Exercises

Homework exercises appear at the end of each chapter (except for the "Additional Topic..." chapters which have exercises after each independent topic). They are grouped into one of four categories:

- **Conceptual exercises** - These questions are quick to answer and require minimal (if any) calculations. They let students practice and assess basic terminology and concepts introduced in the chapter.

- **Guided exercises** - These exercises ask students to perform various stages of an analysis process with prompts for the individual steps.

- **Open-ended exercises** - These problems tend to ask for more complete analyses, without much or any step-by-step direction.

- **Supplemental exercises** - Topics for these exercises go somewhat beyond the scope of the material covered in the chapter.

Chapter 0

What Is a Statistical Model?

The unifying theme of this book is the use of models in statistical data analysis. Statistical models are useful for answering all kinds of questions. For example:

- Can we use the number of miles that a used car has been driven to predict the price that is being asked for the car? How much less can we expect to pay for each additional 1000 miles that the car has been driven? Would it be better to base our price predictions on the age of the car in years, rather than its mileage? Is it helpful to consider both age and mileage, or do we learn about as much about price by considering only one of these? Would the impact of mileage on the predicted price be different for a Honda as opposed to a Porsche?

- Do babies begin to walk at an earlier age if they engage in a regimen of special exercises? Or does any kind of exercise suffice? Or does exercise have no connection to when a baby begins to walk?

- If we find a footprint and a handprint at the scene of a crime, are they helpful for predicting the height of the person who left them? How about for predicting whether the person is male or female?

- Can we distinguish among different species of hawks based solely on the lengths of their tails?

- Do students with a higher grade point average really have a better chance of being accepted to medical school? How much better? How well can we predict whether or not an applicant is accepted, based on his/her GPA? Is there a difference between male and female students' chances for admission? If so, does one sex retain its advantage even after GPA is accounted for?

- Can a handheld device that sends a magnetic pulse into the head reduce pain for migraine sufferers?

- When people serve ice cream to themselves, do they take more if they are using a bigger bowl? What if they are using a bigger spoon?

- Which is more strongly related to average score for professional golfers: driving distance, driving accuracy, putting performance, or iron play? Are all of these useful for predicting a golfer's average score? Which are most useful? How much of the variability in golfers' scores can be explained by knowing all of these other values?

These questions reveal several purposes of statistical modeling:

1. **Making predictions.** Examples include predicting the price of a car based on its age, mileage, and model; predicting the length of a hawk's tail based on its species; predicting the probability of acceptance to medical school based on grade point average.

2. **Understanding relationships.** For example, after taking mileage into account, how is the age of a car related to its price? How does the relationship between foot length and height differ between men and women? How are the various measures of a golfer's performance related to each other and to the golfer's scoring average?

3. **Assessing differences.** For example, is the difference in ages of first walking different enough between an exercise group and a control group to conclude that exercise really does affect age of first walking? Is the rate of headache relief for migraine sufferers who experience a magnetic pulse sufficiently higher than those in the control group to advocate for the magnetic pulse as an effective treatment?

As with all models, statistical models are simplifications of reality. George Box, a renowned statistician, famously said that "all statistical models are wrong, but some are useful." Statistical models are not deterministic, meaning that their predictions are not expected to be perfectly accurate. For example, we do not expect to predict the exact price of a used car based on its mileage. Even if we were to record every imaginable characteristic of the car and include them all in the model, we would still not be able to predict its price exactly. And we certainly do not expect to predict the exact moment that a baby first walks based on the kind of exercise she/he engaged in. Statistical models merely aim to explain as much of the variability as possible in whatever phenomenon is being modeled. In fact, because human beings are notoriously variable and unpredictable, social scientists who develop statistical models are often delighted if the model explains even a small part of the variability.

A distinguishing feature of statistical models is that we pay close attention to possible simplifications and imperfections, seeking to quantify how much the model explains and how much it does not. So, while we do not expect our model's predictions to be exactly correct, we are able to state how confident we are that our predictions fall within a certain range of the truth. And while we do not expect to determine the exact relationship between two variables, we can quantify how far off our model is likely to be. And while we do not expect to assess exactly how much two groups may differ, we can draw conclusions about how likely they are to differ and by what magnitude.

More formally, a statistical model can be written as:

$$DATA = MODEL + ERROR$$

or as

$$Y = f(X) + \epsilon$$

The Y here represents the variable being modeled, X is the variable used to do the modeling, and f is a function[1]. We start in Chapter 1 with just one quantitative, explanatory variable X and with a linear function f. Then we will consider more complicated functions for f, often by transforming Y or X or both. Later we will consider multiple explanatory variables, which can be either quantitative or categorical. In these initial models we assume that the response variable Y is quantitative. Eventually, we will allow the response variable Y to be categorical.

The ϵ term in the model above is called the "error," meaning the part of the response variable Y that remains unexplained after considering the predictor X. Our models will sometimes stipulate a probability distribution for this ϵ term, often a normal distribution. An important aspect of our modeling process will be checking whether the stipulated probability distribution for the error term seems reasonable, based on the data, and making appropriate adjustments to the model if it does not.

0.1 Fundamental Terminology

Before you begin to study statistical modeling, you will find it very helpful to review and practice applying some fundamental terminology.

The **observational units** in a study are the people, objects, or cases on which data are recorded. The **variables** are the characteristics that are measured or recorded about each observational unit.

Example 0.1: *Car Prices*

In the study about predicting the price of a used car, the observational units are the cars. The variables are the car's price, mileage, age (in years), and manufacturer (Porsche or Honda).

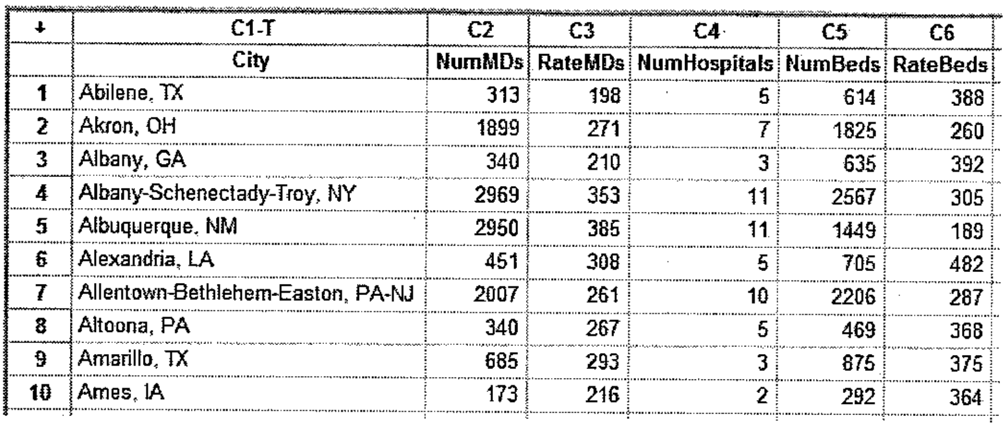
◇

Example 0.2: *Walking Babies*

In the study about babies' walking, the observational units are the babies. The variables are whether or not the baby was put on an exercise regimen, and the age at which the baby first walked.

◇

[1] The term "model" is used to refer to the entire equation or just the structural part that we have denoted by $f(X)$



↓	C1-T	C2	C3	C4	C5	C6
	City	NumMDs	RateMDs	NumHospitals	NumBeds	RateBeds
1	Abilene, TX	313	198	5	614	388
2	Akron, OH	1899	271	7	1825	260
3	Albany, GA	340	210	3	635	392
4	Albany-Schenectady-Troy, NY	2969	353	11	2567	305
5	Albuquerque, NM	2950	385	11	1449	189
6	Alexandria, LA	451	308	5	705	482
7	Allentown-Bethlehem-Easton, PA-NJ	2007	261	10	2206	287
8	Altoona, PA	340	267	5	469	368
9	Amarillo, TX	685	293	3	875	375
10	Ames, IA	173	216	2	292	364

Figure 0.1: Health facilities in U.S. metropolitan areas

Example 0.3: *Metropolitan Health Care*

You may find it helpful to envision the data in a spreadsheet format. The row labels are cities, which are observational units, and the columns correspond to the variables. For example, Figure 0.1 shows part of a Minitab worksheet with data compiled by the U.S. Census Bureau on health care facilities in metropolitan areas. The observational units are the metropolitan areas and the variables count the number of doctors, hospitals and beds in each city as well as rates (number of doctors or beds per 100,000 residents). The full dataset for 83 metropolitan areas is in the file **MetroHealth83** .

◇

Variables can be classified into two types: quantitative and categorical. A **quantitative** variable records numbers about the observational units. It must be sensible to perform ordinary arithmetic operations on these numbers, so zip codes and jersey numbers are not quantitative variables. A **categorical** variable records a category designation about the observational units. If there are only two possible categories, the variable is also said to be **binary**.

Example 0.1 (continued): The price, mileage, and age of a car are all *quantitative* variables. The model of the car is a *categorical* variable.

◇

Example 0.2 (continued): Whether or not a baby was assigned to a special exercise regimen is a *categorical* variable. The age at which the baby first walked is a *quantitative* variable.

