Contemporary Pre-Calculus Through Applications

Contemporary Pre-Calculus Through Applications

Mathematics Department
North Carolina School of Science and Mathematics

September 17, 2018

Contents

Chapter 1

Data

Note to Students:

Contemporary Precalculus Through Applications will introduce you to many new mathematical ideas and methods and will give you opportuniities to use those ideas and methods in creative ways to describe everyday phenomena. At times, you will consider mathematical ideas learned in previous courses from a deeper, more mathematical perspective. Chapter 1 sets the stage for the approach that we will use throughout the year. You will have the opportunity to engage in mathematical modeling activities that open the door to new mathematical concepts. As you work through Chapter 1 with your classmates, don't get lost in the details. Details are important, and you will deal with them in later chapters, but for now, in this introductory chapter, we want you to look at the big picture. We want you to get comfortable with working as a team, explaining your ideas to your peers, making your own decisions about how to proceed and determining for yourself whether your choices have been good one, interpreting your computational results within the context of the problem setting, and making sense out of the mathematics. In mathematics, understanding why a method works is just as essential as knowing how to apply it. And this understanding makes it easier to decide what approach might be most fruitful in different situations.

So, keep your focus on the big ideas and mathematical practices, and welcome to Precalculus at NCSSM.

1.1 Some Thoughts about Models and Mathematics

When children think about models, they are generally considering some kind of a toy, perhaps an airplane or a dinosaur. When scientists and mathematicians think about models, they are generally considering a model as a tool, even though they may be thinking about the same airplane or dinosaur. Scientists and mathematicians use models to help them study and understand the physical world. People in all walks of life use models to help them solve problems; problems in this course will involve models used by bankers, anthropologists, geologists, and many, many others.

So just what is a model? Models are simplified representations of phenomena. To be useful, a model must share important characteristics with the phenomenon it represents, and it must also be simpler than what it represents. A model usually differs significantly from what it represents, but these differences are offset by the advantage that comes from simplifying the phenomenon. A good example is a road map, which models the streets and highways in a particular area. Clearly, a map has a lot in common with the actual streets and highways it shows how roads are oriented and where they intersect. A road map simplifies the situation; it ignores stoplights, hills, and back alleys and instead focuses on major thoroughfares. Such a map is very useful for traveling from one city to another, but is not much good for finding the quickest route to the shopping mall or the best street for skateboarding. Road maps, and most other models, are useful precisely because they ignore some information and thereby allow you to see other information more clearly.

Another fairly common model is an EKG, which models the electrical activity of the heart. The EKG is an excellent model when used to determine the heart rate or to find

which regions of the heart may be damaged after a heart attack. It is not a useful model for determining the volume of blood flowing through the heart. Different models emphasize different aspects of a phenomenon; the choice of which model to use depends on which aspect is under investigation.

The ability to predict is the ultimate test for a model. A good model allows us to make accurate predictions about what will occur under certain conditions. If what actually occurs is very different from our prediction, then the model is of little use. Scientists and mathematicians often need to update or revise models as more is learned about the phenomenon under study. Sometimes a model needs to be completely discarded and replaced with a new one. For example, before Columbus sailed to the Americas, many people believed the world was flat, but that model was quickly abandoned in light of new information.

Even though Isaac Newton's models for the actions of a gravitational field have been replaced by Einstein's relativistic model, we still use Newtonian physics in many situations because it is easier and because it gives reasonably accurate results. The aspects of Einstein's mechanics that are ignored are largely irrelevant in most everyday applications, so the Newtonian model is still a good one.

As we move through the course, we will encounter phenomena that we want to know more about. Our task will be to find a mathematical expression or a graph that mimics the phenomenon we are interested in. This model must accurately represent the aspects of the phenomenon that we care about, but it may be very different from the phenomenon in other ways. To be able to find a model to represent a problem, we need to have a large toolkit of mathematical information and techniques at our disposal. The fundamental concepts of Algebra and Geometry are all a part of our toolkit. We will also use the calculator and computer as tools to construct and analyze models for the phenomena we study. Probably the most important tools necessary for making models are an inquisitive mind and a determined spirit.

