



Unit A: Linear Regression

Simple Linear Regression

Response: Quantitative

Predictor(s): Quantitative

Chapter 1: Inference for Simple Linear Regression

Identify and fit a linear model for a quantitative response based on a quantitative predictor. Check the conditions for a simple linear model and use transformations when they are not met. Detect outliers and influential points.

Chapter 2: Inference for Simple Linear Regression

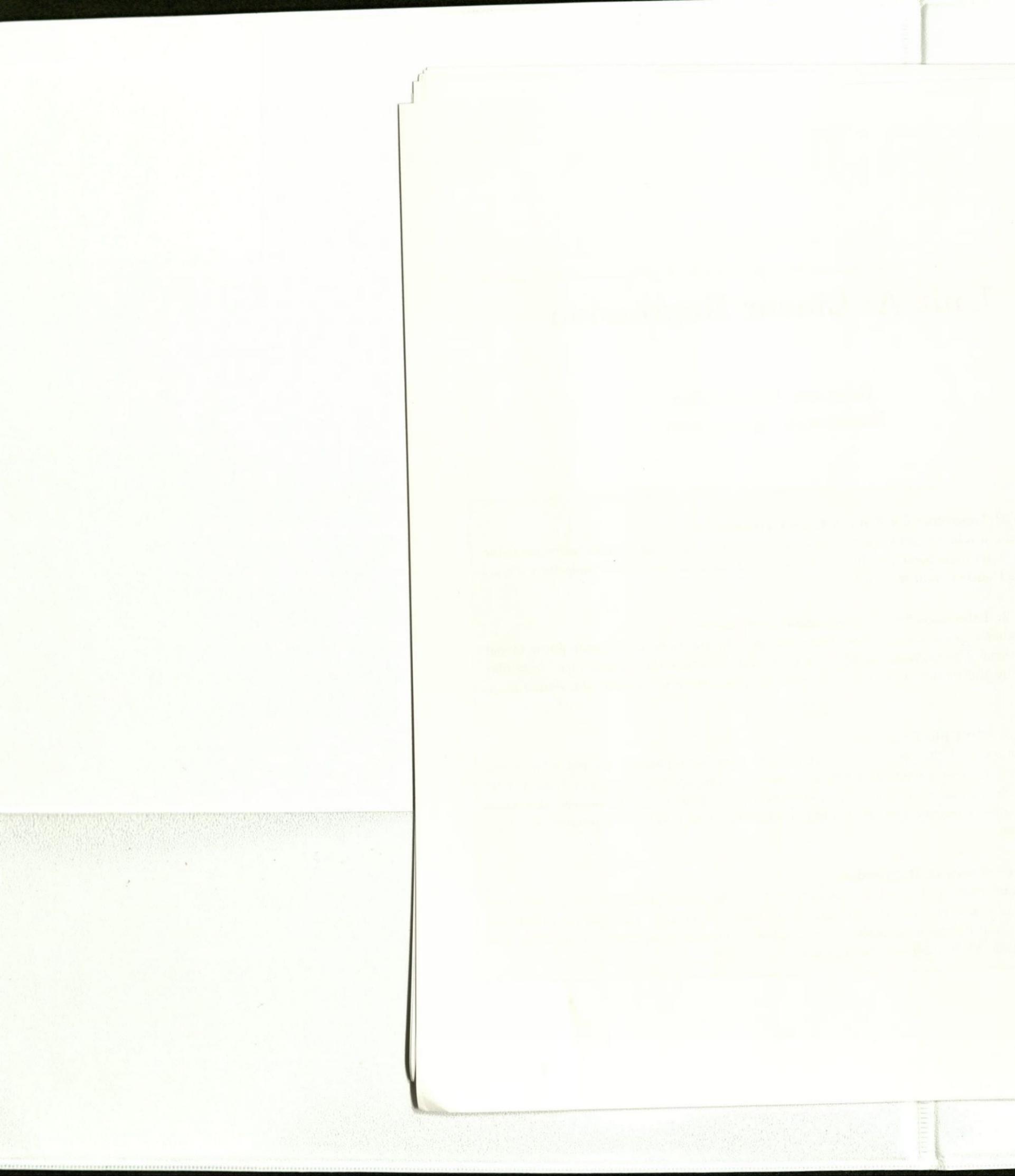
Test hypotheses and construct confidence intervals for the slope of a simple linear model. Partition variability to create an ANOVA table and determine the proportion of variability explained by the model. Construct intervals for predictions made with the simple linear model.

Chapter 3: Multiple Regression

Extend the ideas of the previous two chapters to consider regression models with two or more predictors. Use a multiple regression model to compare two regression lines. Create and assess models using functions of predictors, interactions, and polynomials. Recognize issues of multicollinearity with correlated predictors. Test a subset of predictors with a nested F-test.

Chapter 4: Topics in Regression

Construct and interpret an added variable plot. Consider techniques for choosing predictors to include in a model. Identify unusual and influential points. Incorporate categorical predictors using indicator variables. Use computer simulation techniques (bootstrap and randomization) to do inference for regression parameters.



CHAPTER 1

Simple Linear Regression

How is the price of a used car related to the number of miles it's been driven? Is the number of doctors in a city related to the number of hospitals? How can we predict the price of a textbook from the number of pages?

In this chapter, we consider a single quantitative predictor X for a quantitative response variable Y . A common model to summarize the relationship between two quantitative variables is the *simple linear regression model*. We assume that you have encountered simple linear regression as part of an introductory statistics course. Therefore, we review the structure of this model, the estimation and interpretation of its parameters, the assessment of its fit, and its use in predicting values for the response. Our goal is to introduce and illustrate many of the ideas and techniques of statistical model building that will be used throughout this book in a somewhat familiar setting. In addition to recognizing when a linear model may be appropriate, we also consider methods for dealing with relationships between two quantitative variables that are not linear.

1.1 The Simple Linear Regression Model

Example 1.1: *Porsche prices*

Suppose that we are interested in purchasing a Porsche sports car. If we can't afford the high sticker price of a new Porsche, we might be interested in finding a used one. How much should we expect to pay? Obviously, the price might depend on many factors including the age, condition, and special features on the car. For this example, we will focus on the relationship between $X = \text{Mileage}$ of a used Porsche and $Y = \text{Price}$. We used an Internet sales site¹ to collect data for a sample of 30 used Porsches, with price (in thousands of dollars) and mileage (in thousands of miles), as shown in Table 1.1. The data are also stored in the file named **PorschePrice**. We are interested in predicting the price of a used Porsche based on its mileage, so the explanatory variable is *Mileage*, the response is *Price*, and both variables are quantitative.

¹Source: Autotrader.com, Spring 2007.

Price (\$1000s)	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

Table 1.1: Price and Mileage for used Porsches

Choosing a Simple Linear Model

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

$$\text{Data} = \text{Model} + \text{Error}$$

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.²

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 that 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:

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

²More 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.

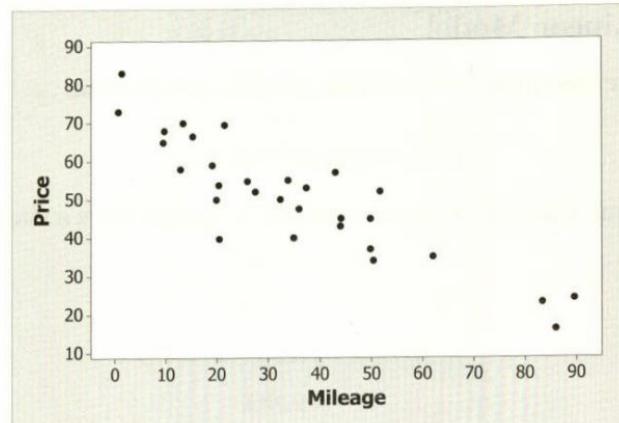


Figure 1.1: Scatterplot of Porsche Price versus Mileage

Thus, the model for actual used Porsche prices would be

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{Mileage} + \epsilon$$

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 denote estimated (fitted) values of these parameters.

A key tool for fitting a model is to compare the values it predicts for the individual data cases³ to the actual values of the response variable in the dataset. The discrepancy in predicting each response is measured by the **residual**:

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

³We generally use a lowercase y when referring to the value of a variable for an individual case and an uppercase Y for the variable itself.

model. A key part of this process is to determine whether the model fits the data at hand. We begin by looking at some diagnostic plots for the fitted regression model. In this section, we focus on the simple linear regression model. The conditions required for this model will be made clear in the next section.

Linearity: The first diagnostic plot is a scatterplot of the response variable versus the predictor variable.

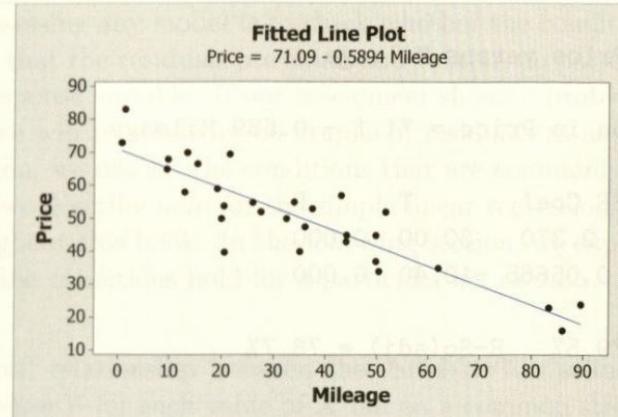


Figure 1.2: Linear regression to predict Porsche Price based on Mileage

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 dataset, 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

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

Thus, for every additional 1000 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

$$S = 7.17029 \quad R-Sq = 79.5\% \quad R-Sq(\text{adj}) = 78.7\%$$

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: Computer 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

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

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

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.⁴ 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 the price of a used Porsche, the typical error will be around \$7170. 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. ◇

⁴If 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.

