between the means and summarize your findings.

10.4.6 Two diets were given to samples of pigs over a period of time, and the following weight gains (in lbs) were recorded.

Diet A	8, 4, 14, 15, 11, 10, 6, 12, 13, 7
Diet B	7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17

Suppose we assume the normal regression model for these data.

(a) Produce a side-by-side boxplot for the data.

(b) Plot the standardized residuals against Diet. Also produce a normal probability plot of the standardized residuals. Does this give you grounds for concern that the model assumptions are incorrect?

(c) Carry out a one-way ANOVA to test for a difference between the conditional means of Y given Diet.

(d) Construct 0.95-confidence intervals for differences between the means.

10.4.7 Ten students were randomly selected from the students in a university who took first-year calculus and first-year statistics. Their grades in these courses are recorded in the following table.

Student	1	2	3	4	5	6	7	8	9	10
Calculus	66	61	77	62	66	68	64	75	59	71
Statistics	66	63	79	63	67	70	71	80	63	74

Suppose we assume the normal regression model for these data.

(a) Produce a side-by-side boxplot for the data.

(b) Treating the calculus and statistics marks as separate samples, carry out a one-way ANOVA to test for any difference between the mean mark in calculus and the mean mark in statistics. Produce the appropriate plots to check for model assumptions.

(c) Now take into account that each student has a calculus mark and a statistics mark and test for any difference between the mean mark in calculus and the mean mark in statistics. Produce the appropriate plots to check for model assumptions. Compare your results with those obtained in part (b).

(d) Estimate the correlation between the calculus and statistics marks.

10.4.8 The following data were recorded in *Statistical Methods*, 6th ed., by G. Snedecor and W. Cochran (Iowa State University Press, Ames, 1967) and represent the average number of florets observed on plants in seven plots. Each of the plants was planted with either high corms or low corms (a type of underground stem).

	Plot 1	Plot 2	Plot 3	Plot 4	Plot 5	Plot 6	Plot 7
Corm High	11.2	13.3	12.8	13.7	12.2	11.9	12.1
Corm Low	14.6	12.6	15.0	15.6	12.7	12.0	13.1

Suppose we assume the normal regression model for these data.

(a) Produce a side-by-side boxplot for the data.

(b) Treating the Corm High and Corm Low measurements as separate samples, carry out a one-way ANOVA to test for any difference between the population means. Produce the appropriate plots to check for model assumptions.

(c) Now take into account that each plot has a Corm High and Corm Low measurement. Compare your results with those obtained in part (b). Produce the appropriate plots to check for model assumptions.

(d) Estimate the correlation between the calculus and statistics marks.

10.4.9 Suppose two measurements, Y_1 and Y_2, corresponding to different treatments, are taken on the same individual who has been randomly sampled from a population Π. Suppose that Y_1 and Y_2 have the same variance and are negatively correlated. Our goal is to compare the treatment means. Explain why it would have been better to have randomly sampled two individuals from Π and applied the treatments to these individuals separately. (Hint: Consider $\text{Var}(Y_1 - Y_2)$ in these two sampling situations.)

10.4.10 List the assumptions that underlie the validity of the one-way ANOVA test discussed in Section 10.4.1.

10.4.11 List the assumptions that underlie the validity of the paired comparison test discussed in Section 10.4.2.

10.4.12 List the assumptions that underlie the validity of the two-way ANOVA test discussed in Section 10.4.3.

10.4.13 List the assumptions that underlie the validity of the test used with the randomized block design, discussed in Section 10.4.4, when $n_{ij} = 1$ for all i and j.

PROBLEMS

10.4.14 Prove that $\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\beta_i\right)^2$ is minimized as a function of the β_i by $\beta_i=\bar{y}_i=\left(y_{i1}+\cdots+y_{in_i}\right)/n_i$ for $i=1,\ldots,a$.

10.4.15 Prove that

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2=\sum_{i=1}^{a}n_i\left(\bar{y}_i-\bar{y}\right)^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$$

where $\bar{y}_i=\left(y_{i1}+\cdots+y_{in_i}\right)/n_i$ and $\bar{y}$ is the grand mean.

10.4.16 Argue that if the relationship between a quantitative response Y and two categorical predictors A and B is given by a linear regression model, then A and B both have an effect on Y whenever A and B interact. (Hint: What does it mean in terms of response curves for an interaction to exist, for an effect due to A to exist?)

10.4.17 Establish that (10.4.2) is the appropriate expression for the standardized residual for the linear regression model with one categorical predictor.

10.4.18 Establish that (10.4.6) is the appropriate expression for the standardized residual for the linear regression model with two categorical predictors.