◇

Example 0.4: *Medical School Admission*

Whether or not an applicant is accepted for medical school is a *binary* variable, as is the gender of the applicant. The applicant's undergraduate grade point average is a *quantitative* variable.

◇

Another important consideration is the role played by each variable in the study. The variable that measures the outcome of interest is called the **response** variable. The variables whose relationship to the response is being studied are called **explanatory** variables. (When the primary goal of the model is to make predictions, the explanatory variables are also called **predictor** variables.)

Example 0.1 (continued): The price of the car is the *response* variable. The mileage, age, and model of the car are all *explanatory* variables. ◇

Example 0.2 (continued): The age at which the baby first walked is the *response* variable. Whether or not a baby was assigned to a special exercise regimen is an *explanatory* variable. ◇

Example 0.4 (continued): Whether or not an applicant is accepted for medical school is the *response* variable. The applicant's undergraduate grade point average and sex are *explanatory* variables. ◇

One reason that these classifications are important is that the choice of the appropriate analysis procedure depends on the type of variables in the study and their roles. Regression analysis (covered in Chapters 1 - 4) is appropriate when the response variable is quantitative and the explanatory variables are also quantitative. In Chapter 3 you will also learn how to incorporate binary explanatory variables into a regression analysis. Analysis of variance (ANOVA, considered in Chapters 5-8) is appropriate when the response variable is quantitative but the explanatory variables are categorical. When the response variable is categorical, logistic regression (considered in Chapters 9-11) can be used with either quantitative or categorical explanatory variables. These various scenarios are displayed in Table 0.1.

Keep in mind that variables are not always clear-cut to measure or even classify. For example, measuring headache relief is not a straightforward proposition and could be done with a quantitative measurement (intensity of pain on a 0-10 scale), a categorical scale (much relief, some relief, no relief) or as a binary categorical variable (relief or not).

Table 0.1: Classifying general types of models

Response	Predictor/explanatory	Procedure	Chapter
Quantitative	Single quantitative	Simple linear regression	1,2
Quantitative	Single categorical	One-way analysis of variance	5
Categorical	Single quantitative	Simple logistic regression	9
Categorical	Single binary	$2x2$ table	9
Quantitative	Multiple quantitative	Multiple linear regression	3,4
Quantitative	Multiple categorical	Multi-way analysis of variance	6,7
Categorical	Multiple quantitative	Multiple logistic regression	10,11
Categorical	Multiple categories	$2xk$ table	11

We collect data and fit models in order to understand **populations**, such as all students who are applying to medical school, and **parameters**, such as the acceptance rate of all students with a grade point average of 3.5. The collected data are a **sample** and a characteristic of a sample, such as the percentage of students with grade point averages of 3.5 who were admitted to medical school, out of those who applied, is a **statistic**. Thus, sample statistics are used to estimate population parameters.

Another crucial distinction is whether a research study is a controlled experiment or an observational study. In a **controlled experiment**, the researcher manipulates the explanatory variable by assigning the explanatory group or value to the observational units. (These observational units may be called **experimental units** or **subjects** in an experiment.) In an **observational study**, the researchers do not assign the explanatory variable but rather passively observe and record its information. This distinction is important because the type of study determines the scope of conclusion that can be drawn. Controlled experiments allow for drawing *cause-and-effect* conclusions. Observational studies, on the other hand, only allow for concluding that variables are *associated.* Ideally, an observational study will anticipate alternative explanations for an association and include the additional relevant variables in the model. These additional explanatory variables are then called **covariates**.

Example 0.5: *Handwriting and SAT Essay Scores*

An article about handwriting appeared in the October 11, 2006 issue of the Washington Post. The article mentioned that among students who took the essay portion of the SAT exam in 2005-6, those who wrote in cursive style scored significantly higher on the essay, on average, than students who used printed block letters. This is an example of an observational study since there was no controlled assignment of the type of writing for each essay. While it shows an association between handwriting and essay scores, we can't tell whether better writers tend to choose to write in cursive or if graders tend to score cursive essays more generously and printed ones more harshly. We might also suspect that students with higher GPAs are more likely to use cursive writing. To examine this carefully, we could fit a model with GPA as a covariate.

The article also mentioned a different study in which the identical essay was shown to many graders, but some graders were shown a cursive version of the essay and the other graders were shown a version with printed block letters. Again, the average score assigned to the essay with the cursive style was significantly higher than the average score assigned to the essay with the printed block letters. This second study involved an experiment since the binary explanatory factor of interest (cursive versus block letters) was controlled by the researchers. In that case we can infer that using cursive writing produces better essay scores, on average, than printing block letters.

◇

0.2 Four-Step Process

We will employ a four-step process for statistical modeling throughout this book. These steps are:

- **Choose** a form for the model. This involves identifying the response and explanatory variable(s) and their types. We usually examine graphical displays to help suggest a model that might summarize relationships between these variables.

- **Fit** that model to the data. This usually entails estimating model parameters based on the sample data. We will almost always use statistical software to do the necessary number-crunching to fit models to data.

- **Assess** how well the model describes the data. One component of this involves comparing the model to other models. Are there elements of the model that are not very helpful in explaining the relationships or do we need to consider a more complicated model? Another component of the assessment step concerns analyzing residuals, which are deviations between the actual data and the model's predictions, to assess how well the model fits the data. This process of assessing model adequacy is as much art as science.

- **Use** the model to address the question that motivated collecting the data in the first place. This might be to make predictions, or explain relationships, or assess differences, bearing in mind possible limitations on the scope of inferences that can be made. For example, if the data were collected as a random sample from a population, then inference can be extended to that population; if treatments were assigned at random to subjects, then a cause-and-effect relationship can be inferred; but if the data arose in other ways, then we have little statistical basis for drawing such conclusions.

The specific details for how to carry out these steps will differ depending on the type of analysis being performed and, to some extent, on the context of the data being analyzed. But these four steps are carried out in some form in all statistical modeling endeavors. To illustrate the process we consider an example in the familiar setting of a two-sample t-procedure.

Example 0.6: *Financial Incentives for Weight Loss*

Losing weight is an important goal for many individuals. An article in the *Journal of the American Medical Association* describes a study (Volpp et. al., 2008) in which researchers investigated whether financial incentives would help people lose weight more successfully. Some participants in the study were randomly assigned to a treatment group which offered financial incentives for achieving weight loss goals, while others were assigned to a control group that did not use financial incentives. All participants were monitored over a four-month period and the net weight change ($Before - After$ in pounds) at the end of this period was recorded for each individual. Note that a positive value corresponds to a weight loss and a negative change is a weight gain. The data are given in Table 0.2 and stored in the file **WeightLossIncentive**.

Table 0.2: Weight Loss After Four Months (pounds)

Control	12.5	12.0	1.0	-5.0	3.0	-5.0	7.5	-2.5	20.0	-1.0
	2.0	4.5	-2.0	-17.0	19.0	-2.0	12.0	10.5	5.0	
Incentive	25.5	24.0	8.0	15.5	21.0	4.5	30.0	7.5	10.0	18.0
	5.0	-0.5	27.0	6.0	25.5	21.0	18.5			

Source: Volpp et. al., *JAMA* 2008

The response variable in this situation (weight change) is quantitative and the explanatory factor of interest (control versus incentive) is categorical and binary. The subjects were assigned to the groups at random so this is a statistical experiment. Thus we may investigate whether there is a statistically significant difference in the distribution of weight changes due to the use of a financial incentive.

CHOOSE

When choosing a model we generally consider the question of interest and types of variables involved, then look at graphical displays and compute summary statistics for the data. Since the weight loss incentive study has a binary explanatory factor and quantitative response, we examine dotplots of the weight losses for each of the two groups (Figure 0.2) and find the sample mean and standard deviation for each group.

```
Variable    Group       N   Mean  StDev
WeightLoss  Control    19   3.92   9.11
            Incentive  17  15.68   9.41
```

The dotplots show a pair of reasonably symmetric distributions with roughly the same variability, although the mean weight loss for the incentive group is larger than the mean for the control group. One model for these data would be for the weight losses to come from a pair of normal distributions, with different means (and perhaps different standard deviations) for the two groups. Let the parameter μ_1 denote the mean weight loss after four months without a financial incentive, μ_2 be the mean with the incentive. If σ_1 and σ_2 are the respective standard deviations and we

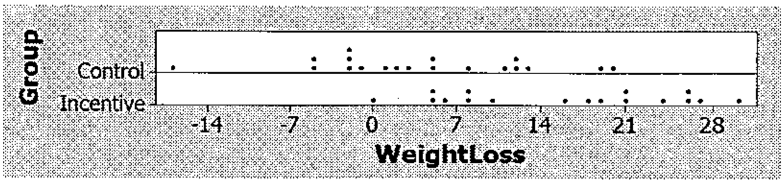


Figure 0.2: Weight Loss for Control versus Incentive groups

let the variable Y denote the weight losses, we can summarize the model as $Y \sim N(\mu_i, \sigma_i)$ where the subscript indicates the group membership[2] and the symbol $\sim$ signifies that the variable has a particular distribution. To see this in the $DATA = MODEL + ERROR$ format, this model could also be written as

$$Y = \mu_i + \epsilon$$

where μ_i is the population mean for the i^{th} group and $\epsilon \sim N(0, \sigma_i)$ is the random error term. Since we only have two groups this model says that

$$\begin{array}{ll} Y = \mu_1 + \epsilon \sim N(\mu_1, \sigma_1) & \text{for individuals in the control group.} \\ Y = \mu_2 + \epsilon \sim N(\mu_2, \sigma_2) & \text{for individuals in the incentive group.} \end{array}$$

FIT

To fit this model we need to estimate four parameters (the means and standard deviations for each of the two groups) using the data from the experiment. The observed means and standard deviations from the two samples provide obvious estimates. We let $\overline{y}_1 = 3.92$ estimate the mean weight loss for the control group and $\overline{y}_2 = 15.68$ estimate the mean for a population getting the incentive. Similarly $s_1 = 9.11$ and $s_2 = 9.41$ estimate the respective standard deviations. The fitted model (a prediction for the typical weight loss in either group) can then be expressed as [3]

$$\hat{y} = \overline{y}_i$$

i.e. that $\hat{y} = 3.92$ pounds for individuals without the incentive and $\hat{y} = 15.68$ pounds for those with the incentive.