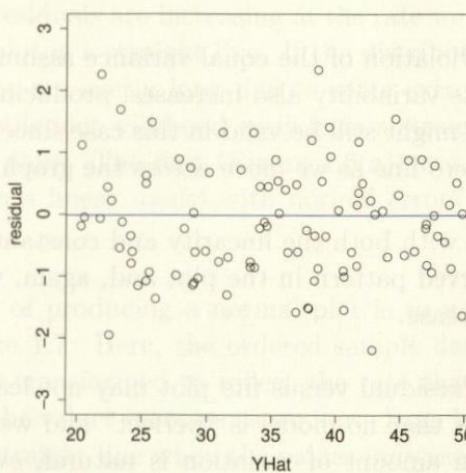


Figure 1.4: Residuals versus fitted values plot when linearity and constant variance conditions hold

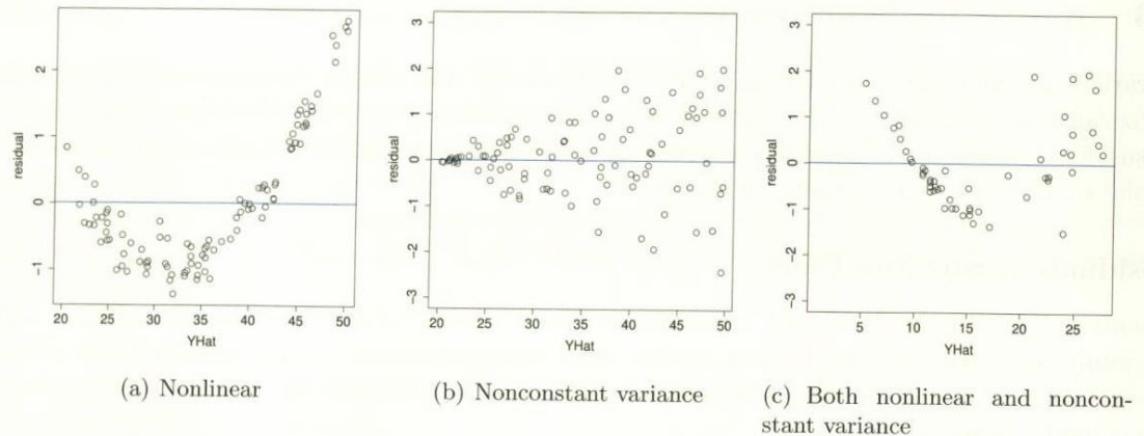


Figure 1.5: Residuals versus fitted values plots illustrating problems with conditions

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 drawing conclusions from the model.

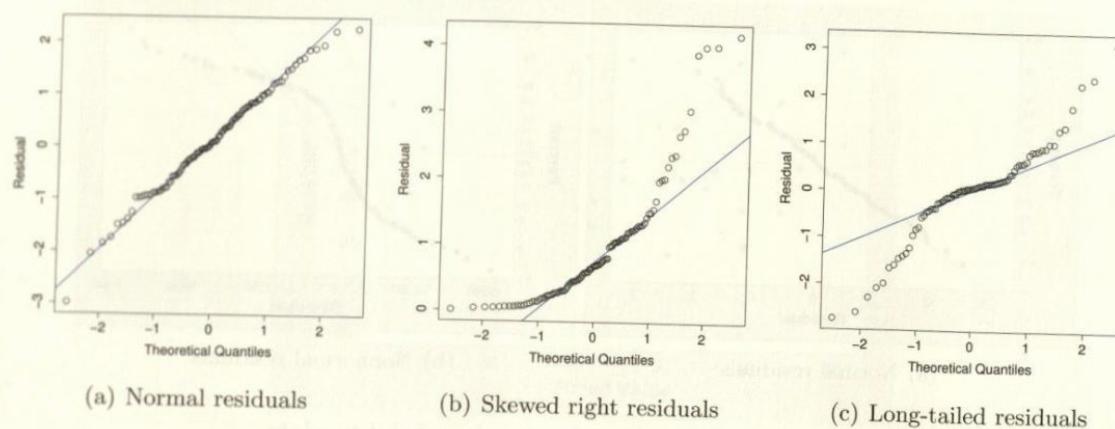


Figure 1.6: Examples of normal quantile plots

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 a number of variations exist, 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 nonnormal 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.

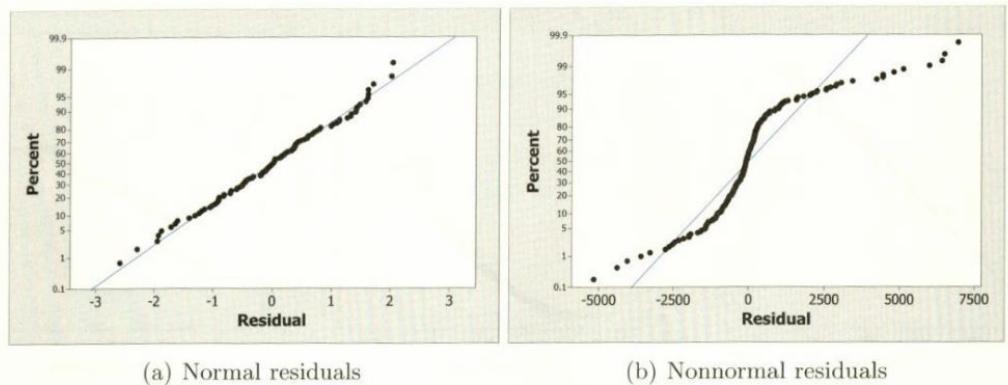


Figure 1.7: Examples of normal probability plots

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.

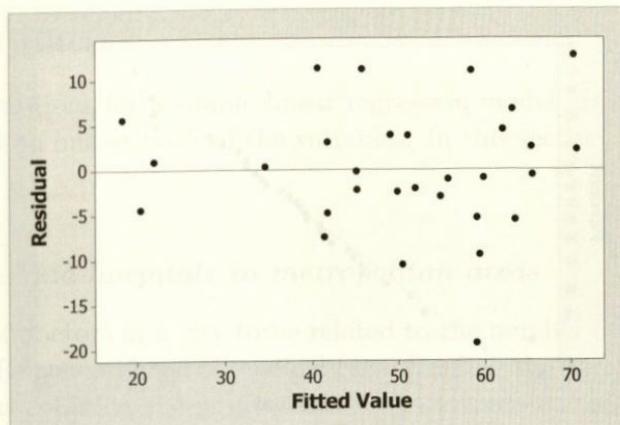


Figure 1.8: Plot of Porsche residuals versus fitted values

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 reasonably bell-shaped pattern is displayed. However, a histogram based on this small sample

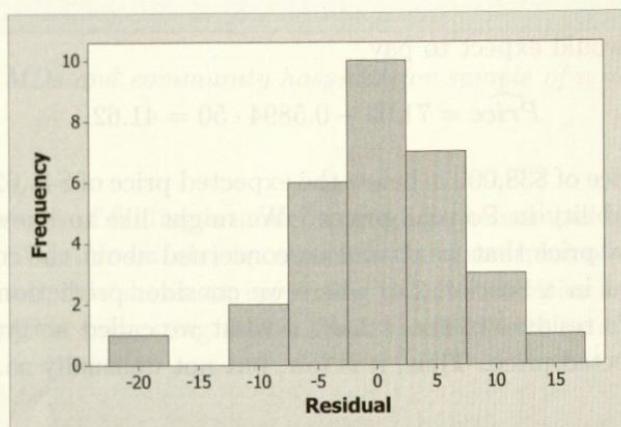


Figure 1.9: Histogram of Porsche residuals

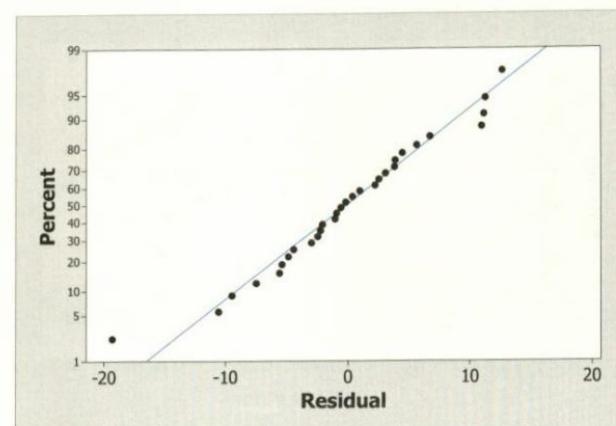


Figure 1.10: *Normal probability plot of residuals for Porsche data*

may not be particularly informative and can change considerably depending on the bins used to 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 that 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 50,000 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

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

or \$41,620. The asking price of \$38,000 is below the expected price of \$41,620, 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 areas⁵ that have at least two community hospitals.

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
:	:	:

Table 1.2: Number of MDs and community hospitals for sample of $n = 83$ metropolitan areas

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.

⁵Source: U.S. Census Bureau, 2006 State and Metropolitan Area Data Book, Table B-6.