10.4.19 Establish that $s^2=0$ for the linear regression model with two categorical predictors when $n_{ij}=1$ for all i and j.

10.4.20 How would you assess whether or not the randomized block design was appropriate after collecting the data?

COMPUTER PROBLEMS

10.4.21 Use appropriate software to carry out Fisher's multiple comparison test on the data in Exercise 10.4.5 so that the family error rate is between 0.04 and 0.05. What individual error rate is required?

10.4.22 Consider the data in Exercise 10.4.3, but now suppose we also take into account that the cheeses were made in lots where each lot corresponded to a production run. Recording the data this way, we obtain the following table.

	Lot 1	Lot 2	Lot 3
Cheese 1	39.02, 38.79	35.74, 35.41	37.02, 36.00
Cheese 2	38.96, 39.01	35.58, 35.52	35.70, 36.04

Suppose we assume the normal regression model for these data with two categorical predictors.

(a) Produce a side-by-side boxplot for the data for each treatment.

(b) Produce a table of cell means.

(c) Produce a normal probability plot of the standardized residuals and a plot of the standardized residuals against each treatment combination (code the treatment combinations so there is a unique integer corresponding to each). Comment on the validity of the model.

(d) Construct the ANOVA table testing first for no interaction between A and B and, if necessary, an effect due to A and an effect due to B.

(e) Based on the results of part (d), construct the appropriate table of means, plot the corresponding response curve, and make all pairwise comparisons among the means.

(f) Compare your results with those obtained in Exercise 10.4.4 and comment on the differences.

10.4.23 A two-factor experimental design was carried out, with factors A and B both categorical variables taking three values. Each treatment was applied four times and the following response values were obtained.

	$A=1$		$A=2$		$A=3$	
$B=1$	19.86	20.88	26.37	24.38	29.72	29.64
	20.15	25.44	24.87	30.93	30.06	35.49
$B=2$	15.35	15.86	22.82	20.98	27.12	24.27
	21.86	26.92	29.38	34.13	34.78	40.72
$B=3$	4.01	4.48	10.34	9.38	15.64	14.03
	21.66	25.93	30.59	40.04	36.80	42.55

Suppose we assume the normal regression model for these data with two categorical predictors.

(a) Produce a side-by-side boxplot for the data for each treatment.

(b) Produce a table of cell means.

(c) Produce a normal probability plot of the standardized residuals and a plot of the standardized residuals against each treatment combination (code the treatment combinations so there is a unique integer corresponding to each). Comment on the validity of the model.

(d) Construct the ANOVA table testing first for no interaction between A and B and, if necessary, an effect due to A and an effect due to B.

(e) Based on the results of part (d), construct the appropriate table of means, plot the corresponding response curves, and make all pairwise comparisons among the means.

10.4.24 A chemical paste is made in batches and put into casks. Ten delivery batches were randomly selected for testing; then three casks were randomly selected from each delivery and the paste strength was measured twice, based on samples drawn from each sampled cask. The response was expressed as a percentage of fill strength. The collected data are given in the following table. Suppose we assume the normal regression model for these data with two categorical predictors.

	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Cask 1	62.8, 62.6	60.0, 61.4	58.7, 57.5	57.1, 56.4	55.1, 55.1
Cask 2	60.1, 62.3	57.5, 56.9	63.9, 63.1	56.9, 58.6	54.7, 54.2
Cask 3	62.7, 63.1	61.1, 58.9	65.4, 63.7	64.7, 64.5	58.5, 57.5
	Batch 6	Batch 7	Batch 8	Batch 9	Batch 10
Cask 1	63.4, 64.9	62.5, 62.6	59.2, 59.4	54.8, 54.8	58.3, 59.3
Cask 2	59.3, 58.1	61.0, 58.7	65.2, 66.0	64.0, 64.0	59.2, 59.2
Cask 3	60.5, 60.0	56.9, 57.7	64.8, 64.1	57.7, 56.8	58.9, 56.8

(a) Produce a side-by-side boxplot for the data for each treatment.

(b) Produce a table of cell means.

(c) Produce a normal probability plot of the standardized residuals and a plot of the standardized residuals against each treatment combination (code the treatment combinations so there is a unique integer corresponding to each). Comment on the validity of the model.

(d) Construct the ANOVA table testing first for no interaction between Batch and Cask and, if necessary, no effect due to Batch and no effect due to Cask.

(e) Based on the results of part (d), construct the appropriate table of means and plot the corresponding response curves.