Often we will not stop after we have developed one model but will form two or three to get a better view of the subject. For example, suppose a rock is thrown vertically into the air. How can its height be modeled? We can use a graph as a model for this phenomenon (see Figure 1).

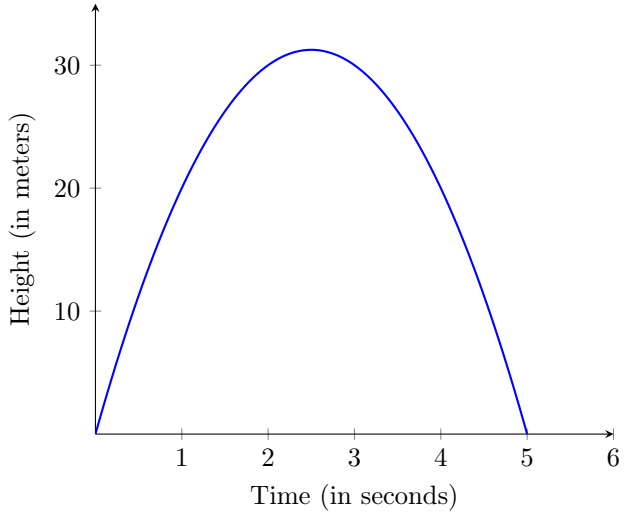

Figure 1.1.1: Height of a rock over time

We can also describe the height by an equation. If the height above the ground is represented by h and the time is represented by t, then a simple model for the height as a function of time is $h = -4.9t^2 + 25.1t + 1$.

Notice that these models do not give complete information about the problem. We cannot tell from the models what type of rock was thrown, who threw it, or why. These aspects of the problem are not relevant, since all we really care about is height and time. Both models give this information. The strength of any model is in what it emphasizes and

what it ignores.

1.2 A Whole Class Modeling Activity

The process of developing a mathematical model is often challenging, and it is almost never a one-step process. In modeling it is important that you think about both the mathematics you know and the phenomenon you are trying to model. Moving back and forth between the two in a thoughtful, organized manner is essential. Precalculus is the study of the basic functions that we use to describe our world. Throughout the course, you will learn how to use your growing knowledge of functions to model real-world situations. In this section, we will work through a sample modeling problem to demonstrate some useful modeling techniques.. You are not expected to be able to do this problem by yourself at this point in the course. By the end of the course, however, you should be comfortable with the modeling process and confident in your ability to solve problems similar to this one.

1.2.1 Problem Setting

The senior class at the local high school wants to raise money to support the athletic program by selling a ticket that will allow the holder to attend all athletic events at the school. The class officers are trying to decide what price to charge for the ticket. Some students might argue for setting the price low, believing that a low price would bring a large response. Others may want to have a higher price, so that even if not many tickets were sold, they would still make money. The students decide to ask the parents what they would be willing to pay for an all sports ticket. They assume the parents want the sale to be a success and will give them accurate information. A survey is sent to all 811 families with students in the school asking, "What is the most you would be willing to pay for an all sports ticket good for this school year?" The results are given in Figure 1.

Maximum Price ($)	Number
50.00	145
75.00	80
90.00	45
95.00	85
115.00	120

Figure 1.2.1: Results of the Survey

Take a minute to think about this problem setting. What do you expect to be the relationship between the price the students set for the tickets and the response to the sale? How can the class officers use this information to determine the "best" price to charge for the ticket? Imagine that you are in charge of the sale and it is your responsibility to determine the price at which the tickets are to be sold. Where do you begin?