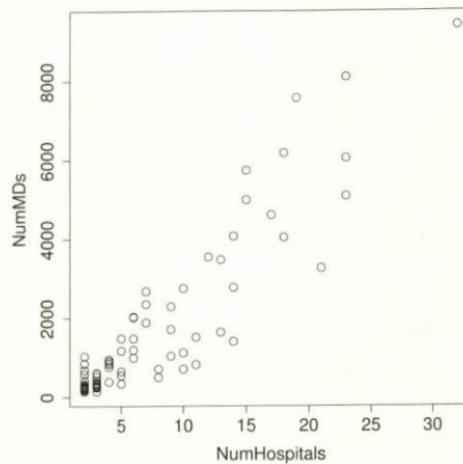


Figure 1.11: Scatterplot for Doctors versus Hospitals

FIT

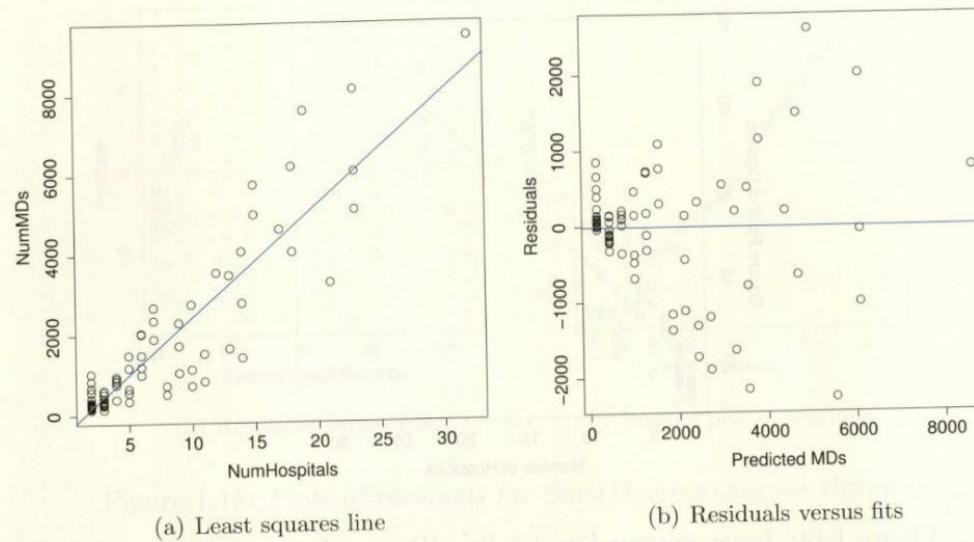
Fitting a least squares line in R produces estimates for the slope and intercept as shown below, giving the prediction equation $\widehat{\text{NumMDs}} = -385.1 + 282.0 \cdot \text{NumHospitals}$. Figure 1.12(a) shows the scatterplot with this regression line as a summary of the relationship.

```
Call:  
lm(formula = NumMDs ~ NumHospitals)
```

```
Coefficients:  
(Intercept)  NumHospitals  
          -385.1           282.0
```

ASSESS

The line does a fairly good job of following the increasing trend in the relationship between number of doctors and number of hospitals. However, a closer look at plots of the residuals shows some considerable departures from our standard regression assumptions. For example, the plot of residuals versus fitted values in Figure 1.12(b) shows a fan shape, with the variability in the residuals tending to increase as the fitted values get larger. This often occurs with count data like the number of MDs and number of hospitals where variability increases as the counts grow larger. We can also observe this effect in a scatterplot with the regression line, Figure 1.12(a).

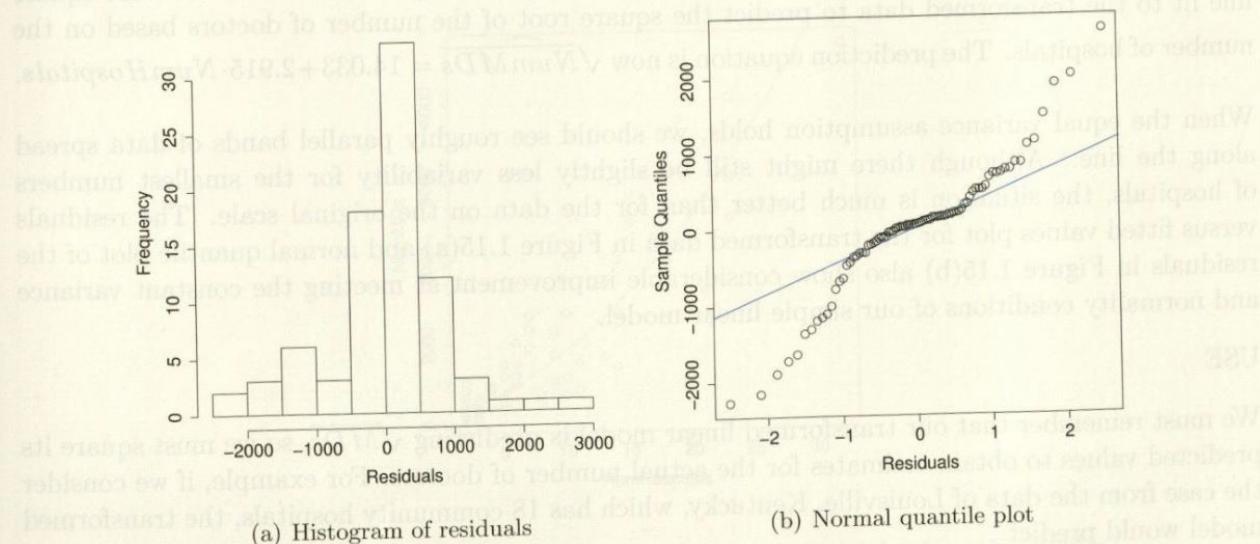


(a) Least squares line

(b) Residuals versus fits

Figure 1.12: Regression for number of doctors based on number of hospitals

We also see from a histogram of the residuals, Figure 1.13(a), and normal quantile plot, Figure 1.13(b), that an assumption of normality would not be reasonable for the residuals in this model. Although the histogram is relatively unimodal and symmetric, the peak is quite narrow with very long “tails.” This departure from normality is seen more clearly in the normal quantile plot that has significant curvature away from a straight line at both ends.



(a) Histogram of residuals

(b) Normal quantile plot

Figure 1.13: Normality plots for residuals of Doctors versus Hospitals

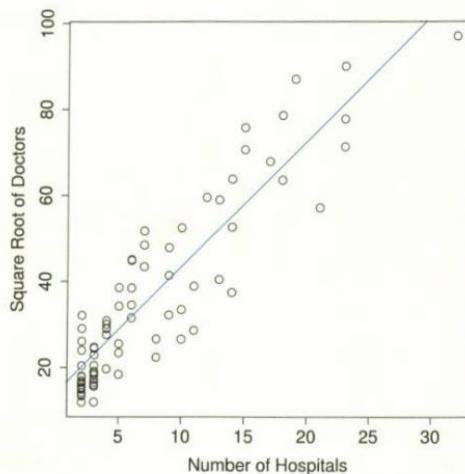


Figure 1.14: Least squares line for $\text{Sqrt}(\text{Doctors})$ versus Hospitals

CHOOSE (again)

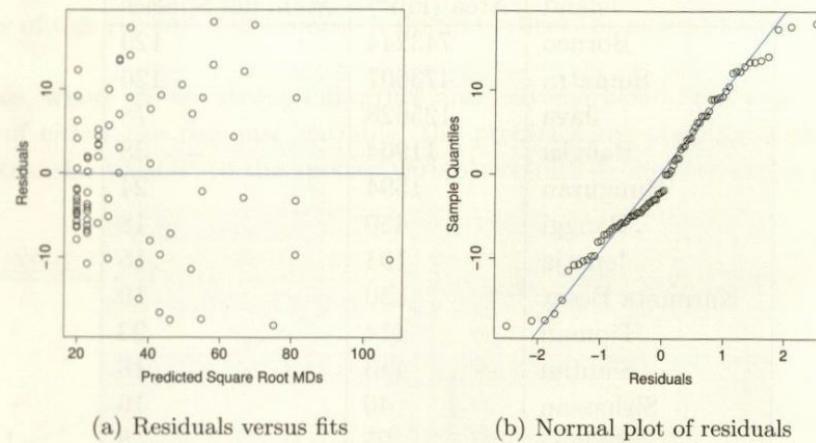
To stabilize the variance in a response (Y) across different values of the predictor (X) we often try transformations on either Y or X . Typical options include raising a variable to a power (such as \sqrt{Y} , X^2 , or $1/X$) or taking a logarithm (e.g., using $\log(Y)$ as the response). For count data, such as the number of doctors or hospitals where the variability increases along with the magnitudes of the variables, a square root transformation is often helpful. Figure 1.14 shows the least square line fit to the transformed data to predict the square root of the number of doctors based on the number of hospitals. The prediction equation is now $\widehat{\text{NumMDs}} = 14.033 + 2.915 \cdot \text{NumHospitals}$.

When the equal variance assumption holds, we should see roughly parallel bands of data spread along the line. Although there might still be slightly less variability for the smallest numbers of hospitals, the situation is much better than for the data on the original scale. The residuals versus fitted values plot for the transformed data in Figure 1.15(a) and normal quantile plot of the residuals in Figure 1.15(b) also show considerable improvement at meeting the constant variance and normality conditions of our simple linear model.