10.4.25 The following data arose from a randomized block design, where factor B is the blocking variable and corresponds to plots of land on which cotton is planted. Each plot was divided into five subplots, and different concentrations of fertilizer were applied to each, with the response being a strength measurement of the cotton harvested. There were three blocks and five different concentrations of fertilizer. Note that there is only one observation for each block and concentration combination. Further discussion of these data can be found in *Experimental Design*, 2nd ed., by W. G. Cochran and G. M. Cox (John Wiley & Sons, New York, 1957, pp. 107–108). Suppose we assume the normal regression model with two categorical predictors.

	$B = 1$	$B = 2$	$B = 3$
$A = 36$	7.62	8.00	7.93
$A = 54$	8.14	8.15	7.87
$A = 72$	7.70	7.73	7.74
$A = 108$	7.17	7.57	7.80
$A = 144$	7.46	7.68	7.21

(a) Construct the ANOVA table for testing for no effect due to fertilizer and which also removes the variation due to the blocking variable.

(b) Beyond the usual assumptions that we are concerned about, what additional assumption is necessary for this analysis?

(c) Actually, the factor A is a quantitative variable. If we were to take this into account by fitting a model that had the same slope for each block but possibly different intercepts, then what benefit would be gained?

(d) Carry out the analysis suggested in part (c) and assess whether or not this model makes sense for these data.

10.5 Categorical Response and Quantitative Predictors

We now consider the situation in which the response is categorical but at least some of the predictors are quantitative. The essential difficulty in this context lies with the quantitative predictors, so we will focus on the situation in which all the predictors

are quantitative. When there are also some categorical predictors, these can be handled in the same way, as we can replace each categorical predictor by a set of dummy quantitative variables, as discussed in Section 10.4.5.

For reasons of simplicity, we will restrict our attention to the situation in which the response variable Y is binary valued, and we will take these values to be 0 and 1. Suppose, then, that there are k quantitative predictors $X_1, \ldots, X_k$. Because $Y \in \{0, 1\}$, we have

$$E(Y \mid X_1 = x_1, \ldots, X_k = x_k) = P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k) \in [0, 1].$$

Therefore, we cannot write $E(Y \mid x_1, \ldots, x_k) = \beta_1 x_1 + \cdots + \beta_k x_k$ without placing some unnatural restrictions on the β_i to ensure that $\beta_1 x_1 + \cdots + \beta_k x_k \in [0, 1]$.

Perhaps the simplest way around this is to use a 1–1 function $l : [0, 1] \to R^1$ and write

$$l(P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k)) = \beta_1 x_1 + \cdots + \beta_k x_k,$$

so that

$$P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k) = l^{-1}(\beta_1 x_1 + \cdots + \beta_k x_k).$$

We refer to l as a *link function*. There are many possible choices for l. For example, it is immediate that we can take l to be any inverse cdf for a continuous distribution.

If we take $l = \Phi^{-1}$, i.e., the inverse cdf of the $N(0, 1)$ distribution, then this is called the *probit link*. A more commonly used link, due to some inherent mathematical simplicities, is the *logistic link* given by

$$l(p) = \ln\left(\frac{p}{1 - p}\right). \tag{10.5.1}$$

The right-hand side of (10.5.1) is referred to as the *logit* or *log odds*. The logistic link is the inverse cdf of the logistic distribution (see Exercise 10.5.1). We will restrict our discussion to the logistic link hereafter.

The logistic link implies that (see Exercise 10.5.2)

$$P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k) = \frac{\exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}}, \tag{10.5.2}$$

which is a relatively simple relationship. We see immediately, however, that

$$\begin{aligned}&\text{Var}(Y \mid X_1 = x_1, \ldots, X_k = x_k)\\ &= P(Y = 1 \mid X_1 = x_1, \ldots, X_k) = x_k(1 - P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k)),\end{aligned}$$

so the variance of the conditional distribution of Y, given the predictors, depends on the values of the predictors. Therefore, these models are not, strictly speaking, regression models as we have defined them. Still when we use the link function given by (10.5.1), we refer to this as the *logistic regression model*.

Now suppose we observe n independent observations $(x_{i1}, \ldots, x_{ik}, y_i)$ for $i = 1, \ldots, n$. We then have that, given $(x_{i1}, \ldots, x_{ik})$, the response y_i is an observation

from the Bernoulli($P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k)$) distribution. Then (10.5.2) implies that the conditional likelihood, given the values of the predictors, is

$$\prod_{i=1}^{n} \left(\frac{\exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}} \right)^{y_i} \left(\frac{1}{1 + \exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}} \right)^{1-y_i}.$$

Inference about the β_i then proceeds via the likelihood methods discussed in Chapter 6. In fact, we need to use software to obtain the MLE's, and, because the exact sampling distributions of these quantities are not available, the large sample methods discussed in Section 6.5 are used for approximate confidence intervals and P-values. Note that assessing the null hypothesis $H_0 : \beta_i = 0$ is equivalent to assessing the null hypothesis that the predictor X_i does not have a relationship with the response.