Note that the error term does not appear in the fitted model since, when predicting a particular weight loss, we don't know whether the random error will be positive or negative. That does not mean that we expect there to be no error, just that the best guess for the *average* weight loss under either condition is the sample group mean, $\overline{y}_i$.

ASSESS

Our model indicates that departures from the mean in each group (the random errors) should follow a normal distribution with mean zero. To check this we examine the sample *residuals* or deviations between what is predicted by the model and the actual data weight losses.

$$residual = \text{observed} - \text{predicted} = y - \hat{y}$$

For subjects in the control group we subtract $\hat{y} = 3.92$ from each weight loss and we subtract $\hat{y} = 15.68$ for the incentive group. Dotplots of the residuals for each group are shown in Figure 0.3.

[2] For this example an assumption that the variances are equal, $\sigma_1^2 = \sigma_2^2$, might be reasonable, but that would lead to the less familiar pooled variance version of the t-test. We explore this situation in more detail in a later chapter.

[3] We use the ^ symbol above a variable name to indicate predicted value, and refer to this as y-hat.

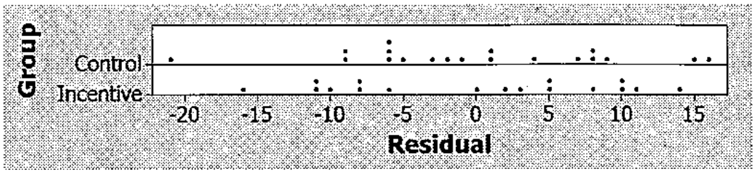
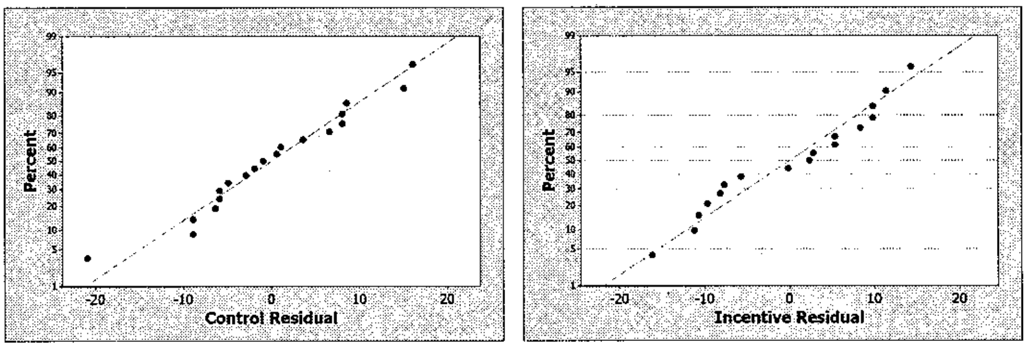


Figure 0.3: Residuals from Group Weight Loss Means

Note that the distributions of the residuals are the same as the original data, except that both are shifted to have a mean of zero. We don't see any significant departures from normality in the dotplots, but it's difficult to judge normality from dotplots with so few points. Normal probability plots (as shown in Figure 0.4) are a more informative technique for assessing normality. Departures from a linear trend in such plots indicate a lack of normality in the data. Normal probability plots will be examined in more detail in the next chapter.



(a) Control

(b) Incentive

Figure 0.4: Normality Probability Plots for Residuals of Weight Loss

As a second component of assessment, we consider whether an alternate (simpler) model might fit the data essentially as well as our model with different means for each group. This is analogous to testing the standard hypotheses for a two sample t-test:

$H_o : \mu_1 = \mu_2$
$H_a : \mu_1 \neq \mu_2$

The null hypothesis (H_o) corresponds to the simpler model $Y = \mu + \epsilon$ which uses the same mean for both the control and incentive groups. The alternative (H_a) reflects the model we have considered here that allows each group to have a different mean. Would the simpler (common mean) model suffice for the weight loss data or do the two separate group means provide a *significantly* better explanation for the data? One way to judge this is with the results of the usual two-sample t-test (as shown in the Minitab output below).

```
Two-sample T for WeightLoss

Group       N   Mean  StDev  SE Mean
Control    19   3.92   9.11      2.1
Incentive  17  15.68   9.41      2.3

Difference = mu (Control) - mu (Incentive)
Estimate for difference:  -11.76
95% CI for difference:  (-18.05, -5.46)
T-Test of difference = 0 (vs not =): T-Value = -3.80  P-Value = 0.001  DF = 33
```

The extreme value for this test statistic ($t = -3.80$) and very small p-value (0.001) provide strong evidence that the means of the two groups are indeed significantly different. If the two group means were really the same (i.e. the common mean model was accurate and the financial incentives had no effect on weight loss) we would expect to see a difference as large as was observed in this experiment for only about 1 in 1000 replications of the experiment. Thus the model with separate means for each group does a substantially better job at explaining the results of the weight loss study.

USE

Since this was a designed experiment with random allocation of the control and incentive conditions to the subjects, we can infer that the financial incentives did produce a difference in the average weight loss over the four-month period; that is, the random allocation of conditions to subjects allows us to draw a cause-and-effect relationship. A person who is on the incentive-based treatment can be expected to lose about 11.8 pounds more ($15.68 - 3.92 = 11.76$), on average, in four months, than control subjects who are not given this treatment. Note that for most individuals, approximately 12 pounds is a substantial amount of weight to lose in four months. Moreover, if we interpret the confidence interval from the Minitab output, we can be 95% confident that the incentive treatment is worth between 5.5 and 18.1 pounds of additional weight loss, on average, over four months.

Before leaving this example we note three cautions. First, all but two of the participants in this study were adult men, so we should avoid making conclusions about the effect of financial incentives on weight loss in women. Second, if the participants in the study did not arise from taking a random sample, we would have difficulty justifying a statistical basis for generalizing the findings to other adults. Any such generalization must be justified on other grounds (such as a belief that most adults respond to financial incentives in similar ways). Third, the experimenters followed up with subjects to see if weight losses were maintained at a point seven months after the start of the study (and three months after any incentives expired). The results from the followup study appear in Exercise 0.10.

◇

0.3 Chapter Summary

In this chapter we reviewed basic terminology, introduced the 4-step approach to modeling that will be used throughout the text, and revisited a common two-sample inference problem.

After completing this chapter you should be able to distinguish between a **sample** and a **population**, describe the difference between a **parameter** and a **statistic**, and identify variables as **categorical** or **quantitative**. Prediction is a major component to modeling so identifying **explanatory** (or predictor) **variables** that can be used to develop a model for **response variable** is an important skill. Another important idea is the distinction between **observational studies** (where researchers simply observe what is happening) and **experiments** (where researchers impose "treatments").

The fundamental idea that a **statistical model** partitions data into two components, one for the model and one for error, was introduced. Even though the models will get more complex as we move through the more advanced settings, this statistical modeling idea will be a major theme throughout the text. The error term and conditions associated with this term are important features in distinguishing statistical models from mathematical models. You saw how to compute **residuals** by comparing the observed data to predictions from a model as a way to begin quantifying the errors.

The **4-step process** of choosing, fitting, assessing, and using a model is vital. Each step in the process requires careful thought and the computations will often be the easiest part of the entire process. Identifying the response and explanatory variable(s) and their types (categorical or quantitative) helps us **choose** the appropriate model(s). Statistical software will almost always be used to **fit** models and obtain estimates. Comparing models and **assessing** the adequacy of these models will require a considerable amount of practice and this is a skill that you will develop over time. Try to remember that **using** the model to make predictions, explain relationships, or assess differences is only one part of the 4-step process.