USE

We must remember that our transformed linear model is predicting $\widehat{\text{MDs}}$, so we must square its predicted values to obtain estimates for the actual number of doctors. For example, if we consider the case from the data of Louisville, Kentucky, which has 18 community hospitals, the transformed model would predict

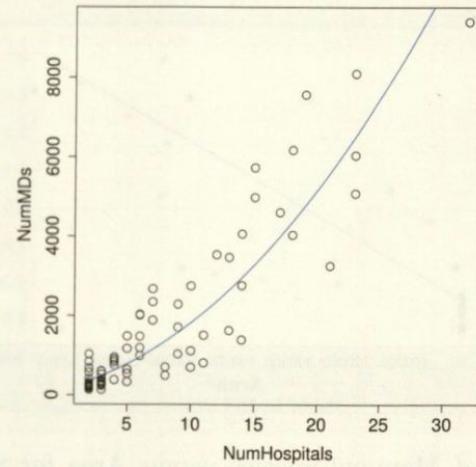
$$\widehat{\text{NumMDs}} = 14.033 + 2.915 \cdot 18 = 66.50$$

Figure 1.15: *Plots of residuals for Sqrt(Doctors) versus Hospitals*

so the predicted number of doctors is $66.50^2 = 4422.3$, while Louisville actually had 4042 doctors at the time of this sample. Figure 1.16 shows the scatterplot with the predicted number of doctors after transforming the linear model for the square roots of the number of doctors back to the original scale so that

$$\widehat{\text{NumMDs}} = (14.033 + 2.915 \cdot \text{NumHospitals})^2$$

Note that we could use this model to make predictions for other cities, but in doing so, we should feel comfortable only to the extent that we believe the sample to be representative of the larger population of cities with at least two community hospitals. ◇

Figure 1.16: *Predicted NumMDs from the linear model for Sqrt(NumMDs)*

Island	Area (km ²)	Mammal Species
Borneo	743244	129
Sumatra	473607	126
Java	125628	78
Bangka	11964	38
Bunguran	1594	24
Banggi	450	18
Jemaja	194	15
Karimata Besar	130	19
Tioman	114	23
Siantan	113	16
Sirhassan	46	16
Redang	25	8
Penebangan	13	13
Perhentian Besar	8	6

Table 1.3: Species and Area for Southeast Asian islands

Example 1.7: Species by area

The data in Table 1.3 (and the file **SpeciesArea**) show the number of mammal species versus area for 13 islands⁶ in Southeast Asia. Biologists have speculated that the number of species is related to the size of an island and would like to be able to predict the number of species based on the size of an island.

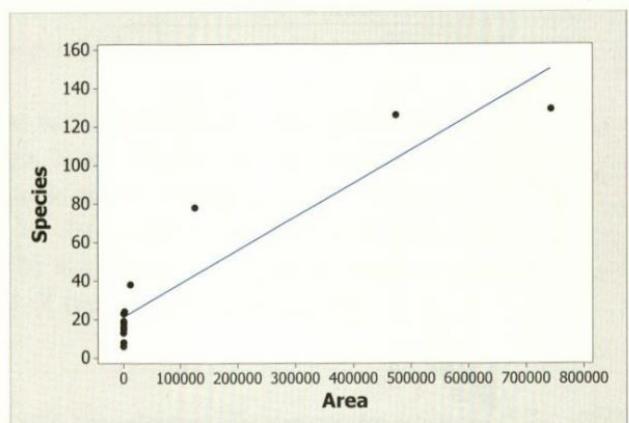


Figure 1.17: Number of Mammal Species versus Area for S.E. Asian islands

⁶Source: Lawrence R. Heaney, (1984), Mammalian Species Richness on Islands on the Sunda Shelf, Southeast Asia, *Oecologia* vol. 61, no. 1, pages 11–17.

Figure 1.17 shows a scatterplot with least squares line added. Clearly, the line does not provide a good summary of this relationship because it doesn't reflect the curved pattern shown in the plot.

In a case like this, where we see strong curvature and extreme values in a scatterplot, a logarithm transformation of either the response variable, the predictor, or possibly both, is often helpful. Applying a log transformation⁷ to the species variable results in the scatterplot of Figure 1.18(a).

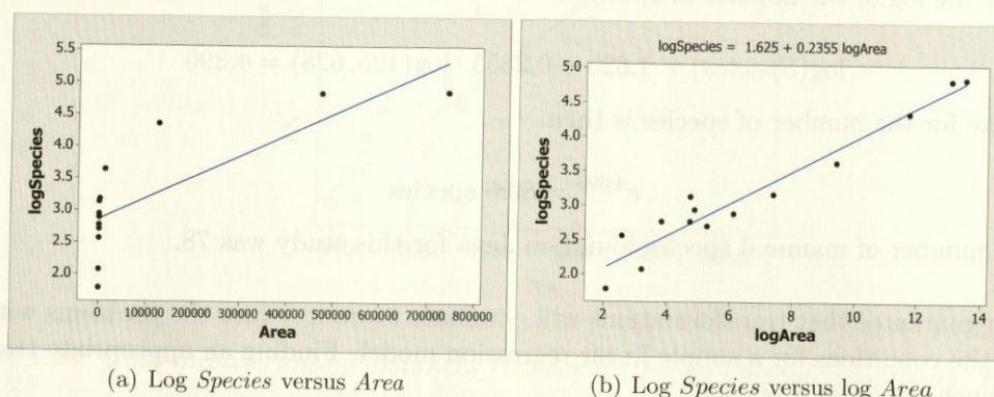


Figure 1.18: *Log transformations of Species versus Area for S.E. Asian islands*

Clearly, this transformation has failed to produce a linear relationship. However, if we also take a log transformation of the area, we obtain the plot illustrated in Figure 1.18(b), which does show a linear pattern. Figure 1.19 shows a residual plot from this regression, which does not indicate any striking patterns.

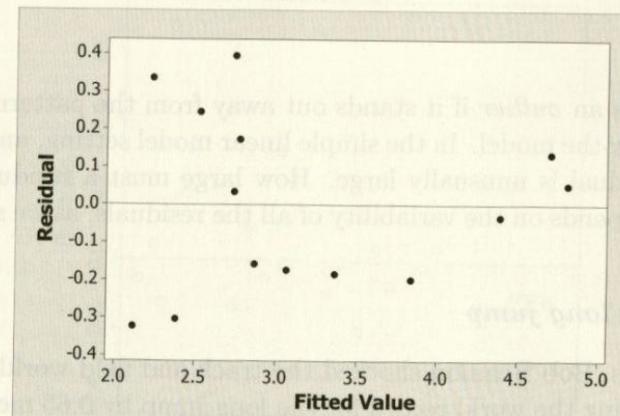


Figure 1.19: *Residual plot after log transform of response and predictor*

⁷In this text we use log to denote the natural logarithm.

Based on the fitted model, we can predict $\log(Species)$ based on $\log(Area)$ for an island with the model

$$\widehat{\log(Species)} = 1.625 + 0.2355 \cdot \log(Area)$$

Suppose we wanted to use the model to find the predicted value for Java, which has an area of 125,628 square kilometers. We substitute 125,628 into the equation as the *Area* and compute an estimate for the log of the number of species:

$$\widehat{\log(Species)} = 1.625 + 0.2355 \cdot \log(125,628) = 4.390$$

Our estimate for the number of species is then

$$e^{4.390} = 80.6 \text{ species}$$

The actual number of mammal species found on Java for this study was 78. ◇

There is no guarantee that transformations will eliminate or even reduce the problems with departures from the conditions for a simple linear regression model. Finding an appropriate transformation is as much an art as a science.

1.5 Outliers and Influential Points

Sometimes, a data point just doesn't fit within a linear trend that is evident in the other points. This can be because the point doesn't fit with the other points in a scatterplot vertically—it is an *outlier*—or a point may differ from the others horizontally and vertically so that it is an *influential* point. In this section, we examine methods for identifying outliers and influential points using graphs and summary statistics.

Outliers

We classify a data point as an *outlier* if it stands out away from the pattern of the rest of the data and is not well described by the model. In the simple linear model setting, an outlier is a point where the magnitude of the residual is unusually large. How large must a residual be for a point to be called an outlier? That depends on the variability of all the residuals, as we see in the next example.

Example 1.8: *Olympic long jump*

During the 1968 Olympics, Bob Beamon shocked the track and field world by jumping 8.9 meters (29 feet 2.5 inches), breaking the world record for the long jump by 0.65 meters (more than 2 feet). Figure 1.20 shows the winning men's Olympic long jump distance (labeled as *Gold*) versus *Year*, together with the least squares regression line, for the $n = 26$ Olympics held during the period 1900–2008. The data are stored in **LongJumpOlympics**.

1.5. OUTLIERS AND INFLUENTIAL POINTS

47

or studentized residuals, to be discussed in the next section. The most important observation for us here is that the 1968 point is an outlier. In fact, it is the only point that is more than about two standard deviations from the regression line. The 1968 point is an outlier because the predicted value for the long jump data point is far above the predicted value for the men's high jump in 1977, while none of the other studentized residuals are even close to zero. A residual indicates how well the model fits the data, so a large residual indicates that the model does not fit the data well.

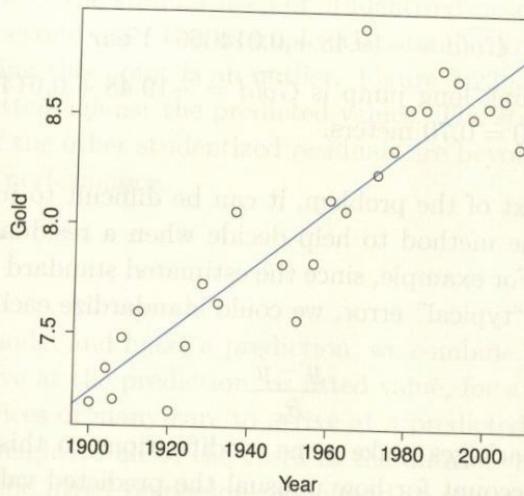


Figure 1.20: *Gold-medal-winning distances (m) for the men's Olympic long jump, 1900–2008*

The 1968 point clearly stands above the others and is far removed from the regression line. Because this point does not fit the general pattern in the scatterplot, it is an outlier. The unusual nature of this point is perhaps even more evident in Figure 1.21, a residual plot for the fitted least squares model.