To determine the price that will bring in the most money to the class, you need to develop a mathematical model relating the amount charged for the ticket and the amount of money, or revenue, brought in by the sale. To develop this model, you must understand the information provided by the data collected. What information about the parents and their support for the sale is contained in the data? Does this data support the conjecture that the more the ticket costs, the fewer families would be interested?

Information presented in a list of numbers is often hidden and difficult to determine. One step in analyzing the relationship between two variables is to make a scatterplot. A scatterplot is simply a graph in a rectangular coordinate system of all ordered pairs of data. Scatterplots display data so we can see the general relationship between two variables. The relationship (or lack thereof) should be more obvious if we plot the data. When making a scatterplot, it really does not matter which variable is plotted on which axis. If we suspect that one variable depends on the other, however, we usually plot the dependent variable

on the vertical axis and the independent variable on the horizontal axis. Figure 2 shows a scatterplot of the ordered pairs (price, number).

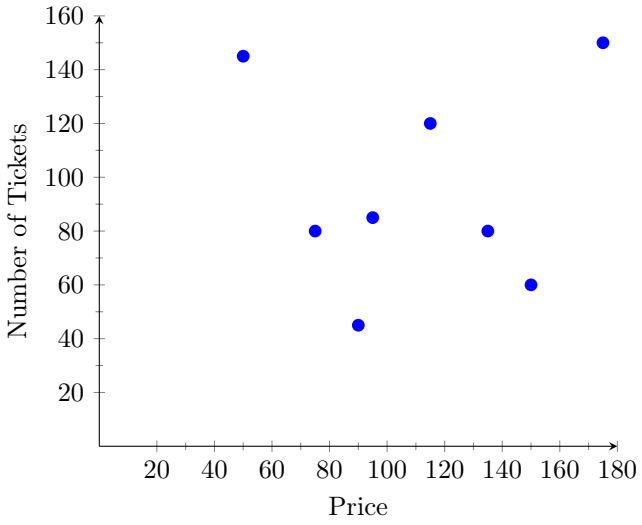

Figure 1.2.2: Scatterplot of the data in Figure 1

It is always a good idea to plot the data, although this particular plot does not seem to give us much useful information. There is no obvious pattern in the data. Perhaps we need to think harder about what the data are telling us.

Did all of the families respond to the survey question? If a family did not respond, what does this mean about their interest in the tickets? All models begin with some simplifying assumptions. One way to think about the families that didn't respond is to assume that they are not interested in supporting the athletic program and will not buy a ticket at any price. There are other assumptions we could make about those families that didn't respond to the survey, of course, and we will consider some of them in the homework exercises. For now, we will assume that only those families that responded will purchase a ticket. With this assumption, we can interpret the data.

According to the results of the survey, there are 150 families willing to pay as much as 175 for the tickets. If they will pay \$175, they will certainly pay less for the tickets as well. In particular, they will buy the tickets if they are priced at \$150. In addition to these 150 families, the 60 families who responded that \$150 is the most they would be willing to pay will also purchase tickets. Summing these numbers, we would expect $150+60 = 210$ families to purchase tickets priced at \$150.

Continuing to sum, we can create a table Figure 3 showing how many families we expect to purchase tickets at each price. If we knew how many families would buy a ticket at each price, we can use that information to predict the price that will bring in the most money.