0.4 Exercises

Conceptual Exercises

0.1 *Categorical or quantitative?* Suppose that a statistics professor records the following for each student enrolled in her class:

- Gender
- Major
- Score on first exam
- Number of quizzes taken (a measure of class attendance)
- Time spent sleeping on the previous night
- Handedness (left- or right-handed)
- Political inclination (liberal, moderate, or conservative)
- Time spent on the final exam
- Score on the final exam

For the following questions, identify the response variable and the explanatory variable(s). Also classify each variable as quantitative or categorical. For categorical variables, also indicate whether the variable is binary.

a. Do the various majors differ with regard to average sleeping time?

b. Is a student's score on the first exam useful for predicting his/her score on the final exam?

c. Do male and female students differ with regard to the average time they spend on the final exam?

d. Do the proportions of left-handers differ between males and females on campus?

e. Are sleeping time, exam 1 score, and number of quizzes taken useful for predicting time spent on final exam?

f. Does knowing a student's gender help to predict his/her major?

g. Does knowing a student's political inclination and time spent sleeping help to predict his/her gender?

0.2 *Sports projects.* For each of the following sports-related projects, identify observational units and the response and explanatory variables when appropriate. Also classify the variables as quantitative or categorical.

a. Interested in predicting how long it takes to play a Major League Baseball game, an individual recorded the following information for all 15 games played on August 26, 2008: time to complete the game, total number of runs scored, margin of victory, total number of pitchers used, ballpark attendance at the game, and which league (National or American) the game was played in.

b. Over the course of several years, a golfer kept track of the length of all of his putts and whether or not he made the putt. He was interested in predicting whether or not he would make a putt based on how long it was.

c. Some students recorded lots of information about all of the football games played by LaDainian Tomlinson during the 2006 season. They recorded his rushing yardage, number of rushes, rushing touchdowns, receiving yardage, number of receptions, and receiving touchdowns.

d. A volleyball coach wants to see if a player using a jump serve is more likely to lead to winning a point than when using a standard overhand serve.

e. To investigate whether the "home-field advantage" differs across major team sports, researchers kept track of how often the home team won a game for all games played in the 2007 and 2008 seasons in Major League Baseball, National Football League, National Basketball Association, and National Hockey League.

f. A student compared men and women professional golfers on how far they drive a golf ball and how often their drive hits the fairway.

0.3 *Scooping ice cream.* In a study reported in the *Journal of Preventative Medicine*, 85 nutrition experts were asked to scoop themselves as much ice cream as they wanted. Some of them were randomly given a large bowl (34 ounces) as they entered the line, and the others were given a smaller bowl (17 ounces). Similarly, some were randomly given a large spoon (3 ounces) and the others were given a small spoon (2 ounces). Researchers then recorded how much ice cream each subject scooped for herself/himself. Their conjecture was that those given a larger bowl would tend to scoop more ice cream, as would those given a larger spoon.

a. Identify the observational units in this study.

b. Is this an observational study or a controlled experiment? Explain how you know.

c. Identify the response variable in this study, and classify it as quantitative or categorical.

d. Identify the explanatory variables in this study, and classify it/them as quantitative or categorical.

0.4 *Wine model.* In his book *SuperCrunchers: Why Thinking by Numbers is the New Way to be Smart*, Ian Ayres writes about Orley Achenfelter, who has gained fame and generated considerable controversy by using statistical models to predict the quality of wine. Ashenfelter developed a model based on decades of data from France's Bordeaux region, which Ayers reports as:

$$\text{WineQuality} = 12.145 + .00117\text{WinterRain} + .0614\text{AverageTemp} - .00386\text{HarvestRain} + \epsilon$$

where *WineQuality* is a function of the price, rainfall is measured in millimeters, and temperature is measured in °C.

a. Identify the response variable in this model. Is it quantitative or categorical?

b. Identify the explanatory variables in this model. Are they quantitative or categorical?

c. According to this model, is higher wine quality associated with more or with less winter rainfall?

d. According to this model, is higher wine quality associated with more or with less harvest rainfall?

e. According to this model, is higher wine quality associated with more or with less average growing season temperature?

f. Are the data that Ashenfelter analyzed observational or experimental? Explain.

0.5 *Measuring students.* The registrar at a small liberal arts college computes descriptive summaries for all members of the entering class on a regular basis. For example, the mean and standard deviation of the high school GPAs for all entering students in particular year were 3.16 and 0.5247, respectively. The Mathematics Department is interested in helping all students who want to take mathematics to identify the appropriate course, so they offer a placement exam. A randomly selected subset of students taking this exam during the past decade had an average score of 71.05 with a standard deviation of 8.96.

a. What is the population of interest to the registrar at this college?

b. Are the descriptive summaries computed by the registrar (3.16 and 0.5247) statistics or parameters? Explain.

c. What is the population of interest to the Mathematics Department?

d. Are the numerical summaries (71.05 and 8.96) statistics or parameters? Explain.

Guided Exercises

0.6 *Scooping ice cream.* Refer to Exercise 0.3 on self-serving ice cream. The following table reports the average amounts of ice cream scooped (in ounces) for the various treatments:

	17-ounce bowl	34-ounce bowl
2-ounce spoon	4.38	5.07
3-ounce spoon	5.81	6.58

a. Does it appear that the size of the bowl had an effect on amount scooped? Explain.

b. Does it appear that the size of the spoon had an effect on amount scooped? Explain.

c. Which appears to have more of an effect: size of bowl or size of spoon? Explain.

d. Does it appear that the effect of the bowl size is similar for both spoon sizes, or does it appear that the effect of the bowl size differs substantially for the two spoon sizes? Explain.

0.7 *Diet plans.* An article in the *Journal of the American Medical Association* (Dansinger et al., 2005) reported on a study in which 160 subjects were randomly assigned to one of four popular diet plans: Atkins, Ornish, Weight Watchers, and Zone. Among the variables measured were:

- which diet the subject was assigned to
- whether or not the subject completed the twelve-month study
- the subject's weight loss after two months, six months, and twelve months (in kilograms, with a negative value indicating weight gain)
- the degree to which the subject adhered to the assigned diet, taken as the average of 12 monthly ratings, each on a 1-10 scale (with 1 indicating complete non-adherence and 10 indicating full adherence)

a. Classify each of these variables as quantitative or categorical.

b. The primary goal of the study was to investigate whether weight loss tends to differ significantly among the four diets. Identify the explanatory and response variables for investigating this question.

c. A secondary goal of the study was to investigate whether weight loss is affected by the adherence level. Identify the explanatory and response variables for investigating this question.

d. Is this an observational study or a controlled experiment? Explain how you know.

e. If the researchers' analysis of the data leads them to conclude that there is a significant difference in weight loss among the four diets, can they legitimately conclude that the difference is because of the diet? Explain why or why not.

f. If the researchers' analysis of the data analysis leads them to conclude that there is a significant association between weight loss and adherence level, can they legitimately conclude that there is a cause-and-effect association between them? Explain why or why not.

0.8 *Predicting NFL wins.* Consider the following model for predicting the number of games that a National Football League (NFL) team wins in a season:

$$wins = 4.6 + 0.5PF - 0.3PA + \epsilon$$

where PF stands for average points a team scores per game over an entire season and PA stands for points allowed per game. Currently each team plays 16 games in a season.

a. According to this model, how many more wins is a team expected to achieve in a season if they increase their scoring by an average of 3 points per game?

b. According to this model, how many more wins is a team expected to achieve in a season if they decrease their points allowed by an average of 3 points per game?

c. Based on your answers to (a) and (b), does it seem that a team should focus more on improving its offense or improving its defense?

d. Use this model to predict the number of wins for the 2009 New Orleans Saints, who scored 510 points and allowed 341 points.

e. The Saints actually won 13 games in 2009. Determine their residual from this model, and interpret what this means.

f. The largest residual value from this model belongs to the Indianapolis Colts, with a residual value of 3.21 games. The Colts actually won 14 games. Determine this model's predicted number of wins for the Colts.

g. The largest negative residual value from this model belongs to the Baltimore Ravens, with a residual value of -1.95 games. Interpret what this residual means.

0.9 *Roller coasters.* The Roller Coaster Database (*rcdb.com*) contains lots of information about roller coasters all over the world. The following statistical model for predicting the top speed (in miles per hour) of a coaster was based on more than 100 roller coasters in the United States and data displayed on the database in November of 2003:

$$TopSpeed = 54 + 7.6TypeCode + \epsilon$$

where $TypeCode = 1$ for steel roller coasters and $TypeCode = 0$ for wooden roller coasters.

a. What top speed does this model predict for a wooden roller coaster?

b. What top speed does this model predict for a steel roller coaster?

c. Determine the difference in predicted speeds in mph for the two types of coasters. Also identify where this number appears in the model equation, and explain why that makes sense.

Some other predictor variables available at the database include: age, total length, maximum height, and maximum vertical drop. Suppose that we include all of these predictor variables in a statistical model for predicting the top speed of the coaster.

d. For each of these predictor variables, indicate whether you expect its coefficient to be positive or negative. Explain your reasoning for each variable.

e. Which of these predictor variables do you expect to be the best single variable for predicting a roller coaster's top speed? Explain why you think that.

The following statistical model was produced from these data:

$$Speed = 33.4 + 0.10Height + 0.11Drop + 0.0007Length - 0.023Age - 2.0TypeCode + \epsilon$$

f. Comment on whether the signs of the coefficients are as you expect.

g. What top speed would this model predict for a steel roller coaster that is 10 years old, with maximum height of 150 feet, maximum vertical drop of 100 feet, and length of 4000 feet?