We illustrate the use of logistic regression via an example.

EXAMPLE 10.5.1
The following table of data represent the

(number of failures, number of successes)

for ingots prepared for rolling under different settings of the predictor variables, $U =$ soaking time and $V =$ heating time, as reported in *Analysis of Binary Data*, by D. R. Cox (Methuen, London, 1970). A failure indicates that an ingot is not ready for rolling after the treatment. There were observations at 19 different settings of these variables.

	$V = 7$	$V = 14$	$V = 27$	$V = 51$
$U = 1.0$	(0, 10)	(0, 31)	(1, 55)	(3, 10)
$U = 1.7$	(0, 17)	(0, 43)	(4, 40)	(0, 1)
$U = 2.2$	(0, 7)	(2, 31)	(0, 21)	(0, 1)
$U = 2.8$	(0, 12)	(0, 31)	(1, 21)	(0, 0)
$U = 4.0$	(0, 9)	(0, 19)	(1, 15)	(0, 1)

Including an intercept in the model and linear terms for U and V leads to three predictor variables $X_1 \equiv 1$, $X_2 = U$, $X_3 = V$, and the model takes the form

$$P(Y = 1 \mid X_2 = x_2, X_3 = x_3) = \frac{\exp\{\beta_1 + \beta_2 x_2 + \beta_3 x_3\}}{1 + \exp\{\beta_1 + \beta_2 x_2 + \beta_3 x_3\}}.$$

Fitting the model via the method of maximum likelihood leads to the estimates given in the following table. Here, z is the value of estimate divided by its standard error. Because this is approximately distributed $N(0, 1)$ when the corresponding β_i equals 0, the P-value for assessing the null hypothesis that $\beta_i = 0$ is $P(|Z| > |z|)$ with $Z \sim N(0, 1)$.

Coefficient	Estimate	Std. Error	z	P-value
β_1	5.55900	1.12000	4.96	0.000
β_2	−0.05680	0.33120	−0.17	0.864
β_3	−0.08203	0.02373	−3.46	0.001

Of course, we have to feel confident that the model is appropriate before we can proceed to make formal inferences about the β_i. In this case, we note that the number

of successes $s(x_2, x_3)$ in the cell of the table, corresponding to the setting $(X_2, X_3) = (x_2, x_3)$, is an observation from a

$$\text{Binomial}(m(x_2, x_3), P(Y = 1 \mid X_2 = x_2, X_3 = x_3))$$

distribution, where $m(x_2, x_3)$ is the sum of the number of successes and failures in that cell. So, for example, if $X_2 = U = 1.0$ and $X_3 = V = 7$, then $m(1.0, 7) = 10$ and $s\,(1.0, 7) = 10$. Denoting the estimate of $P(Y = 1 \mid X_2 = x_2, X_3 = x_3)$ by $\hat{p}(x_2, x_3)$, obtained by plugging in the MLE, we have that (see Problem 10.5.8)

$$X^2 = \sum_{(x_2, x_3)} \frac{(s(x_2, x_3) - m(x_2, x_3)\hat{p}(x_2, x_3))^2}{m(x_2, x_3)\hat{p}(x_2, x_3)} \tag{10.5.3}$$

is asymptotically distributed as a $\chi^2(19 - 3) = \chi^2(16)$ distribution when the model is correct. We determine the degrees of freedom by counting the number of cells where there were observations (19 in this case, as no observations were obtained when $U = 2.8, V = 51$) and subtracting the number of parameters estimated. For these data, $X^2 = 13.543$ and the P-value is $P(\chi^2(16) > 13.543) = 0.633$. Therefore, we have no evidence that the model is incorrect and can proceed to make inferences about the β_i based on the logistic regression model.

From the preceding table, we see that the null hypothesis $H_0 : \beta_2 = 0$ is not rejected. Accordingly, we drop X_2 and fit the smaller model given by

$$P(Y = 1 \mid X_3 = x_3) = \frac{\exp\{\beta_1 + \beta_3 x_3\}}{1 + \exp\{\beta_1 + \beta_3 x_3\}}.$$

This leads to the estimates $\hat{\beta}_1 = 5.4152$ and $\hat{\beta}_3 = -0.08070$. Note that these are only marginally different from the previous estimates. In Figure 10.5.1, we present a graph of the fitted function over the range where we have observed X_3. ∎

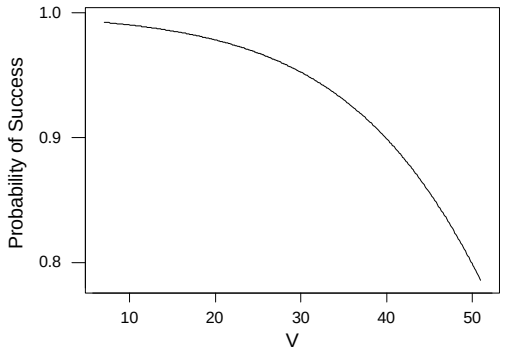

Figure 10.5.1: The fitted probability of obtaining an ingot ready to be rolled as a function of heating time in Example 10.5.1.