Maximum Price (\$)	Number
50.00	765
75.00	620
90.00	540
95.00	495
115.00	410

Figure 1.2.3: Price and number of tickets expected to be sold

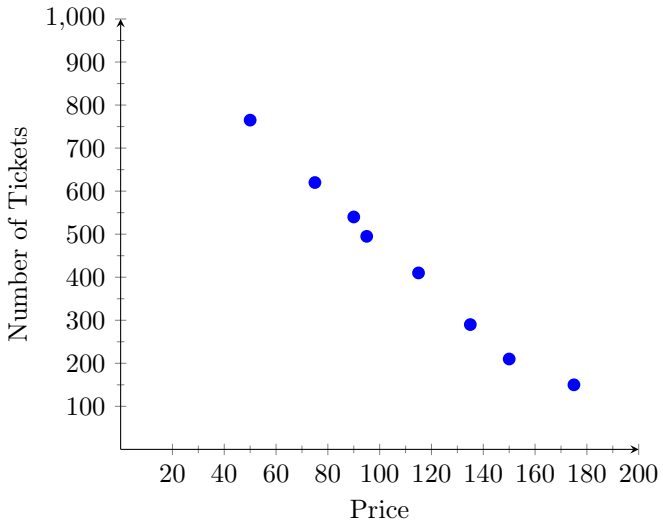

Figure 1.2.4: Scatterplot of the data in Figure 3

The scatterplot of this data, given in Figure 4, gives useful information about the relationship between the price charged for the ticket and the number of tickets the students can expect to sell. Notice that this graph supports the conjecture that higher prices result in lower sales. If we could find a mathematical equation relating price and number of tickets sold, we could approximate the number of tickets to be sold at prices that are not on the list.

The general pattern of the points in the graph in Figure 4 is linear. If we place a pencil over the graph of the data, we see that the pencil does a good job of modeling the relationship between the two variables.

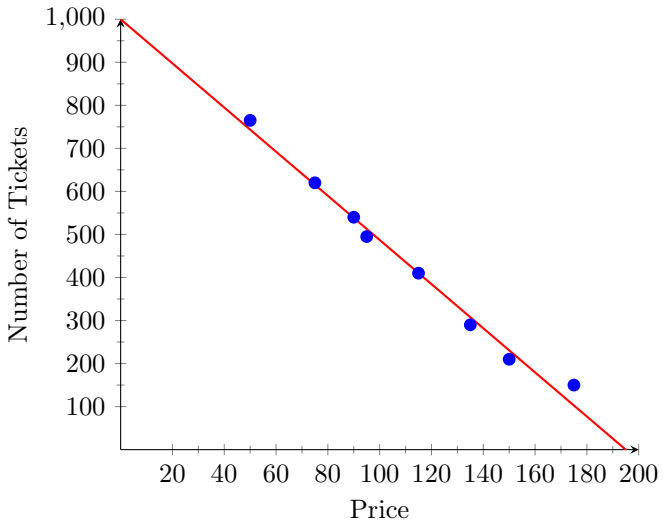

Figure 1.2.5: Pencil Model

What is the equation of the line represented by the pencil? There are a number of ways to find an equation of a line that does a good job of modeling the data set. We will look at two standard techniques later in this chapter. For now, a quick estimate will do. It appears from the graph that the price-intercept (where the ticket sales are zero) is approximately \$195 dollars. The ticketintercept (where the price is zero) is approximately 1000 tickets. This means that the two points $(195, 0)$ and $(0, 1000)$ lie on our line. The equation of the

line passing through these two points is

$$\text{Tickets} = -\frac{200}{39} \cdot \text{Price} + 1000.$$

Using function notation, we say that the number of tickets sold, T, is a function of price, P, and write

$$T(P) = -\frac{200}{39} \cdot P + 1000.$$

Using this function, we can predict how many tickets will be sold for different prices. For example, if we charge \$60 per ticket, we would expect to sell around $T(60) \approx 692$ tickets, while if we charge \$110, we would expect to sell only around $T(100) \approx 436$.

When working on multiple-step problems, it is easy to lose your focus and forget how the present process helps reach the final goal. It is important to stop periodically to compare where we are in the process of solving the problem with our original goal. The students want to find a relationship between the price they charge and the revenue from the sale of the tickets. They want this model so they can determine the price to charge to make the most money. The function $T(P) = -\frac{200}{39} + 1000$ doesn't tell us this directly. It only tells us how many tickets we can expect to sell for a specified price. It is important to note that this is not what we wanted to find, but it is what we could find from the data the students gathered. We can now use this function to answer the question we really want to know, that is, what price will bring in the most money?