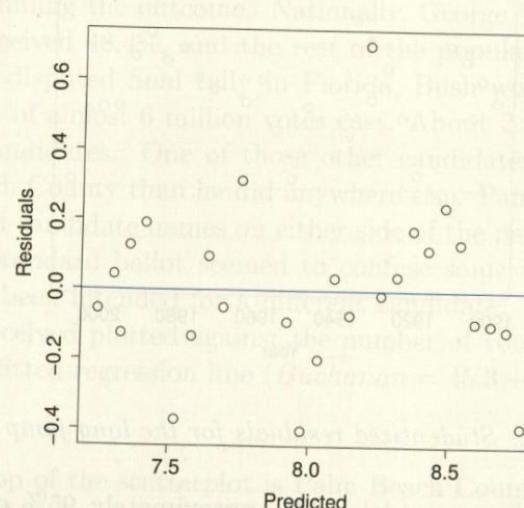


Figure 1.21: *Residual plot for long jump model*

The fitted regression model is

$$\widehat{Gold} = -19.48 + 0.014066 \cdot Year$$

Thus, the predicted 1968 winning long jump is $\widehat{Gold} = -19.48 + 0.014066 \cdot 1968 = 8.20$ meters. The 1968 residual is $8.90 - 8.20 = 0.70$ meters.

Even when we know the context of the problem, it can be difficult to judge whether a residual of 0.70 m is unusually large. One method to help decide when a residual is extreme is to put the residuals on a standard scale. For example, since the estimated standard deviation of the regression error, $\widehat{\sigma}_\epsilon$, reflects the size of a “typical” error, we could standardize each residual using

$$\frac{y - \widehat{y}}{\widehat{\sigma}_\epsilon}$$

In practice, most statistical packages make some modifications to this formula when computing a **standardized** residual to account for how unusual the predicted value is for a particular case. Since an extreme outlier might have a significant effect on the estimation of σ_ϵ , another common adjustment is to estimate the standard deviation of the regression error using a model that is fit after omitting the point in question. Such residuals are often called **studentized**⁸ (or **deleted-t**) residuals.

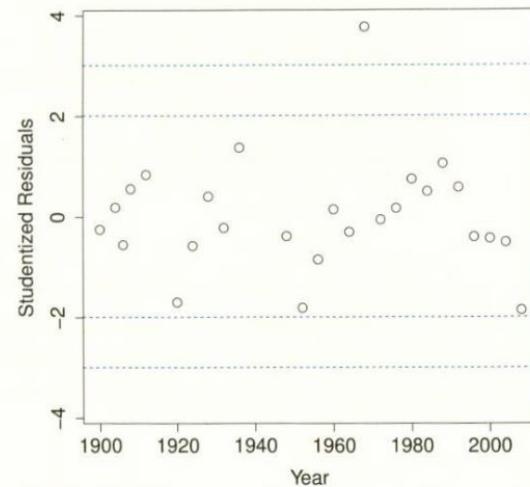


Figure 1.22: Studentized residuals for the long jump model

If the conditions of a simple linear model hold, approximately 95% of the residuals should be within 2 standard deviations of the residual mean of zero, so we would expect most standardized

⁸You may recall the t-distribution, which is sometimes called Student's t, from an introductory statistics course.

or studentized residuals to be less than 2 in absolute value. We may be slightly suspicious about points where the magnitude of the standardized or studentized residual is greater than 2 and even more wary about points beyond ± 3 . For example, the standardized residual for Bob Beamon's 1968 jump is 3.03, indicating this point is an outlier. Figure 1.22 shows the studentized residuals for the long jump data plotted against the predicted values. The studentized residual for the 1968 jump is 3.77, while none of the other studentized residuals are beyond ± 2 , clearly pointing out the exceptional nature of that performance. \diamond

Influential Points

When we fit a regression model and make a prediction, we combine information across several observations, or cases, to arrive at the prediction, or fitted value, for a particular case. For example, we use the mileages and prices of many cars to arrive at a predicted price for a particular car. In doing this, we give equal weight to all of the cases in the dataset; that is, every case contributes equally to the creation of the fitted regression model and to the subsequent predictions that are based on that model.

Usually, this is a sensible and useful thing to do. Sometimes, however, this approach can be problematic, especially when the data contain one or more extreme cases that might have a significant impact on the coefficient estimates in the model.

Example 1.9: *Butterfly ballot*

The race for the presidency of the United States in the fall of 2000 was very close, with the electoral votes from Florida determining the outcome. Nationally, George W. Bush received 47.9% of the popular vote, Al Gore received 48.4%, and the rest of the popular vote was split among several other candidates. In the disputed final tally in Florida, Bush won by just 537 votes over Gore (48.847% to 48.838%) out of almost 6 million votes cast. About 2.3% of the votes cast in Florida were awarded to other candidates. One of those other candidates was Pat Buchanan, who did much better in Palm Beach County than he did anywhere else. Palm Beach County used a unique "butterfly ballot" that had candidate names on either side of the page with "chads" to be punched in the middle. This nonstandard ballot seemed to confuse some voters, who punched votes for Buchanan that may have been intended for a different candidate. Figure 1.23 shows the number of votes that Buchanan received plotted against the number of votes that Bush received for each county, together with the fitted regression line ($Buchanan = 45.3 + 0.0049 * Bush$). The data are stored in **PalmBeach**.

The data point near the top of the scatterplot is Palm Beach County, where Buchanan picked up over 3000 votes. Figure 1.24 is a plot of the residuals versus fitted values for this model; clearly, Palm Beach County stands out from the rest of the data. Using this model, Minitab computes the standardized residual for Palm Beach to be 7.65 and the studentized residual to be 24.08! No

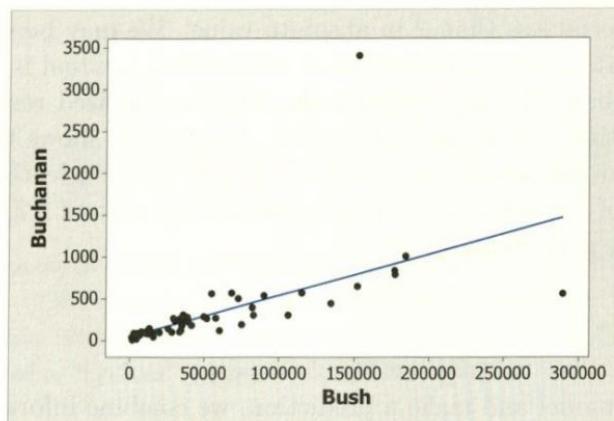


Figure 1.23: 2000 presidential election totals in Florida counties

question that this point should be considered an outlier. Also, the data point at the far right on the plots (Dade County) has a large negative residual of -907.5 , which gives a standardized residual of -3.06 ; certainly something to consider as an outlier, although not as dramatic as Palm Beach.

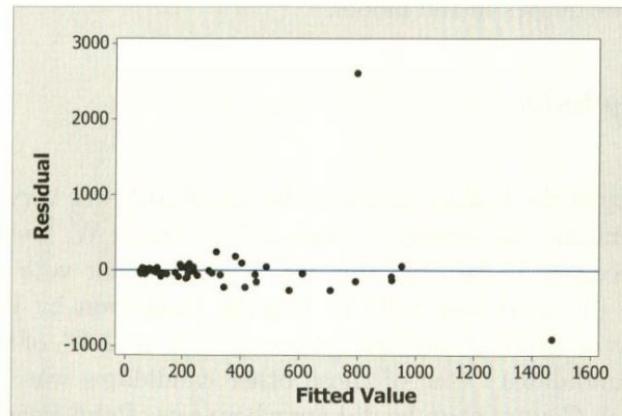


Figure 1.24: Residual plot for the butterfly ballot data

Other than recognizing that the model does a poor job of predicting Palm Beach County (and to a lesser extent Dade County), should we worry about the effect that such extreme values have on the rest of the predictions given by the model? Would removing Palm Beach County from the dataset produce much change in the regression equation? Portions of the Minitab output for fitting the simple linear model with and without the Palm Beach County data point are shown below. Figure 1.25 shows both regression lines, with the steeper slope (0.0049) occurring when Palm Beach County is included and the shallower slope (0.0035) when that point is omitted. Notice that the effect of the extreme value for Palm Beach is to “pull” the regression line in its direction.

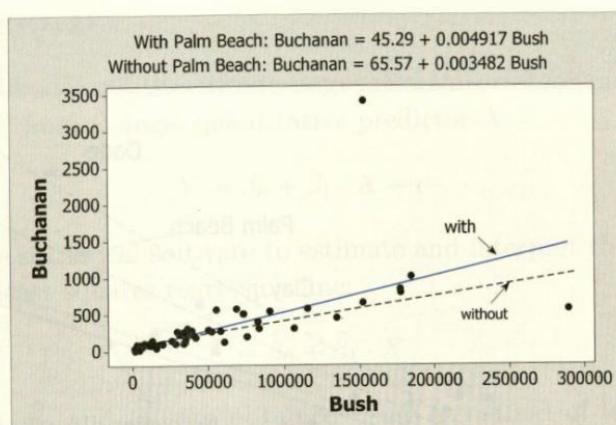


Figure 1.25: Regression lines with and without Palm Beach

Regression output with Palm Beach County:

The regression equation is

$$\text{Buchanan} = 45.3 + 0.00492 \text{ Bush}$$

Predictor	Coef	SE Coef	T	P
Constant	45.29	54.48	0.83	0.409
Bush	0.0049168	0.0007644	6.43	0.000

$$S = 353.922 \quad R-\text{Sq} = 38.9\% \quad R-\text{Sq}(\text{adj}) = 38.0\%$$

Regression output without Palm Beach County:

The regression equation is

$$\text{Buchanan} = 65.6 + 0.00348 \text{ Bush}$$

Predictor	Coef	SE Coef	T	P
Constant	65.57	17.33	3.78	0.000
Bush	0.0034819	0.0002501	13.92	0.000

$$S = 112.453 \quad R-\text{Sq} = 75.2\% \quad R-\text{Sq}(\text{adj}) = 74.8\%$$