Summary of Section 10.5

- We have examined the situation in which we have a single binary-valued response variable and a number of quantitative predictors.
- One method of expressing a relationship between the response and predictors is via the use of a link function.
- If we use the logistic link function, then we can carry out a logistic regression analysis using likelihood methods of inference.

EXERCISES

10.5.1 Prove that the function $f : R^1 \to R^1$, defined by $f(x) = e^{-x}(1 + e^{-x})^{-2}$ for $x \in R^1$, is a density function with distribution function given by $F(x) = (1 + e^{-x})^{-1}$ and inverse cdf given by $F^{-1}(p) = \ln p - \ln(1 - p)$ for $p \in [0, 1]$. This is called the *logistic distribution*.

10.5.2 Establish (10.5.2).

10.5.3 Suppose that a logistic regression model for a binary-valued response Y is given by

$$P(Y = 1 \mid x) = \frac{\exp\{\beta_1 + \beta_2 x\}}{1 + \exp\{\beta_1 + \beta_2 x\}}.$$

Prove that the log odds at $X = x$ is given by $\beta_1 + \beta_2 x$.

10.5.4 Suppose that instead of the inverse logistic cdf as the link function, we use the inverse cdf of a Laplace distribution (see Problem 2.4.22). Determine the form of $P(Y = 1|X_1 = x_1, \ldots, X_k = x_k)$.

10.5.5 Suppose that instead of the inverse logistic cdf as the link function, we use the inverse cdf of a Cauchy distribution (see Problem 2.4.21). Determine the form of $P(Y = 1|X_1 = x_1, \ldots, X_k = x_k)$.

COMPUTER EXERCISES

10.5.6 Use software to replicate the results of Example 10.5.1.

10.5.7 Suppose that the following data were obtained for the quantitative predictor X and the binary-valued response variable Y.

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	0, 0	0, 0	0, 0	0, 0	1, 0	0, 0	1, 0	0, 1	1, 1	1, 1	1, 1

(a) Using these data, fit the logistic regression model given by

$$P(Y = 1 \mid x) = \frac{\exp\{\beta_1 + \beta_2 x + \beta_3 x^2\}}{1 + \exp\{\beta_1 + \beta_2 x + \beta_3 x^2\}}.$$

(b) Does the model fit the data?

(c) Test the null hypothesis $H_0 : \beta_3 = 0$.

(d) If you decide there is no quadratic effect, refit the model and test for any linear effect.
(e) Plot $P(Y = 1 \mid x)$ as a function of x.

PROBLEMS

10.5.8 Prove that (10.5.3) is the correct form for the chi-squared goodness-of-fit test statistic.

10.6 Further Proofs (Advanced)

Proof of Theorem 10.3.1

We want to prove that, when $E(Y \mid X = x) = \beta_1 + \beta_2 x$ and we observe the independent values $(x_1, y_1), \ldots, (x_n, y_n)$ for (X, Y), then the least-squares estimates of β_1 and β_2 are given by $b_1 = \bar{y} - b_2\bar{x}$ and

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

whenever $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$.

We need an algebraic result that will simplify our calculations.

Lemma 10.6.1 If $(x_1, y_1), \ldots, (x_n, y_n)$ are such that $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$ and $q, r \in R^1$, then $\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)(q + r x_i) = 0$.

PROOF We have

$$\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i) = n\bar{y} - nb_1 - nb_2\bar{x} = n(\bar{y} - \bar{y} + b_2\bar{x} - b_2\bar{x}) = 0,$$

which establishes that $\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)q = 0$ for any q. Now using this, and the formulas in Theorem 10.3.1, we obtain

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)x_i \\
&= \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)(x_i - \bar{x}) \\
&= \sum_{i=1}^{n}(y_i - \bar{y} - b_2(x_i - \bar{x}))(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = 0.
\end{aligned}$$