Open-ended Exercises

0.10 *Incentive for weight loss.* The study (Volpp et. al., 2008) on financial incentives for weight loss in Example 0.6 on page 7 used a follow-up weight check after seven months to see whether weight losses persisted after the original four months of treatment. The results are given in Table 0.3 and in the variable *Month7Loss* of the **WeightLossIncentive** data file. Note that a few participants dropped out and were not re-weighed at the seven-month point. As with the earlier example, the data are the change in weight (in pounds) from the beginning of the study and positive values correspond to weight losses. Using Example 0.6 as an outline, follow the four-step process to see whether the data provide evidence that the beneficial effects of the financial incentives still apply to the weight losses at the seven-month point.

Table 0.3: Weight Loss After Seven Months (pounds)

Control	-2.0	7.0	19.5	-0.5	-1.5	-10.0	0.5	5.0	8.5	$\overline{y}_1 = 4.64$
	18.0	16.0	-9.0	4.5	23.5	5.5	6.5	-9.5	1.5	$s_1 = 9.84$
Incentive	11.5	20.0	-22.0	2.0	7.5	16.5	19.0	18.0	-1.0	$\overline{y}_2 = 7.80$
	5.5	24.5	9.5	10.0	-8.5	4.5				$s_2 = 12.10$

Source: Volpp, John, Troxel, et. al., *JAMA* 2008

0.11 *Statistics students survey.* An instructor at a small liberal arts college distributed the data collection card similar to what is shown below on the first day of class. The data for two different sections of the course are shown in the file **Day1Survey**. Note that the names have not been entered into the data set.

Data Collection Card

Directions: Please answer each question and return to me.
1. Your name (as you prefer): ____
2. What is your current class standing? ______________________________
3. Sex: Male _____ Female _____
4. How many miles (approximately) did you travel to get to campus? ____________
5. Height (estimated) in inches: ____________________________________
6. Handedness (Left, Right, Ambidextrous): ___________________________
7. How much money, in coins (not bills), do you have with you? $______________
8. Estimate the length of the white string (in inches): _______________________
9. Estimate the length of the black string (in inches): _______________________
10. How much do you expect to read this semester (in pages/week)? ______________
11. How many hours do you watch TV in a typical week? __________________________
12. What is your resting pulse? _______________________________________
13. How many text messages have you sent and received in the last 24 hours? _____

The data for this survey are stored in **Day1Survey**.

a. Apply the four-step process to the survey data to address the question: "Is there evidence that the mean resting pulse rates for women is different from the mean resting pulse rate for men?"

b. Pick another question that interests you from the survey and compare the responses of men and women.

0.12 *Statistics student survey (continued).* Refer to the survey of statistics students described in Exercise 0.11 with data in **Day1Survey**. Use the survey data to address the question: "Do women expect to do more reading than men?"

0.13 *Marathon training.* Training records for a marathon runner are provided in the file **Marathon**. The *Date*, *Miles* run, *Time* (in minutes:seconds:hundredths), and running *Pace* (in minutes:seconds:hundredths per mile) are given for a five-year period from 2002 to 2006. The time and pace have been converted to decimal minutes in *TimeMin* and *PaceMin*, respectively. The brand of the running shoe is added for 2005 and 2006. Use the four-step process to investigate if a runner has a tendency to go faster on short runs (5 or less miles) than long runs. The variable *Short* in the dataset is coded with 1 for short runs and 0 for longer runs. Assume that the data for this runner can be viewed as a sample for runners of a similar age and ability level.

0.14 *Marathon training (continued).* Refer to the data described in Exercise 0.13 that contains five years worth of daily training information for a runner. One might expect that the running patterns might change as the runner gets older. The file **Marathon** also contains a variable called $After2004$ which has the value 0 for any runs during the years 2002-2004 and 1 for runs during 2005 and 2006. Use the four-step process to see if there is evidence of a difference between these two time periods in the following aspects of the training runs.

a. the average running pace ($PaceMin$).

b. the average distance run per day ($Miles$).

Supplementary Exercises

0.15 *Pythagorean theorem of baseball.* Renowned baseball statistician Bill James devised a model for predicting a team's winning percentage. Dubbed the "Pythagorean Theorem of Baseball," this model predicts a team's winning percentage as:

$$\text{Winning Percentage} = \frac{(\text{runs scored})^2}{(\text{runs scored})^2 + (\text{runs against})^2} \times 100 + \epsilon$$

a. Use this model to predict the winning percentage for the New York Yankees, who scored 915 runs and allowed 753 runs in the 2009 season.

b. The New York Yankees actually won 103 games and lost 59 in the 2009 season. Determine the winning percentage, and also determine the residual from the Pythagorean model (by taking the observed winning percentage minus the predicted winning percentage).

c. Interpret what this residual value means for the 2009 Yankees. [Hints: Did the team do better or worse than expected, given their runs scored and runs allowed? By how much?]

d. Repeat (a)-(c) for the 2009 San Diego Padres, who scored 638 runs and allowed 769 runs.

e. Which team (Yankees or Padres) exceeded their Pythagorean expectations by more?

Table 0.4 provides data, predictions, and residuals for all 30 Major League Baseball teams in 2009.

f. Which team exceeded their Pythagorean expectations the most? Describe how this team's winning percentage compares to what is predicted by their runs scores and runs allowed.

g. Which team fell furthest below their Pythagorean expectations? Describe how this team's winning percentage compares to what is predicted by their runs scored and runs allowed.

Table 0.4: Winning percentage and Pythagorean predictions for baseball teams in 2009

TEAM	W	L	WinPct	RunScored	RunsAgainst	Predicted	Residual
Arizona Diamondbacks	70	92	43.21	720	782	45.88	-2.67
Atlanta Braves	86	76	53.09	735	641	56.80	-3.71
Baltimore Orioles	64	98	39.51	741	876	41.71	-2.20
Boston Red Sox	95	67	58.64	872	736	58.40	0.24
Chicago Cubs	83	78	51.55	707	672	52.54	-0.98
Chicago White Sox	79	83	48.77	724	732	49.45	-0.69
Cincinnati Reds	78	84	48.15	673	723	46.42	1.73
Cleveland Indians	65	97	40.12	773	865	44.40	-4.28
Colorado Rockies	92	70	56.79	804	715	55.84	0.95
Detroit Tigers	86	77	52.76	743	745	49.87	2.90
Florida Marlins	87	75	53.70	772	766	50.39	3.31
Houston Astros	74	88	45.68	643	770	41.08	4.59
Kansas City Royals	65	97	40.12	686	842	39.90	0.23
Los Angeles Angels	97	65	59.88	883	761	57.38	2.50
Los Angeles Dodgers	95	67	58.64	780	611	61.97	-3.33
Milwaukee Brewers	80	82	49.38	785	818	47.94	1.44
Minnesota Twins	87	76	53.37	817	765	53.28	0.09
New York Mets	70	92	43.21	671	757	44.00	-0.79
New York Yankees	103	59	63.58	915	753		
Oakland Athletics	75	87	46.30	759	761	49.87	-3.57
Philadelphia Phillies	93	69	57.41	820	709	57.22	0.19
Pittsburgh Pirates	62	99	38.51	636	768	40.68	-2.17
San Diego Padres	75	87	46.30	638	769		
San Francisco Giants	88	74	54.32	657	611	53.62	0.70
Seattle Mariners	85	77	52.47	640	692	46.10	6.37
St. Louis Cardinals	91	71	56.17	730	640	56.54	-0.37
Tampa Bay Rays	84	78	51.85	803	754	53.14	-1.29
Texas Rangers	87	75	53.70	784	740	52.88	0.82
Toronto Blue Jays	75	87	46.30	798	771	51.72	-5.42
Washington Nationals	59	103	36.42	710	874	39.76	-3.34

Source: $www.baseball - reference.com$

Table 1.1: Price and Mileage for Used Porsches

Price ($1,000's)	Mileage (thousands)
69.4	21.5
56.9	43.0
49.9	19.9
47.4	36.0
42.9	44.0
36.9	49.8
83.0	1.3
72.9	0.7
69.9	13.4
67.9	9.7
66.5	15.3
64.9	9.5
58.9	19.1
57.9	12.9
54.9	33.9
54.7	26.0
53.7	20.4
51.9	27.5
51.9	51.7
49.9	32.4
44.9	44.1
44.8	49.8
39.9	35.0
39.7	20.5
34.9	62.0
33.9	50.4
23.9	89.6
22.9	83.4
16.0	86.0
52.9	37.4

Source: Autotrader.com, Spring 2007.

Choosing a Simple Linear Model

Recall that data can be represented by a **model** plus an **error** term:

$$Data = Model + Error$$

When the data involve a quantitative response variable Y and we have a single quantitative predictor X the model becomes

$$\begin{aligned} Y &= f(X) + \epsilon \\ &= \mu_Y + \epsilon \end{aligned}$$

where $f(X)$ is a function that gives the mean value of Y, μ_Y, at any value of X and ϵ represents the error (deviation) from that mean[1].