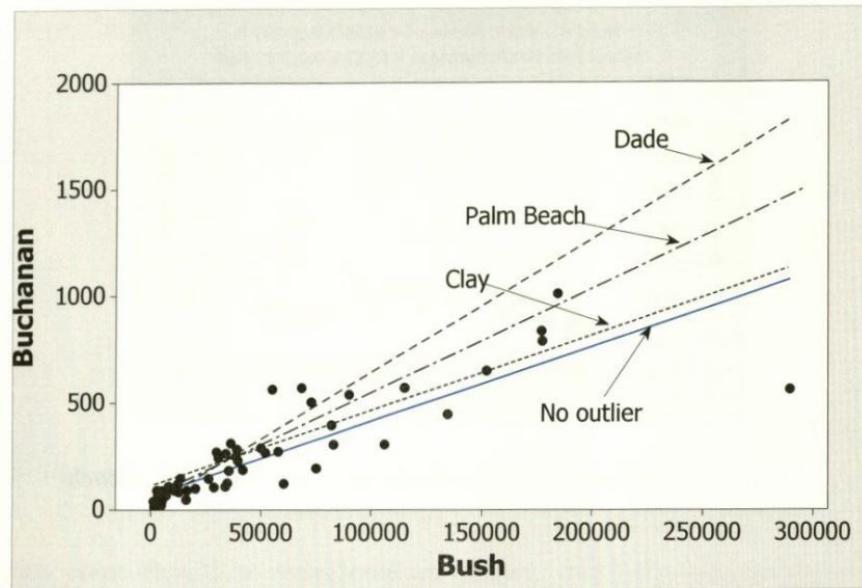


Figure 1.26: Regression lines with an outlier of 3407 “moved” to different counties

The amount of **influence** that a single point has on a regression fit depends on how well it aligns with the pattern of the rest of the points and on its value for the predictor variable. Figure 1.26 shows the regression lines we would have obtained if the extreme value (3407 Buchanan votes) had occurred in Dade County (with 289,456 Bush votes), Palm Beach County (152,846 Bush votes), Clay County (41,745 Bush votes), or not occurred at all. Note that the more extreme values for the predictor (large Bush counts in Dade or Palm Beach) produced a bigger effect on the slope of the regression line than when the outlier was placed in a more “average” Bush county such as Clay.

Generally, points farther from the mean value of the predictor (\bar{x}) have greater potential to influence the slope of a fitted regression line. This concept is known as the **leverage** of a point. Points with high leverage have a greater capacity to pull the regression line in their direction than do low leverage points near the predictor mean. Although in the case of a single predictor, we could measure leverage as just the distance from the mean, we introduce a somewhat more complicated statistic in Section 4.3 that can be applied to more complicated regression settings.

1.6. CHAPTER SUMMARY

1.6 Chapter Summary

In this chapter, we considered a **simple linear regression model** for predicting a single quantitative response variable Y from a single quantitative predictor X :

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

You should be able to use statistical software to estimate and interpret the slope and intercept for this model to produce a least squares regression line:

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

The coefficient estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are obtained using a method of least squares that selects them to provide the smallest possible sum of squared residuals (SSE). A **residual** is the observed response (y) minus the predicted response (\hat{y}), or the vertical distance from a point to the line. You should be able to interpret, in context, the intercept and slope, as well as residuals.

The **scatterplot**, which shows the relationship between two quantitative variables, is an important visual tool to help choose a model. We look for the direction of association (positive, negative, or a random scatter), the strength of the relationship, and the degree of linearity. Assessing the fit of the model is a very important part of the modeling process. The conditions to check when using a simple linear regression model include **linearity**, **zero mean** (of the residuals), **constant variance** (about the regression line), **independence**, **random selection** (or random assignment), and **normality** (of the residuals). We can summarize several of these conditions by specifying the distribution of the error term, $\epsilon \sim N(0, \sigma_\epsilon)$. In addition to a scatterplot with the least squares line, various residual plots, such as a **histogram of the residuals** or a **residuals versus fits plot**, are extremely helpful in checking these conditions. Once we are satisfied with the fit of the model, we use the estimated model to make inferences or predictions.

Special plots, known as a **normal quantile plot** or **normal probability plot**, are useful in assessing the normality condition. These two plots are constructed using slightly different methods, but the interpretation is the same. A linear trend indicates that normality is reasonable, and departures from linearity indicate trouble with this condition. Be careful not to confuse linearity in normal plots with the condition of a linear relationship between the predictor and response variable.

You should be able to estimate the **standard deviation of the error** term, σ_ϵ , the third parameter in the simple linear regression model. The estimate is based on the sum of squared errors (SSE) and the associated degrees of freedom ($n - 2$):

$$\widehat{\sigma_\epsilon} = \sqrt{\frac{SSE}{n - 2}}$$

and is the typical error, often referred to as the **regression standard error**.

If the conditions are not satisfied, then **transformations** on the predictor, response, or both variables should be considered. Typical transformations include the square root, reciprocal, logarithm, and raising the variable(s) to another power. Identifying a useful transformation for a particular dataset (if one exists at all) is as much art as science. Trial and error is often a good approach.

You should be able to identify obvious **outliers** and **influential points**. Outliers are points that are unusually far away from the overall pattern shown by the other data. Influential points exert considerable impact on the estimated regression line. We will look at more detailed methods for identifying outliers and influential points in Section 4.3. For now, you should only worry about recognizing very extreme cases and be aware that they can affect the fitted model and analysis. One common guideline is to tag all observations with **standardized** or **studentized residuals** smaller than -2 or larger than 2 as possible outliers. To see if a point is influential, fit the model with and without that point to see if the coefficients change very much. In general, points far from the average value of the predictor variable have greater potential to influence the regression line.

1.7 Exercises

Conceptual Exercises

1.1 Equation of a line. Consider the fitted regression equation $\hat{Y} = 100 + 15 \cdot X$. Which of the following is *false*?

- a. The sample slope is 15.
- b. The predicted value of Y when $X = 0$ is 100.
- c. The predicted value of Y when $X = 2$ is 110.
- d. Larger values of X are associated with larger values of Y .

1.2–1.5 Sparrows. Priscilla Erickson from Kenyon College collected data on a stratified random sample of 116 Savannah sparrows at Kent Island. The weight (in grams) and wing length (in mm) were obtained for birds from nests that were reduced, controlled, or enlarged. The data⁹ are in the file **Sparrows**. Use the computer output below in Exercises 1.2–1.5.

The regression equation is Weight = 1.37 + 0.467 WingLength

Predictor	Coef	SE Coef	T	P
Constant	1.3655	0.9573	1.43	0.156
WingLength	0.4674	0.03472	13.46	0.000

S = 1.39959 R-Sq = 61.4% R-Sq(adj) = 61.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	355.05	355.05	181.25	0.000
Residual Error	114	223.31	1.96		
Total	115	578.36			

1.2 Sparrows slope. Based on the regression output, what is the slope of the least squares regression line for predicting sparrow weight from wing length?

1.3 Sparrows intercept. Based on the regression output, what is the intercept of the least squares regression line for predicting sparrow weight from wing length?

1.4 Sparrows regression standard error. Based on the regression output, what is the size of the typical error when predicting weight from wing length?

⁹We thank Priscilla Erickson and Professor Robert Mauck from the Department of Biology at Kenyon College for allowing us to use these data.

1.5 Sparrow degrees of freedom. What are the degrees of freedom associated with the standard regression error when predicting weight from wing length for these sparrows?

1.6 Computing a residual. Consider the fitted regression equation $\hat{Y} = 25 + 7 \cdot X$. If $x_1 = 10$ and $y_1 = 100$, what is the residual for the first data point?

1.7 Residual plots to check conditions. For which of the following conditions for inference in regression does a residual plot *not* aid in assessing whether the condition is satisfied?

- a. Linearity
- b. Constant variance
- c. Independence
- d. Zero mean

Guided Exercises

1.8 Breakfast cereal. The number of calories and number of grams of sugar per serving were measured for 36 breakfast cereals. The data are in the file **Cereal**. We are interested in trying to predict the calories using the sugar content.

- a. Make a scatterplot and comment on what you see.
- b. Find the least squares regression line for predicting calories based on sugar content.
- c. Interpret the value (not just the sign) of the slope of the fitted model in the context of this setting.

1.9 More breakfast cereal. Refer to the data on breakfast cereals in **Cereal** that is described in Exercise 1.8. The number of calories and number of grams of sugar per serving were measured for 36 breakfast cereals. The data are in the file **Cereal**. We are interested in trying to predict the calories using the sugar content.

- a. How many calories would the fitted model predict for a cereal that has 10 grams of sugar?
- b. Cheerios has 110 calories but just 1 gram of sugar. Find the residual for this data point.
- c. Does the linear regression model appear to be a good summary of the relationship between calories and sugar content of breakfast cereals?

1.10 Sparrow residuals. Exercises 1.2–1.5 deal with a model for the weight (in grams) of sparrows using the wing length as a predictor and the data in **Sparrows**. Construct and interpret the following plots for the residuals of this model. In each case, discuss what the plot tells you about potential problems (if any) with the regression conditions.

- Histogram of the residuals.
- Normal probability plot of the residuals.
- Scatterplot that includes the least squares line. Are there any obvious outliers or influential points in this plot?

1.11 Capacitor voltage. A capacitor was charged with a 9-volt battery and then a voltmeter recorded the voltage as the capacitor was discharged. Measurements were taken every 0.02 seconds. The data are in the file **Volts**.

- Make a scatterplot with *Voltage* on the vertical axis versus *Time* on the horizontal axis. Comment on the pattern.
- Transform *Voltage* using a log transformation and then plot $\log(Voltage)$ versus *Time*. Comment on the pattern.
- Regress $\log(Voltage)$ on *Time* and write down the prediction equation.
- Make a plot of residuals versus fitted values from the regression from part (c). Comment on the pattern.