This establishes the lemma. ∎

Returning to the proof of Theorem 10.3.1, we have

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\beta_1-\beta_2x_i)^2 &= \sum_{i=1}^{n}(y_i-b_1-b_2x_i-(\beta_1-b_1)-(\beta_2-b_2)x_i)^2\\ &= \sum_{i=1}^{n}(y_i-b_1-b_2x_i)^2-2\sum_{i=1}^{n}(y_i-b_1-b_2x_i)\{(\beta_1-b_1)+(\beta_2-b_2)x_i\}\\ &\quad+\sum_{i=1}^{n}((\beta_1-b_1)+(\beta_2-b_2)x_i)^2\\ &= \sum_{i=1}^{n}(y_i-b_1-b_2x_i)^2+\sum_{i=1}^{n}((\beta_1-b_1)+(\beta_2-b_2x_i))^2,\end{aligned}$$

as the middle term is 0 by Lemma 10.6.1. Therefore,

$$\sum_{i=1}^{n}(y_i-\beta_1-\beta_2x_i)^2 \geq \sum_{i=1}^{n}(y_i-b_1-b_2x_i)^2$$

and $\sum_{i=1}^{n}(y_i-\beta_1-\beta_2x_i)^2$ takes its minimum value if and only if

$$\sum_{i=1}^{n}((\beta_1-b_1)+(\beta_2-b_2)x_i)^2=0.$$

This occurs if and only if $(\beta_1-b_1)+(\beta_2-b_2)x_i=0$ for every i. Because the x_i are not all the same value, this is true if and only if $\beta_1=b_1$ and $\beta_2=b_2$, which completes the proof. ∎

Proof of Theorem 10.3.2

We want to prove that, if $E(Y\mid X=x)=\beta_1+\beta_2x$ *and we observe the independent values* $(x_1,y_1),\ldots,(x_n,y_n)$ *for* (X,Y), *then*
(i) $E(B_1\mid X_1=x_1,\ldots,X_n=x_n)=\beta_1$,
(ii) $E(B_2\mid X_1=x_1,\ldots,X_n=x_n)=\beta_2$.

From Theorem 10.3.1 and $E(\bar{Y}\mid X_1=x_1,\ldots,X_n=x_n)=\beta_1+\beta_2\bar{x}$, we have that

$$\begin{aligned}E(B_2\mid X_1 &= x_1,\ldots,X_n=x_n)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_1+\beta_2x_i-\beta_1-\beta_2\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\ &= \beta_2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\beta_2.\end{aligned}$$

Also, from Theorem 10.3.1 and what we have just proved,

$$E(B_1\mid X_1=x_1,\ldots,X_n=x_n)=\beta_1+\beta_2\bar{x}-\beta_2\bar{x}=\beta_1.$$

∎

Proof of Theorem 10.3.3

We want to prove that, if $E(Y \mid X = x) = \beta_1 + \beta_2 x$, $Var(Y \mid X = x) = \sigma^2$ *for every* x, *and we observe the independent values* $(x_1, y_1), \ldots, (x_n, y_n)$ *for* (X, Y), *then*
(i) $Var(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2))$,
(ii) $Var(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2/\sum_{i=1}^n (x_i - \bar{x})^2$,
(iii) $Cov(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = -\sigma^2\bar{x}/\sum_{i=1}^n (x_i - \bar{x})^2$.

We first prove (ii). Observe that b_2 is a linear combination of the $y_i - \bar{y}$ values, so we can evaluate the conditional variance once we have obtained the conditional variances and covariances of the $Y_i - \bar{Y}$ values. We have that

$$Y_i - \bar{Y} = \left(1 - \frac{1}{n}\right) Y_i - \frac{1}{n} \sum_{j \neq i} Y_j,$$

so the conditional variance of $Y_i - \bar{Y}$ is given by

$$\sigma^2 \left(1 - \frac{1}{n}\right)^2 + \sigma^2 \frac{n-1}{n^2} = \sigma^2 \left(1 - \frac{1}{n}\right).$$

When $i \neq j$, we can write

$$Y_i - \bar{Y} = \left(1 - \frac{1}{n}\right) Y_i - \frac{1}{n} Y_j - \frac{1}{n} \sum_{k \neq i,j} Y_k,$$

and the conditional covariance between $Y_i - \bar{Y}$ and $Y_j - \bar{Y}$ is then given by

$$-2\sigma^2 \left(1 - \frac{1}{n}\right) \frac{1}{n} + \sigma^2 \frac{n-2}{n^2} = -\frac{\sigma^2}{n}$$