We generally use graphs to help visualize the nature of the relationship between the response and potential predictor variables. Scatterplots are the major tool for helping us choose a model when both the response and predictor are quantitative variables. If the scatterplot shows a consistent linear trend, then we use in our model a mean which follows a straight line relationship with the predictor. This gives a **simple linear regression** model where the function, $f(X)$, is a linear function of X. If we let β_0 and β_1 represent the intercept and slope, respectively, of that line we have

$$\mu_Y = f(X) = \beta_0 + \beta_1 X$$

and

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Example 1.2: *Porsche prices (continued)*

CHOOSE

A scatterplot of price versus mileage for the sample of used Porsches is shown in Figure 1.1. The plot indicates a negative association between these two variables. It is generally understood that cars with lots of miles cost less on average than cars with only limited miles and the scatterplot supports this understanding. Since the rate of decrease in the scatterplot is relatively constant as the mileage increases, a linear model might provide a good summary of the relationship between the average prices and mileages of used Porsches for sale on this internet site. In symbols, we express the mean price as a linear function of mileage.

$$\mu_{Price} = \beta_0 + \beta_1 \cdot Mileage$$

[1] More formal notation for the mean value of Y at a given value of X is $\mu_{Y|X}$. To minimize distractions in most formulas we will use just μ_Y when the role of the predictor is clear.

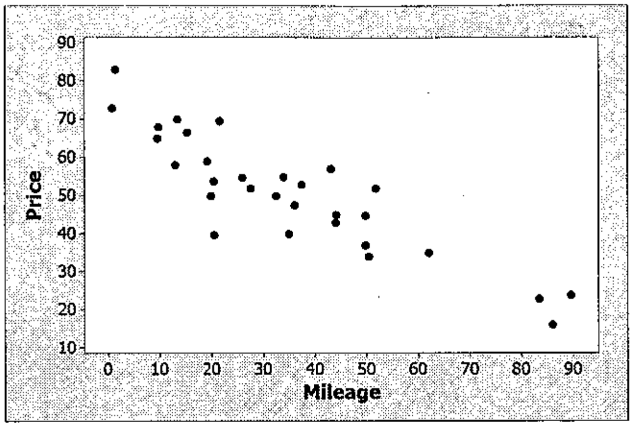


Figure 1.1: Scatterplot of Porsche price versus mileage

Thus the model for actual used Porsche prices would be

$$Price = \beta_0 + \beta_1 \cdot Mileage + \epsilon$$

This model indicates that Porsche prices should be scattered around a straight line with deviations from the line determined by the random error component, ϵ. We now turn to the question of how to choose the slope and intercept for the line that best summarizes this relationship. ◇

Fitting a Simple Linear Model

We want the best possible estimates of β_0 and β_1. Thus, we use least squares regression to fit the model to the data. This chooses coefficient estimates to minimize the sum of the squared errors and leads to the best set of predictions when we use our model to predict the data. In practice, we rely on computer technology to compute the least squares estimates for the parameters. The fitted model is represented by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

In general, we use Greek letters (β_0, β_1, etc.) to denote parameters and hats ($\hat{\beta}_0, \hat{\beta}_1$, etc.) are added to denoted estimated (fitted) values of these parameters.

A key tool for fitting a model is to compare the values it predicts for the individual data cases[2] to the actual values of the response variable in the dataset. The discrepancy in predicting each response is measured by the **residual**.

$$residual = observed\, y - predicted\, y = y - \hat{y}$$

[2] We generally use a lower case y when referring to the value of a variable for an individual case and an upper case Y for the variable itself.

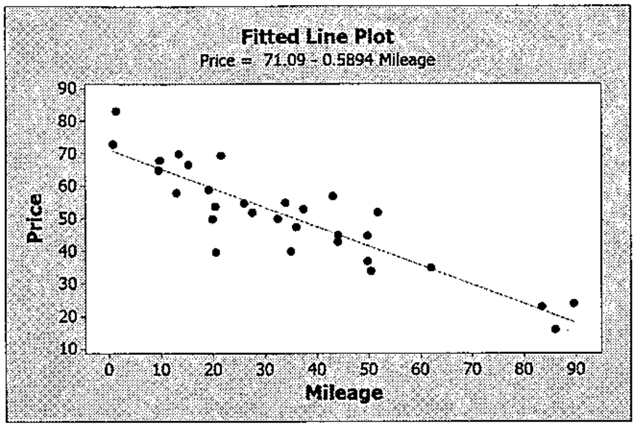


Figure 1.2: Linear regression to predict Porsche price based on mileage

The **sum of squared residuals** provides a measure of how well the line predicts the actual responses for a sample. We often denote this quantity as **SSE** for the sum of the squared errors. Statistical software calculates the fitted values of the slope and intercept so as to minimize this sum of squared residuals, hence we call this the **least squares line.**

Example 1.3: *Porsche prices (continued)*

FIT

For the i^{th} car in the data set, with mileage x_i, the model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The parameters, β_0 and β_1 in the model, represent the true, population-wide intercept and slope for all Porsches for sale. The corresponding statistics, $\hat{\beta}_0$ and $\hat{\beta}_1$, are estimates derived from this particular sample of 30 Porsches. (These estimates are determined from statistical software, for example in the Minitab fitted line plot shown in Figure 1.2 or the output shown in Figure 1.3).

The least squares line is

$$\widehat{Price} = 71.09 - 0.5894 \cdot Mileage$$

Thus, for every additional 1,000 miles on a used Porsche the predicted price goes down by about \$589. Also, if a (used!) Porche had zero miles on it, we would predict the price to be \$71,090. In many cases the the intercept lies far from the data used to fit the model and has no practical interpretation.

```
Regression Analysis: Price versus Mileage

The regression equation is Price = 71.1 - 0.589 Mileage

Predictor      Coef  SE Coef       T      P
Constant     71.090    2.370   30.00  0.000
Mileage    -0.58940  0.05665  -10.40  0.000


S = 7.17029   R-Sq = 79.5%   R-Sq(adj) = 78.7%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  5565.7  5565.7  108.25  0.000
Residual Error  28  1439.6    51.4
Total           29  7005.2
```

Figure 1.3: Minitab output for regression of Porsche *Price* on *Mileage*

Note that car #1 in Table 1.1 had a mileage level of 21.5 (21,500 miles) and a price of 69.4 ($69,400), whereas the fitted line predicts a price of

$$\widehat{Price} = 71.09 - 0.5894 \cdot Mileage = 58.4$$

The residual here is $Price - \widehat{Price} = 69.4 - 58.4 = 11.0$.

If we do a similar calculation for each of the 30 cars, square each of the resulting residuals, and sum the squares we get 1439.6. If you were to choose any other straight line to make predictions for these Porsche prices based on the mileages you could never obtain an SSE less than 1439.6. ◇

1.2 Conditions for a Simple Linear Model

We know that our model won't fit the data perfectly. The discrepancies that result from fitting the model represent what the model did not capture in each case. We want to check whether our model is reasonable and captures the main features of the dataset. Are we justified in using our model? Do the assumptions of the model appear to be reasonable? How much can we trust predictions that come from the model? Do we need to adjust or expand the model to better explain features of the data or could it be simplified without much loss of predictive power?

In specifying any model certain conditions must be satisfied for the model to make sense. We often make assumptions about the nature of the relationship between variables and the distribution of the errors. A key part of assessing any model is to check whether the conditions are reasonable for the data at hand. We hope that the residuals are small and contain no pattern that could be exploited to better explain the response variable. If our assessment shows a problem, then the model should be refined. Typically, we will rely heavily on graphs of residuals to assess the appropriateness of the model. In this section we discuss the conditions that are commonly placed on a simple linear model. The conditions we describe here for the simple linear regression model are typical of those that will be used throughout this book. In the following section we explore ways to use graphs to help us assess whether the conditions hold for a particular set of data.

Linearity – The overall relationship between the variables has a linear pattern. The average values of the response Y for each value of X fall on a common straight line.

The other conditions deal with the distribution of the errors.

Zero Mean – The error distribution is centered at zero. This means that the points are scattered at random above and below the line. (Note: By using least squares regression, we force the residual mean to be zero. Other techniques would not necessarily satisfy this condition.)

Constant Variance – The variability in the errors is the same for all values of the predictor variable. This means that the spread of points around the line remains fairly constant.

Independence – The errors are assumed to be independent from one another. Thus, one point falling above or below the line has no influence on the location of another point.

When we are interested in using the model to make formal inferences (conducting hypothesis tests or providing confidence intervals), additional assumptions are needed.

Random – The data are obtained using a random process. Most commonly this arises either from random sampling from a population of interest or from the use of randomization in a statistical experiment.

Normality – In order to use standard distributions for confidence intervals and hypothesis tests we often need to assume that the random errors follow a normal distribution.

We can summarize these conditions for a simple linear model using the following notation.

Simple Linear Regression Model

For a quantitative response variable Y and a single quantitative explanatory variable X the **simple linear regression model** is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ follows a normal distribution, that is $\epsilon \sim N(0, \sigma_\epsilon)$, and the errors are independent from one another.