1.12–1.15 Caterpillars. Student and faculty researchers at Kenyon College conducted numerous experiments with *Manduca Sexta* caterpillars to study biological growth.¹⁰ A subset of the measurements from some of the experiments is in the file **Caterpillars**. Exercises 1.12–1.15 deal with some relationships in these data. The variables in the dataset include:

Variable	Description
<i>Instar</i>	A number from 1 (smallest) to 5 (largest) indicating stage of the caterpillar's life
<i>ActiveFeeding</i>	An indicator (Y or N) of whether or not the animal is actively feeding
<i>Fgp</i>	An indicator (Y or N) of whether or not the animal is in a free growth period
<i>Mgp</i>	An indicator (Y or N) of whether or not the animal is in a maximum growth period
<i>Mass</i>	Body mass of the animal in grams
<i>Intake</i>	Food intake in grams/day
<i>WetFrass</i>	Amount of frass (solid waste) produced by the animal in grams/day
<i>DryFrass</i>	Amount of frass, after drying, produced by the animal in grams/day
<i>Cassim</i>	CO_2 assimilation (ingestion–excretion)
<i>Nfrass</i>	Nitrogen in frass
<i>Nassim</i>	Nitrogen assimilation (ingestion–excretion)

Log (base 10) transformations are also provided as *LogMass*, *LogIntake*, *LogWetFrass*, *LogDryFrass*, *LogCassim*, *LogNfrass*, and *LogNassim*.

¹⁰We thank Professors Harry Itagaki, Drew Kerkhoff, Chris Gillen, Judy Holdener, and their students for sharing this data from research supported by NSF InSTaRs grant No. 0827208.

1.12 Caterpillar waste versus mass. We might expect that the amount of waste a caterpillar produces per day (*WetFrass*) is related to its size (*Mass*). Use the data in **Caterpillars** to examine this relationship as outlined below.

- a. Produce a scatterplot for predicting *WetFrass* based on *Mass*. Comment on any patterns.
- b. Produce a similar plot using the log (base 10) transformed variables, *LogWetFrass* versus *LogMass*. Again, comment on any patterns.
- c. Would you prefer the plot in part (a) or part (b) to predict the amount of wet frass produced for caterpillars? Fit a linear regression model for the plot you chose and write down the prediction equation.
- d. Add a plotting symbol for the grouping variable *Instar* to the scatterplot that you chose in (c). Does the linear trend appear consistent for all five stages of a caterpillar's life? (Note: We are not ready to fit more complicated models yet, but we will return to this experiment in Chapter 3.)
- e. Repeat part (d) using plotting symbols (or colors) for the groups defined by the free growth period variable *Fgp*. Does the linear trend appear to be better when the caterpillars are in a free growth period? (Again, we are not ready to fit more complicated models, but we are looking at the plot for linear trend in the two groups.)

1.13 Caterpillar nitrogen assimilation versus mass. The *Nassim* variable in the **Caterpillars** dataset measures nitrogen assimilation, which might be associated with the size of the caterpillars as measured with *Mass*. Use the data to examine this relationship as outlined below.

- a. Produce a scatterplot for predicting nitrogen assimilation (*Nassim*) based on *Mass*. Comment on any patterns.
- b. Produce a similar plot using the log (base 10) transformed variables, *LogNassim* versus *LogMass*. Again, comment on any patterns.
- c. Would you prefer the plot in part (a) or part (b) to predict the nitrogen assimilation of caterpillars with a linear model? Fit a linear regression model for the plot you chose and write down the prediction equation.
- d. Add a plotting symbol for the grouping variable *Instar* to the scatterplot that you chose in (c). Does the linear trend appear consistent for all five stages of a caterpillar's life? (Note: We are not ready to fit more complicated models yet, but we will return to this experiment in Chapter 3.)
- e. Repeat part (d) using plotting symbols (or colors) for the groups defined by the free growth period variable *Fgp*. Does the linear trend appear to be better when the caterpillars are in a free growth period? (Again, we are not ready to fit more complicated models, but we are looking at the plot for linear trend in the two groups.)

1.14 Caterpillar body mass and food intake. We might expect that larger caterpillars would consume more food. Use the data in **Caterpillars** to look at using food intake to predict *Mass* as outlined below.

- a. Plot body mass (*Mass*) as the response variable versus food intake (*Intake*) as the explanatory variable. Comment on the pattern.
- b. Plot the log (base 10) transformed variables, *LogMass* versus *LogIntake*. Again, comment on any patterns.
- c. Do you think the linear model should be used to model either of the relationships in part (a) or (b)? Explain.

1.15 Caterpillar body mass and food intake—again. This exercise is similar to Exercise 1.14 except that we will reverse the roles of predictor and response, using caterpillar size (*Mass*) to predict food intake (*Intake*) for the data in **Caterpillars**.

- a. Plot *Intake* as the response variable versus *Mass* as the explanatory variable. Comment on the pattern.
- b. Plot the log (base 10) transformed variables, *LogIntake* versus *LogMass*. Again, comment on any patterns.
- c. Would you prefer the plot in part (a) or (b) to fit with a linear model? Fit a linear regression model for the plot you chose and write down the prediction equation.
- d. Add plotting symbols (or colors) for the grouping variable *Instar* to the scatterplot that you chose in (c). Is a linear model more appropriate for this relationship during some of the stages of caterpillar development?

1.16 U.S. stamp prices. Historical prices¹¹ of U.S. stamps for mailing a letter weighing less than 1 ounce are provided in the file **USStamps**.

- a. Plot *Price* (in cents) versus *Year* and comment on any patterns.
- b. Regular increases in the postal rates started in 1958. Remove the first four observations from the dataset and fit a regression line for predicting *Price* from *Year*. What is the equation of the regression line?
- c. Plot the regression line along with the prices from 1958 to 2012. Does the regression line appear to provide a good fit?
- d. Analyze appropriate residual plots for the linear model relating stamp price and year. Are the conditions for the regression model met?

¹¹The data were obtained from Wikipedia, the URL is http://en.wikipedia.org/wiki/History_of_United_States_postage_rates.

- e. Identify any unusual residuals.

1.17 Enrollment in mathematics courses. Total enrollments in mathematics courses at a small liberal arts college¹² were obtained for each semester from Fall 2001 to Spring 2012. The academic year at this school consists of two semesters, with enrollment counts for *Fall* and *Spring* each year as shown in Table 1.4. The variable *AYear* indicates the year at the beginning of the academic year. The data are also provided in the file **MathEnrollment**.

AYear	Fall	Spring
2001	259	246
2002	301	206
2003	343	288
2004	307	215
2005	286	230
2006	273	247
2007	248	308
2008	292	271
2009	250	285
2010	278	286
2011	303	254

Table 1.4: *Math enrollments*

- Plot the mathematics enrollment for each semester against time. Is the trend over time roughly the same for both semesters? Explain.
- A faculty member in the Mathematics Department believes that the fall enrollment provides a very good predictor of the spring enrollment. Do you agree?
- After examining a scatterplot with the least squares regression line for predicting spring enrollment from fall enrollment, two faculty members begin a debate about the possible influence of a particular point. Identify the point that the faculty members are concerned about.
- Fit the least squares line for predicting math enrollment in the spring from math enrollment in the fall, with and without the point you identified in part (c). Would you tag this point as influential? Explain.

¹²The data were obtained from <http://Registrar.Kenyon.edu> on June 1, 2012.

1.18 Pines. The dataset **Pines** contains data from an experiment conducted by the Department of Biology at Kenyon College at a site near the campus in Gambier, Ohio.¹³ In April 1990, student and faculty volunteers planted 1000 white pine (*Pinus strobes*) seedlings at the Brown Family Environmental Center. These seedlings were planted in two grids, distinguished by 10- and 15-foot spacings between the seedlings. Several variables, described below, were measured and recorded for each seedling over time.

Variable	Description
<i>Row</i>	Row number in pine plantation
<i>Col</i>	Column number in pine plantation
<i>Hgt90</i>	Tree height at time of planting (cm)
<i>Hgt96</i>	Tree height in September 1996 (cm)
<i>Diam96</i>	Tree trunk diameter in September 1996 (cm)
<i>Grow96</i>	Leader growth during 1996 (cm)
<i>Hgt97</i>	Tree height in September 1997 (cm)
<i>Diam97</i>	Tree trunk diameter in September 1997 (cm)
<i>Spread97</i>	Widest lateral spread in September 1997 (cm)
<i>Needles97</i>	Needle length in September 1997 (mm)
<i>Deer95</i>	Type of deer damage in September 1995: 0 = none, 1 = browsed
<i>Deer97</i>	Type of deer damage in September 1997: 0 = none, 1 = browsed
<i>Cover95</i>	Thorny cover in September 1995: 0 = none; 1 = some; 2 = moderate; 3 = lots
<i>Fert</i>	Indicator for fertilizer: 0 = no, 1 = yes
<i>Spacing</i>	Distance (in feet) between trees (10 or 15)

- Construct a scatterplot to examine the relationship between the initial height in 1990 and the height in 1996. Comment on any relationship seen.
- Fit a least squares line for predicting the height in 1996 from the initial height in 1990.
- Are you satisfied with the fit of this simple linear model? Explain.

1.19 Pines: 1997 versus 1990. Refer to the **Pines** data described in Exercise 1.18. Examine the relationship between the initial seedling height and the height of the tree in 1997.

- Construct a scatterplot to examine the relationship between the initial height in 1990 and the height in 1997. Comment on any relationship seen.
- Fit a least squares line for predicting the height in 1997 from the initial height in 1990.
- Are you satisfied with the fit of this simple linear model? Explain.