(note that you can assume that the means of the expectations of the Y's are 0 for this calculation). Therefore, the conditional variance of B_2 is given by

$$\begin{aligned}
&\mathrm{Var}(B_2 \mid x_1, \ldots, x_n) \\
&= \sigma^2 \left(1 - \frac{1}{n}\right) \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - \frac{\sigma^2}{n} \frac{\sum_{i \neq j} (x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},
\end{aligned}$$

because

$$\begin{aligned}
\sum_{i \neq j} (x_i - \bar{x})(x_j - \bar{x}) &= \left(\sum_{i=1}^n (x_i - \bar{x})\right)^2 - \sum_{i=1}^n (x_i - \bar{x})^2 \\
&= -\sum_{i=1}^n (x_i - \bar{x})^2.
\end{aligned}$$

For (iii), we have that

$$\begin{aligned}&\text{Cov}(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \text{Cov}(\bar{Y} - B_2\bar{X}, B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \text{Cov}(\bar{Y}, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) - \bar{x}\,\text{Var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\end{aligned}$$

and

$$\begin{aligned}&\text{Cov}(\bar{Y}, B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \frac{\sum_{i=1}(x_i - \bar{x})\text{Cov}((Y_i - \bar{Y}), \bar{Y} \mid X_1 = x_1, \ldots, X_n = x_n)}{\sum_{i=1}^n (x_i - \bar{x})^2}\\ &= \sigma^2(1 - 1/n)\frac{\sum_{i=1}(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0.\end{aligned}$$

Therefore, $\text{Cov}(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = -\sigma^2\bar{x}/\sum_{i=1}^n (x_i - \bar{x})^2$.

Finally, for (i), we have,

$$\begin{aligned}&\text{Var}(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \text{Var}(\bar{Y} - B_2\bar{x} \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \text{Var}(\bar{Y} \mid X_1 = x_1, \ldots, X_n = x_n) + \bar{x}^2\,\text{Var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\\ &\quad - 2\,\text{Cov}(\bar{Y}, B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\end{aligned}$$

where $\text{Var}(\bar{Y} \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2/n$. Substituting the results for (ii) and (iii) completes the proof of the theorem. ∎

Proof of Corollary 10.3.1

We need to show that

$$\text{Var}(B_1 + B_2x \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

For this, we have that

$$\begin{aligned}&\text{Var}(B_1 + B_2x \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \text{Var}(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) + x^2\,\text{Var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\\ &\quad + 2x\,\text{Cov}(B_1, B_2 \mid X_1 = x_1, \ldots, X_n = x_n)\\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2 + x^2 - 2x\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).\end{aligned}$$ ∎

Proof of Theorem 10.3.4

We want to show that, if $E(Y \mid X = x) = \beta_1 + \beta_2 x$, $Var(Y \mid X = x) = \sigma^2$ *for every* x, *and we observe the independent values* $(x_1, y_1), \ldots, (x_n, y_n)$ *for* (X, Y), *then*

$$E(S^2 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2.$$

We have that

$$\begin{aligned}
&E((n-2)S^2 \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= E\left(\sum_{i=1}^{n}(Y_i - B_1 - B_2 x_i)^2 \mid X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \sum_{i=1}^{n} E((Y_i - \bar{Y} - B_2(x_i - \bar{x}))^2 \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_{i=1}^{n} \mathrm{Var}(Y_i - \bar{Y} - B_2(x_i - \bar{x}) \mid X_1 = x_1, \ldots, X_n = x_n)
\end{aligned}$$

because

$$\begin{aligned}
&E(Y_i - \bar{Y} - B_2(x_i - \bar{x}) \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \beta_1 + \beta_2 x_i - \beta_1 - \beta_2 \bar{x} - \beta_2(x_i - \bar{x}) = 0.
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathrm{Var}(Y_i - \bar{Y} - B_2(x_i - \bar{x}) \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \mathrm{Var}(Y_i - \bar{Y} \mid X_1 = x_1, \ldots, X_n = x_n) \\
&\quad - 2(x_i - \bar{x})\,\mathrm{Cov}((Y_i - \bar{Y}), B_2 \mid X_1 = x_1, \ldots, X_n = x_n) \\
&\quad + (x_i - \bar{x})^2\,\mathrm{Var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n)
\end{aligned}$$

and, using the results established about the covariances of the $Y_i - \bar{Y}$ in the proof of Theorem 10.3.3, we have that

$$\mathrm{Var}(Y_i - \bar{Y} \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2(1 - 1/n)$$

and

$$\begin{aligned}
&\mathrm{Cov}(Y_i - \bar{Y}, B_2 \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{j=1}^{n}(x_j - \bar{x})\,\mathrm{Cov}(Y_i - \bar{Y}, Y_j - \bar{Y} \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left(\left(1 - \frac{1}{n}\right)(x_i - \bar{x}) - \frac{1}{n}\sum_{j \neq i}(x_j - \bar{x})\right) = \frac{\sigma^2(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},
\end{aligned}$$