We know that if we charge \$60, we would expect to sell around 692 tickets and bring in around $\$41,520$. If we sell the tickets at \$110 each, we will sell only around 436 of them, but 436 tickets sold for \$110 each brings in $\$47,960$, so a price of \$110 is better than \$60. The revenue expected from the sale of the tickets is the product of the number of tickets expected to be sold, given by $T(P)$, and the price, P. Written in function notation, we say that the revenue, R, generated from the sale of $T(P)$ tickets is

$$R(P) = P \cdot P(T) = P\left(-\frac{200}{39}P + 1000\right).$$

In this case, revenue is a quadratic function of price. A view of the revenue function is given by looking at its graph Figure 6, which we recognize as a parabola.

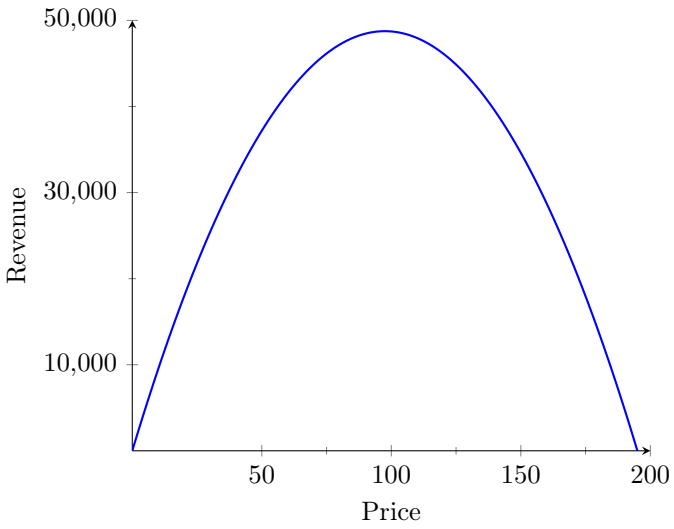

Figure 1.2.6: Graph of revenus function $R(P) = P\left(-\frac{200}{39}P + 1000\right)$

To find the price that will generate the maximum revenue, we need to recall what we learned about quadratic functions in earlier courses. We know that a parabola has a vertex, which in this problem represents the maximum revenue for the students' project. We could use our calculators to approximate the coordinates of the vertex, but it is quicker to recall

that the vertex of a parabola is mid-way between the zeros. From our function, we know that $R(P) = 0$ at $P = 0$ and $P = 195$, so the vertex is located at $P = \frac{195-0}{2} = 97.5$. Thus, we should charge \$97.50 for the tickets. If we charge \$97.50, we expect to sell around 500 tickets and to receive $\$48,750$ in revenue.

1.2.2 Reviewing the Process

Let's step back and look at what we have done in this problem. We were faced with a question, "What price should the seniors charge to maximize the money brought in by the project?" We had some ideas about the relationship between the price and participation. We believed that the more they charged for a ticket, the fewer tickets they would sell but these ideas were not quantified. We needed to determine how much participation would drop with each dollar increase in price.

To quantify this relationship, we looked at the information from the parents' survey. We realized that the data generated from the survey did not directly lead to the desired relationship. To find the number of tickets they could expect to sell at each price, we had to create a new data set by accumulating the survey data. After creating a data set that represented this relationship, we looked at the graph of the scatterplot and observed a linear pattern. We then used our knowledge of lines to fit a linear model to approximate the number of tickets to be sold for prices that were not on the survey.

Once we had a linear model for the expected level of participation, we used this to generate a quadratic function that modeled the expected revenue. Lastly, we used our knowledge of quadratic functions to find the optimum price and the maximum revenue.

Throughout the process, we had to stay focused on the question at hand. We made progress by calling upon mathematics at some points and our understanding of the problem setting at others. In this particular problem, the mathematics of finding the equation of a line and finding the vertex of a parabola should be familiar. However, the process of modeling and knowing when and how to use those mathematical techniques may be new. Don't be concerned if you could not have done this problem on your own. Learning precalculus mathematics and how to use that mathematics in problem settings such as this one is what this course is all about.