1.20 Pines: 1997 versus 1996. Refer to the **Pines** data described in Exercise 1.18. Consider fitting a line for predicting height in 1997 from height in 1996.

¹³Thanks to the Kenyon College Department of Biology for sharing these data.

- a. Before doing any calculations, do you think that the height in 1996 will be a better predictor than the initial seedling height in 1990? Explain.
- b. Fit a least squares line for predicting height in 1997 from height in 1996.
- c. Does this simple linear regression model provide a good fit? Explain.

1.21 Caterpillar CO_2 assimilation and food intake. Refer to the data in **Caterpillars** that is described on page 57 for Exercises 1.12–1.15. Consider a linear model to predict CO_2 assimilation (*Cassim*) using food intake (*Intake*) for the caterpillars.

- a. Plot *Cassim* versus *Intake* and comment on the pattern.
- b. Find the least squares regression line for predicting CO_2 assimilation from food intake.
- c. Are the conditions for inference met? Comment on the appropriate residual plots.

1.22 Caterpillar nitrogen assimilation and wet frass. Repeat the analysis described in Exercise 1.21 for a model to predict nitrogen assimilation (*Nassim*) based on the amount of solid waste (*WetFrass*) in the **Caterpillars** data.

1.23 Fluorescence experiment. Suzanne Rohrback used a novel approach in a series of experiments to examine calcium-binding proteins. The data from one experiment¹⁴ are provided in **Fluorescence**. The variable *Calcium* is the log of the free calcium concentration and *ProteinProp* is the proportion of protein bound to calcium.

- a. Find the regression line for predicting the proportion of protein bound to calcium from the transformed free calcium concentration.
- b. What is the regression standard error?
- c. Plot the regression line and all of the points on a scatterplot. Does the regression line appear to provide a good fit?
- d. Analyze appropriate residual plots. Are the conditions for the regression model met?

1.24 Goldenrod galls. Biology students collected measurements on goldenrod galls at the Brown Family Environmental Center.¹⁵ The file **Goldenrod** contains the gall diameter (in mm), stem diameter (in mm), wall thickness (in mm), and codes for the fate of the gall in 2003 and 2004.

- a. Are stem diameter and gall diameter positively associated in 2003?

¹⁴Thanks to Suzanne Rohrback for providing these data from her honors experiments at Kenyon College.

¹⁵Thanks to the Kenyon College Department of Biology for sharing these data.

- b. Plot wall thickness against stem diameter and gall diameter on two separate scatterplots for the 2003 data. Based on the scatterplots, which variable has a stronger linear association with wall thickness? Explain.
- c. Fit a least squares regression line for predicting wall thickness from the variable with the strongest linear relationship in part (b).
- d. Find the fitted value and residual for the first observation using the fitted model in (c).
- e. What is the value of a typical residual for predicting wall thickness based on your linear model in part (c)?

1.25 More goldenrod galls. Refer to the data on goldenrod galls described in Exercise 1.24. Repeat the analysis in that exercise for the measurements made in 2004 instead of 2003. The value of *Wall04* is missing for the first observation, so use the second case for part (e).

Open-ended Exercises

1.26 Textbook prices. Two undergraduate students at Cal Poly took a random sample¹⁶ of 30 textbooks from the campus bookstore in the fall of 2006. They recorded the price and number of pages in each book, in order to investigate the question of whether the number of pages can be used to predict price. Their data are stored in the file **TextPrices** and appear in Table 1.5.

Pages	Price	Pages	Price	Pages	Price
600	95.00	150	16.95	696	130.50
91	19.95	140	9.95	294	7.00
200	51.50	194	5.95	526	41.25
400	128.50	425	58.75	1060	169.75
521	96.00	51	6.50	502	71.25
315	48.50	930	70.75	590	82.25
800	146.75	57	4.25	336	12.95
800	92.00	900	115.25	816	127.00
600	19.50	746	158.00	356	41.50
488	85.50	104	6.50	248	31.00

Table 1.5: Pages and price for textbooks

- a. Produce the relevant scatterplot to investigate the students' question. Comment on what the scatterplot reveals about the question.
- b. Determine the equation of the regression line for predicting price from number of pages.

¹⁶Source: Cal Poly Student project.

- c. Produce and examine relevant residual plots, and comment on what they reveal about whether the conditions for inference are met with these data.

1.27 Baseball game times. What factors can help to predict how long a Major League Baseball game will last? The data in Table 1.6 were collected at www.baseball-reference.com for the 15 games played on August 26, 2008, and stored in the file named **BaseballTimes**. The *Time* is recorded in minutes. *Runs* and *Pitchers* are totals for both teams combined. *Margin* is the difference between the winner's and loser's scores.

Game	League	Runs	Margin	Pitchers	Attendance	Time
CLE-DET	AL	14	6	6	38774	168
CHI-BAL	AL	11	5	5	15398	164
BOS-NYY	AL	10	4	11	55058	202
TOR-TAM	AL	8	4	10	13478	172
TEX-KC	AL	3	1	4	17004	151
OAK-LAA	AL	6	4	4	37431	133
MIN-SEA	AL	5	1	5	26292	151
CHI-PIT	NL	23	5	14	17929	239
LAD-WAS	NL	3	1	6	26110	156
FLA-ATL	NL	19	1	12	17539	211
CIN-HOU	NL	3	1	4	30395	147
MIL-STL	NL	12	12	9	41121	185
ARI-SD	NL	11	7	10	32104	164
COL-SF	NL	9	5	7	32695	180
NYM-PHI	NL	15	1	16	45204	317

Table 1.6: Major League Baseball game times

- First, analyze the distribution of the response variable (*Time* in minutes) alone. Use a graphical display (dotplot, histogram, boxplot) as well as descriptive statistics. Describe the distribution. Also, identify the outlier (which game is it?) and suggest a possible explanation for it.
- Examine scatterplots to investigate which of the quantitative predictor variables appears to be the best single predictor of time. Comment on what the scatterplots reveal.
- Choose the one predictor variable that you consider to be the best predictor of time. Determine the regression equation for predicting time based on that predictor. Also, interpret the slope coefficient of this equation.
- Analyze appropriate residual plots and comment on what they reveal about whether the conditions for inference appear to be met here.

1.28 Baseball game times (continued). Refer to the previous Exercise 1.27 on the playing time of baseball games.

- a. Which game has the largest residual (in absolute value) for the model that you selected? Is this the same game that you identified as an outlier based on your analysis of the time variable alone?
- b. Repeat the entire analysis from the previous exercise, with the outlier omitted.
- c. Comment on the extent to which omitting the outlier changed the analysis and your conclusions.

1.29 Retirement SRA. A faculty member opened a supplemental retirement account in 1997 to investment money for retirement. Annual contributions were adjusted downward during sabbatical years in order to maintain a steady family income. The annual contributions are provided in the file **Retirement**.

- a. Fit a simple linear regression model for predicting the amount of the annual contribution (*SRA*) using *Year*. Identify the two sabbatical years that have unusually low *SRA* residuals and compute the residual for each of those cases. Are the residuals for the two sabbatical years outliers? Provide graphical and numerical evidence to support your conclusion.
- b. Sabbaticals occur infrequently and are typically viewed by faculty to be different from other years. Remove the two sabbatical years from the dataset and refit a linear model for predicting the amount of the annual contribution (*SRA*) using *Year*. Does this model provide a better fit for the annual contributions? Make appropriate graphical and numerical comparisons for the two models.

1.30 Metabolic rate of caterpillars. Marisa Stearns collected and analyzed body size and metabolic rates for *Manduca Sexta* caterpillars.¹⁷ The data are in the file **MetabolicRate** and the variables are:

Variable	Description
<i>Computer</i>	Number of the computer used to obtain metabolic rate
<i>BodySize</i>	Size of the animal in grams
<i>CO2</i>	Carbon dioxide concentration in parts per million
<i>Instar</i>	Number from 1 (smallest) to 5 (largest) indicating stage of the caterpillar's life
<i>Mrate</i>	Metabolic rate

The dataset also has variables *LogBodySize* and *LogMrate* containing the logs (base 10) of the size and metabolic rate variables. The researchers would like to build a linear model to predict metabolic rate (either *Mrate* directly or on a log scale with *LogMrate*) using a measure of body size for the caterpillars (either *BodySize* directly or on a log scale with *LogBodySize*).

¹⁷We thank Professor Itagaki and his research students for sharing these data.

- a. Which variables should they use as the response and predictor for the model? Support your choice with appropriate plots.
- b. What metabolic rate does your fitted model from (a) predict for a caterpillar that has a body size of 1 gram?

1.31 More metabolic rate of caterpillars. Refer to Exercise 1.30 that considers linear models for predicting metabolic rates (either *Mrate* or *LogMrate*) for caterpillars using a measure of body size (either *BodySize* or *LogBodySize*) for the data in **MetabolicRate**. Produce a scatterplot for the model you chose in Exercise 1.30 and add a plotting symbol for the grouping variable *Instar* to show the different stages of development. Does the linear trend appear to be consistent across all five stages of a caterpillar's life? (Note: We are not ready to fit more complicated models yet, but we will return to this experiment in Chapter 3.)

Supplemental Exercises

1.32 Zero mean. One of the neat consequences of the least squares line is that the sample means (\bar{x}, \bar{y}) always lie on the line so that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. From this, we can get an easy way to calculate the intercept if we know the two means and the slope:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

We could use this formula to calculate an intercept for *any* slope, not just the one obtained by least squares estimation. See what happens if you try this. Pick any dataset with two quantitative variables, find the mean for both variables, and assign one to be the predictor and the other the response. Pick any slope you like and use the formula above to compute an intercept for your line. Find predicted values and then residuals for each of the data points using this line as a prediction equation. Compute the sample mean of your residuals. What do you notice?