Estimating the standard deviation of the error term

The simple linear regression model has three unknown parameters: the slope, β_1; the intercept, β_0; and the standard deviation σ_ϵ of the errors around the line. We have already seen that software will find the least squares estimates of the slope and intercept. Now we must consider how to estimate σ_ϵ, the standard deviation of the distribution of errors. Since the residuals estimate how much Y varies about the regression line, the sum of the squared residuals (SSE) is used to compute the estimate, $\widehat{\sigma_\epsilon}$. The value of $\widehat{\sigma_\epsilon}$ is referred to as the **regression standard error** and is interpreted as the size of a "typical" error.

Regression Standard Error

For a simple linear regression model, the estimated standard deviation of the error term based on the least squares fit to a sample of n observations is

$$\widehat{\sigma_\epsilon} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

The predicted values and resulting residuals are based on a sample slope and intercept that are calculated from the data. Therefore, we have $n-2$ **degrees of freedom** for estimating the regression standard error[3]. In general we lose an additional degree of freedom in the denominator for each new beta parameter that is estimated in the prediction equation.

Example 1.4: *Porsche prices (continued)*

The sum of squared residuals for the Porsche data is shown in Figure 1.3 as 1439.6 (see the SS column of the Residual Error line of the Analysis of Variance table in the Minitab output). Thus, the regression standard error is

$$\widehat{\sigma_\epsilon} = \sqrt{\frac{1439.6}{30-2}} = 7.17$$

Using mileage to predict price of a used Porsche, the typical error will be around \$7,170. So we have some feel for how far individual cases might spread above or below the regression line. Note that this value is labeled S in the Minitab output of Figure 1.3. ◇

[3] If a scatterplot only has 2 points, then it's easy to fit a straight line with residuals of zero, but we have no way of estimating the variability in the distribution of the error term. This corresponds to having zero degrees of freedom.

1.3 Assessing Conditions

A variety of plots are used to assess the conditions of the simple linear model. Scatterplots, histograms, and dotplots will be helpful to begin the assessment process. However, plots of residuals versus fitted values and normal plots will provide more detailed information, and these visual displays will be used throughout the text.

Residuals versus Fits Plots

A scatterplot with the fitted line provides one visual method of checking linearity. Points will be randomly scattered above and below the line when the linear model is appropriate. Clear patterns, for example clusters of points above and below the line in a systematic fashion, indicate that the linear model is not appropriate.

A more informative way of looking at how the points vary about the regression line is a scatterplot of the residuals versus the fitted values for the prediction equation. This plot reorients the axes so that the regression line is represented as a horizontal line through zero. Positive residuals represent points that are above the regression line. The residuals versus fits plot allows us to focus on the estimated errors and look for any clear patterns without the added complexity of a sloped line.

The residual versus fits plot is especially useful for assessing the linearity and constant variance conditions of a simple linear model. The ideal pattern will be random variation above and below zero in a band of relatively constant width. Figure 1.4 shows a typical residual versus fits plot when these two conditions are satisfied.

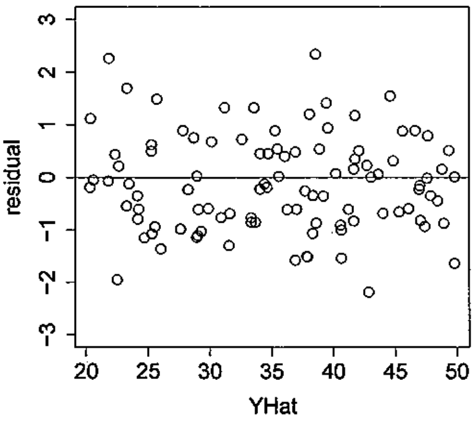
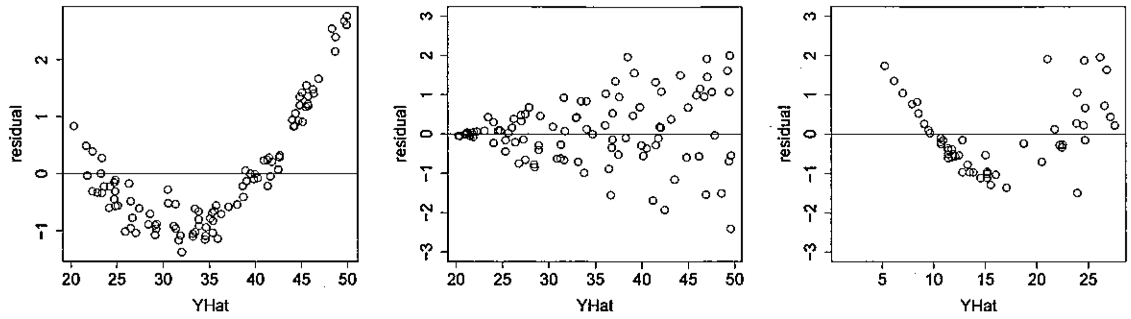


Figure 1.4: Residuals versus fitted values plot when linearity and constant variance conditions hold



(a) Nonlinear

(b) Nonconstant variance

(c) Both nonlinear & nonconstant variance

Figure 1.5: Residuals versus fitted values plots illustrating problems with conditions

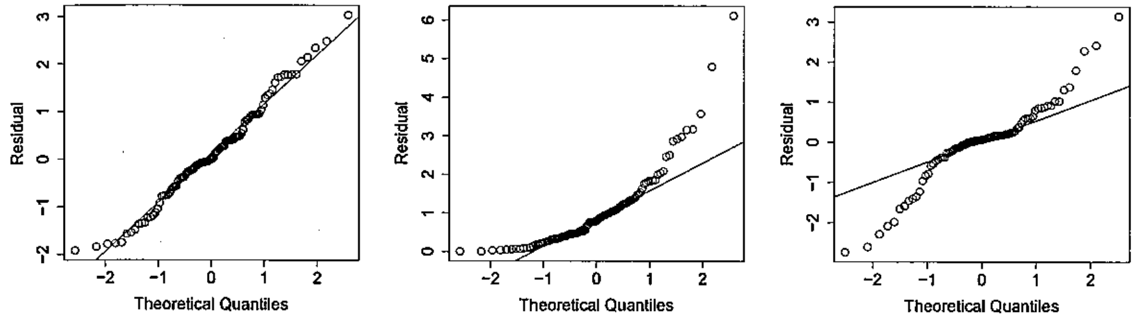
Figure 1.5 shows some examples of residual versus fits plots that exhibit some typical patterns indicating a problem with linearity, constant variance, or both conditions.

Figure 1.5(a) illustrates a curved pattern demonstrating a lack of linearity in the relationship. The residuals are mostly positive at either extreme of the graph and negative in the middle, indicating more of a curved relationship. Despite this pattern, the vertical width of the band of residuals is relatively constant across the graph, showing that the constant variance condition is probably reasonable for this model.

Figure 1.5(b) shows a common violation of the equal variance assumption. In many cases as the predicted response gets larger its variability also increases, producing a fan shape as in this plot. Note that a linearity assumption might still be valid in this case since the residuals are still equally dispersed above and below the zero line as we move across the graph.

Figure 1.5(c) indicates problems with both the linearity and constant variance conditions. We see a lack of linearity due to the curved pattern in the plot and, again, variance in the residuals that increases as the fitted values increase.

In practice, the assessment of a residual versus fits plot may not lead to as obvious a conclusion as in these examples. Remember that no model is "perfect" and we should not expect to always obtain the ideal plot. A certain amount of variation is natural, even for sample data that are generated from a model that meets all of the conditions. The goal is to recognize when departures from the model conditions are sufficiently evident in the data to suggest that an alternative model might be preferred or we should use some caution when the drawing conclusions from the model.



(a) Normal residuals (b) Skewed right residuals (c) Long-tailed residuals

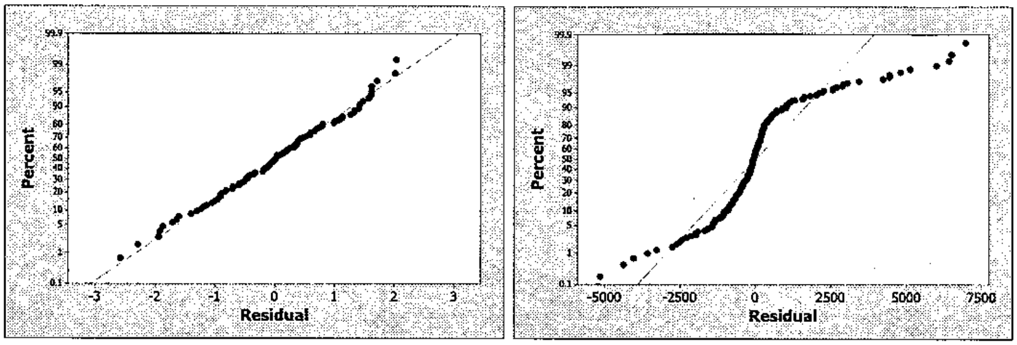
Figure 1.6: Examples of normal quantile plots

Normal plots

Data from a normal distribution should exhibit a "bell-shaped" curve when plotted as a histogram or dotplot. However, we often need a fairly large sample to see this shape accurately and even then it may be difficult to assess whether the symmetry and curvature of the tails are consistent with a true normal curve. As an alternative, a **normal plot** shows a different view of the data where an ideal pattern for a normal sample is a straight line. Although there are a number of a variations, there are generally two common methods for constructing a normal plot.