Exercises

1. In the linear equation for the number of tickets sold, $T(P) = -\frac{200}{39}P + 1000$, interpret the meaning of the slope, the P-intercept, and the T-intercept in the context of the model.

2. What question would the students ask if they wanted to generate the values in Figure 3 directly from their questionnaire?

3. Suppose we thought that the responses received from the questionnaire represented a sample of the 811 families. Those who didn't return their surveys are still interested, but they either forgot or didn't have time to fill out the survey. In this case, we could assume that the number of families interested follows proportionally from the results of the survey. That is, since 80 of the 765 responses, or 10.46%, reported that the most they would pay would be \$75, then 10.46% of the total population of 811 families would be willing to pay at most \$75 for the tickets. Rework the problem based on this assumption and determine the "best" price.

1.3 The Average as a Mathematical Model

It is common for students to have studied simple methods of data analysis in earlier mathematics classes, including experience with mean, median, standard deviation, "lines of best fit" and correlation, among other topics. Many of the ideas in this chapter may be familiar to you. In this course, we will take some care to not just look at how computations are made, but to look behind the computation to the mathematical models underlying them,

and to think carefully how they may be used and interpreted.

As a simple example to get started, consider the average, which is a computation you have done routinely since elementary school. In what way is the average a good representation of the data from which it is being computed? The average, or mean, of a set of numbers is often called a "measure of central tendency" and is commonly interpreted as a "typical value" in the set. Our question isn't about how to compute this and other measures of central tendency, but in what way do they "measure" central tendency? In what way does each represent a typical value? What aspects of the data are being modeled by the mean, median, and mode?

Example 1.3.1 . Suppose your quiz grades (scored 0 to 12) in Precalculus are $\{2, 3, 3, 6, 7, 7, 7\}$. Your teacher needs to use a single number to convey information to your parents about your quiz performance. What number should she use?

You have probably learned several different numbers that could be used in this situation. Of course, the most common is the average or mean of the data. You might also have studied the median, the mode, and perhaps the midrange values. The mean of the data is 5, the median is 6, the midrange is 4.5, and the mode is 7. Each measure answers a different question about the data. We will look at these different measures and consider "what question about the data set $\{2, 3, 3, 6, 7, 7, 7\}$ does each measure answer?
"

The Mode: The mode is the value that is seen most often. If no value is seen more than others, then the set has no mode. If there are two values that appear more than others, we say the data is bi-model. In what way does the mode model the data? Suppose you were given another quiz and asked to guess what score you would likely get, what would you say? Surely not the average of 5, since you have never yet made a 5 on any quiz.

The mode models the data in the following manner: Imagine we write down each data value on a slip of paper and put them in a bag. If you pull a slip of paper out of the bag at random, what number are you most likely to see? Since there are more 7's in the bag than any other number, it is the most likely observation. The mode answers the question, what is most likely, so it is, in a certain sense, a measure of typicality based on probability. The mode defines "typical" as most likely, but doesn't pretend to be representative of the center of the data, so it is not truly a measure of central tendency since it considers only one (or two if the data is bi-modal) data values and ignores all the rest.

The Midrange: Another common measure of centrality or typicality is the midrange. The midrange is the value half-way between the smallest and largest value in the set. How does the midrange model the data? Consider, again, the numbers on slips of paper in a bag. If we choose a single number to represent all the numbers in the bag, then pull a number out of the bag, what is the largest error we can make? For example, if we choose the mean of 5 and we pull out the 2, we are off by 3. If we choose the mode 7, our largest error would be 5. If we choose the median 6, our maximum error would be 4. We would like to choose the number that minimizes the maximum possible error. Minimizing the maximum error is called the mini-max criterion. The midrange satisfies the mini-max criterion for your quiz scores. In this example, that is $\frac{7+2}{2} = 4.5$. No matter what you pull out of the bag, the error can be no more than 2.5. Since the mini-max value in this context is the middle of the range, it is commonly called the midrange.