because $\sum_{j \neq i}(x_j - \bar{x}) = -(x_i - \bar{x})$. Therefore,

$$\begin{aligned}
&\mathrm{Var}(Y_i - \bar{Y} - B_2(x_i - \bar{x}) \mid X_1 = x_1, \ldots, X_n = x_n) \\
&= \sigma^2\left(1 - \frac{1}{n}\right) - 2\frac{\sigma^2(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sigma^2(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)
\end{aligned}$$

and

$$E(S^2 \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{\sigma^2}{n-2} \sum_{i=1}^{n} \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \sigma^2,$$

as was stated. ∎

Proof of Lemma 10.3.1

We need to show that, if $(x_1, y_1), \ldots, (x_n, y_n)$ *are such that* $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$, *then*

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = b_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2.$$

We have that

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\
&= \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i + b_1 + b_2 x_i)^2 - n\bar{y}^2 \\
&= \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2 + \sum_{i=1}^{n}(b_1 + b_2 x_i)^2 - n\bar{y}^2
\end{aligned}$$

because $\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)(b_1 + b_2 x_i) = 0$ by Lemma 10.6.1. Then, using Theorem 10.3.1, we have

$$\sum_{i=1}^{n}(b_1 + b_2 x_i)^2 - n\bar{y}^2 = \sum_{i=1}^{n}(\bar{y} + b_2(x_i - \bar{x}))^2 - n\bar{y}^2 = b_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

and this completes the proof. ∎

Proof of Theorem 10.3.6

We want to show that, if Y, *given* $X = x$, *is distributed* $N(\beta_1 + \beta_2 x, \sigma^2)$ *and we observe the independent values* $(x_1, y_1), \ldots, (x_n, y_n)$ *for* (X, Y), *then the conditional distributions of* B_1, B_2, *and* S^2, *given* $X_1 = x_1, \ldots, X_n = x_n$, *are as follows.*

(i) $B_1 \sim N(\beta_1, \sigma^2(1/n + \bar{x}^2/\sum_{i=1}^{n}(x_i - \bar{x})^2))$

(ii) $B_2 \sim N(\beta_2, \sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2)$

(iii)

$$B_1 + B_2 x \sim N\left(\beta_1 + \beta_2 x, \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

(iv) $(n-2)\,S^2/\sigma^2 \sim \chi^2(n-2)$ *independent of* (B_1, B_2)

We first prove (i). Because B_1 can be written as a linear combination of the Y_i, Theorem 4.6.1 implies that the distribution of B_1 must be normal. The result then follows from Theorems 10.3.2 and 10.3.3. A similar proof establishes (ii) and (iii). The proof of (iv) is similar to the proof of Theorem 4.6.6, and we leave this to a further course in statistics. ∎

Proof of Corollary 10.3.2

We want to show
(i) $(B_1 - \beta_1)/S\left(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2} \sim t(n-2)$

(ii) $(B_2 - \beta_2)\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}/S \sim t(n-2)$

(iii)
$$\frac{B_1 + B_2 x - \beta_1 - \beta_2 x}{S\left((1/n + (x-\bar{x})^2)/\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}} \sim t(n-2)$$

(iv) *If F is defined as in (10.3.8), then $H_0 : \beta_2 = 0$ is true if and only if* $F \sim F(1, n-2)$.

We first prove (i). Because B_1 and S^2 are independent

$$\frac{B_1 - \beta_1}{\sigma\left(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}} \sim N(0,1)$$

independent of $(n-2)S^2/\sigma^2 \sim \chi^2(n-2)$. Therefore, applying Definition 4.6.2, we have

$$\frac{B_1 - \beta_1}{\sigma\left(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}\left((n-2)\,S^2/(n-2)\,\sigma^2\right)^{1/2}}$$
$$= \frac{B_1 - \beta_1}{S\left(1/n + \bar{x}^2/\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}} \sim t(n-2).$$

For (ii), the proof proceeds just as in the proof of (i).

For (iii), the proof proceeds just as in the proof of (i) and also using Corollary 10.3.1.

We now prove (iv). Taking the square of the ratio in (ii) and applying Theorem 4.6.11 implies

$$G = \frac{(B_2 - \beta_2)^2}{S^2\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{-1}} = \frac{(B_2 - \beta_2)^2 \sum_{i=1}^n (x_i - \bar{x})^2}{S^2} \sim F(1, n-2).$$

Now observe that F defined by (10.3.8) equals G when $\beta_2 = 0$. The converse that $F \sim F(1, n-2)$ only if $\beta_2 = 0$ is somewhat harder to prove and we leave this to a further course. ∎