The first, called a **normal quantile plot**, is a scatterplot of the ordered observed data versus values (the theoretical quantiles) that we would expect to see from a "perfect" normal sample of the same size. If the ordered residuals are increasing at the rate we would expect to see for a normal sample, the resulting scatterplot is a straight line. If the distribution of the residuals is skewed in one direction or has tails that are overly long due to some extreme outliers at both ends of the distribution the normal quantile plot will bend away from a straight line. Figure 1.6 shows several examples of normal quantile plots. The first (Figure 1.6(a)) was generated from residuals where the data were generated from a linear model with normal errors and the other two from models with non-normal errors.

The second common method of producing a normal plot is to use a **normal probability plot** such as those shown in Figure 1.7. Here the ordered sample data are plotted on the horizontal axis while the vertical axis is transformed to reflect the rate that normal probabilities grow. As with a normal quantile plot, the values increase as we move from left to right across the graph, but the revised scale produces a straight line when the values increase at the rate we would expect for a sample from a normal distribution. Thus, the interpretation is the same. A linear pattern (as in Figure 1.7(a)) indicates good agreement with normality and curvature or bending away from a straight line (as in Figure 1.7(b)) shows departures from normality.



(a) Normal residuals

(b) Non-normal residuals

Figure 1.7: Examples of normal probability plots

Since both normal plot forms have a similar interpretation, we will use them interchangeably. The choice we make for a specific problem often depends on the options that are most readily available in the statistical software we are using.

Example 1.5: *Porsche prices (continued)*

ASSESS

We illustrate these ideas by checking the conditions for the model to predict Porsche prices based on mileage.

Linearity: Figure 1.2 shows that the linearity condition is reasonable as the scatterplot shows a consistent decline in prices with mileage and no obvious curvature. A plot of the residuals versus fitted values is shown in Figure 1.8. The horizontal band of points scattered randomly above and below the zero line illustrates that a linear model is appropriate for describing the relationship between price and mileage.

Zero mean: We used least squares regression which forces the sample mean of the residuals to be zero when estimating the intercept β_0. Also note that the residuals are scattered on either side of zero in the residual plot of Figure 1.8 and a histogram of the residuals, Figure 1.9, is centered at zero.

Constant variance: The fitted line plot in Figure 1.2 shows the data spread in roughly equal width bands on either side of the least squares line. Looking left to right in the plot of residuals versus fitted values in Figure 1.8 reinforces this finding as we see a fairly constant spread of the residuals above and below zero (where zero corresponds to actual prices that fall on the least squares regression line). This supports the constant variance condition.

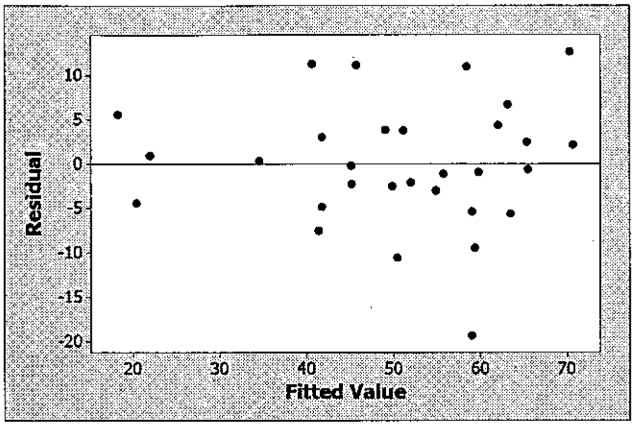


Figure 1.8: Plot of Porsche residuals versus fitted values

Independence and Random: We cannot tell from examining the data whether these conditions are satisfied. However, the context of the situation and the way the data were collected make these reasonable assumptions. There is no reason to think that one seller changing the asking price for a used car would necessarily influence the asking price of another seller. We were also told that these data were randomly selected from the Porsches for sale on the Autotrader.com website. So, at the least, we can treat it as a random sample from the population of all Porsches on that site at the particular time the sample was collected. We might want to be cautious about extending the findings to cars from a different site, an actual used car lot, or a later point in time.

Normality: In assessing normality we can refer to the histogram of residuals in Figure 1.9 where a reaqsonably bell-shaped pattern is displayed. However, a histogram based on this small sample may not be particularly informative and can change considerably depending on the bins used to

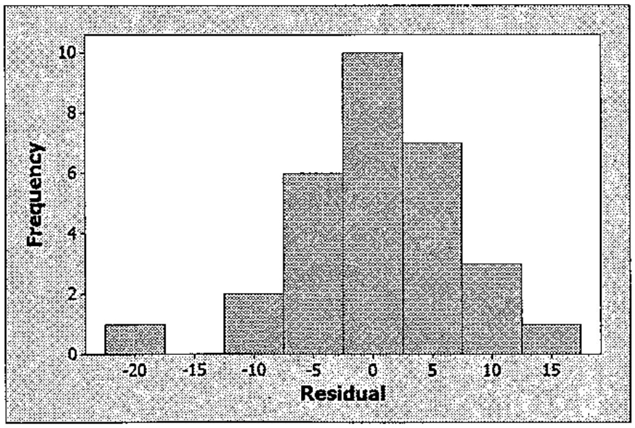


Figure 1.9: Histogram of Porsche residuals

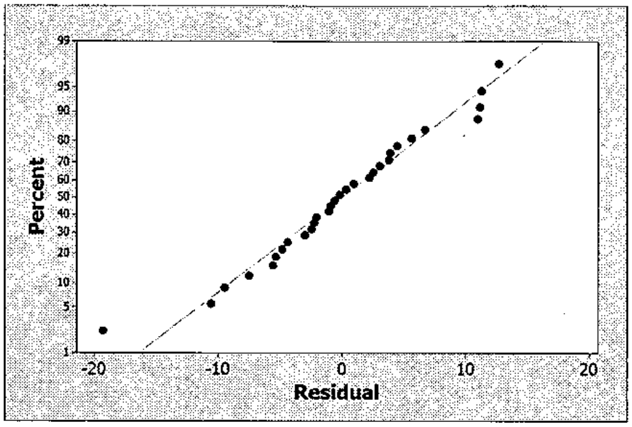


Figure 1.10: Normal Probability Plot of residuals for Porsche data

determine the bars. A more reliable plot for assessing normality is the normal probability plot of the residuals shown in Figure 1.10. This graph shows a consistent linear trend which supports the normality condition. We might have a small concern about the single point in the lower left corner of the plot, but we are looking more at the overall pattern when assessing normality.

USE

After we have decided on a reasonable model, we interpret the implications for the question of interest. For example, suppose we find a used Porsche for sale with 50 thousand miles and we believe that it is from the same population from which our sample of 30 used Porsches was drawn. What should we expect to pay for this car? Would it be an especially good deal if the owner was asking \$38,000?

Based on our model, we would expect to pay

$$\widehat{Price} = 71.09 - 0.5894 \cdot 50 = 41.62$$

or \$41,620. The asking price of \$38,000 is below the expected price of \$41,6200, but is this difference large relative to the variability in Porsche prices? We might like to know if this is a really good deal or perhaps such a low price that we should be concerned about the condition of the car. This question will be addressed in a Section 2.4 where we consider prediction intervals. For now, we can observe that the car's residual is about half of what we called a "typical error" ($\widehat{\sigma}_\epsilon = \$7.17$ thousand) below the expected price. Thus it is low, but not unusually so. ◇

1.4 Transformations

If one or more of the conditions for a simple linear regression model are not satisfied, then we can consider transformations on one or both of the variables. In this section we provide two examples where this is the case.

Example 1.6: *Doctors and hospitals in metropolitan areas*

We expect the number of doctors in a city to be related to the number of hospitals, reflecting both the size of the city and the general level of medical care. Finding the number of hospitals in a given city is relatively easy, but counting the number of doctors is a more challenging task. Fortunately, the U.S. Census Bureau regularly collects such data for many metropolitan areas in the United States. The data in Table 1.2 show values for these two variables (and the *City* names) from the first few cases in the data file **MetroHealth83**, which has a sample of 83 metropolitan areas that have at least two community hospitals.

CHOOSE

As usual, we start the process of finding a model to predict the number of MDs ($NumMDs$) from the number of hospitals ($NumHospitals$) by examining a scatterplot of the two variables as seen in Figure 1.11. As expected this shows an increasing trend with cities having more hospitals also tending to have more doctors, suggesting that a linear model might be appropriate.

Table 1.2: Number of MDs and Community Hospitals for sample of $n = 83$ Metropolitan Areas

City	NumMDs	NumHospitals
Holland-Grand Haven, MI	349	3
Louisville, KY-IN	4042	18
Battle Creek, MI	256	3
Madison, WI	2679	7
Fort Smith, AR-OK	502	8
Sarasota-Bradenton-Venice, FL	2352	7
Anderson, IN	200	2
Honolulu, HI	3478	13
Asheville, NC	1489	5
Winston-Salem, NC	2018	6
⋮	⋮	⋮

Source: U.S. Census Bureau: 2006 State and Metropolitan Area Data Book (Table B-6)