Mean and Median: The mean (average) and median also are solutions to some question about minimizing error. The basic idea is that we want to use a single number to represent a set of different numbers. We want the number to "represent" the set in some way, perhaps by estimating the center of the set or by representing a typical value. When your teacher says that your quiz average is 5, she doesn't mean that all your quiz scores have been a 5, but you're your scores are centered around 5, or that 5 is a typical score for

you on quizzes.

Mathematically, we think about it this way: We want to choose a value v to represent the set $\{2, 3, 3, 6, 7, 7, 7\}$. So, instead of the set $\{2, 3, 3, 6, 7, 7, 7\}$, we will use $\{v, v, v, v, v, v, v\}$. Your teacher might inform your parents, "in general, Melinda tends to make a score of v on her quizzes". How wrong can the teacher be if she says this? This depends on how we measure how wrong a choice is. Minimizing how wrong you would be to use v instead of the individual values is the basic principle upon which the mean and median are based. They model the total error by comparing the choice, v, to all the elements in the set and try to minimize that total error.

We can define the Total Error as the sum of all the errors when using v instead of the actual values, so

$$TE(v) = (2 - v) + (3 - v) + (3 - v) + (6 - v) + (7 - v) + (7 - v) + (7 - v)$$

TE can be simplified to $TE(v) = (2 - v) + 2 \cdot (3 - v) + (6 - v) + 3 \cdot (7 - v) = 35 - 7v$. So, one model for the total error is the equation $TE(v) = 35 - 7v$. If we choose $v = 3$, as our representative value, then $TE(3) = 14$. If we choose $v = 10$, then $TE(10) = 35$. Can we find a value of v that generates the smallest possible Total Error? There is a problem with this formulation of the model: there is no minimum value of TE!

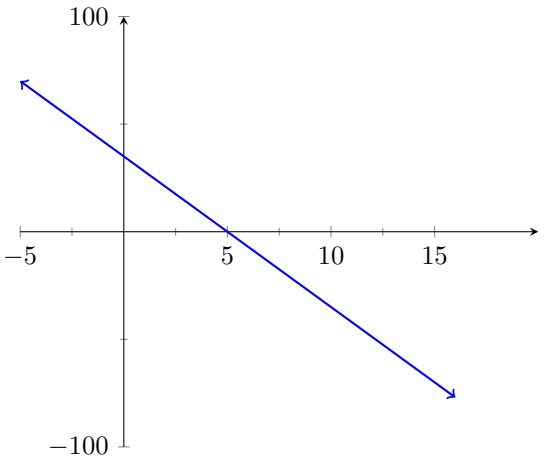

Figure 1.3.2: Total Error: $TE = 35 - 7v$

We see that the equation $TE = 35 - 7v$ is a linear equation, and we can make the equation as small as we choose by choosing v to be large, since negative values are smaller than positive values. When thinking about errors being small, we usually mean being close to zero, not -100. So, we need to alter our definition of total error to include "close to zero" as our measure (called a metric) of small. We want to exclude negative values from consideration. There are two simple ways to alter the definition to accomplish this: by using the absolute value or by squaring. We will consider three different models for the new Total Error function.

$$TE_1 = |(2 - v) + (3 - v) + (3 - v) + (6 - v) + (7 - v) + (7 - v) + (7 - v)| = |35 - 7v|$$

TE_1 is read "the absolute value of the sum of the errors".

$$TE_2 = |2 - v| + |3 - v| + |3 - v| + |6 - v| + |7 - v| + |7 - v| + |7 - v|$$

TE_2 is read "the sum of the absolute value of errors".

$$TE_3 = (2-v)^2+(3-v)^2+(3-v)^2+(6-v)^2+(7-v)^2+(7-v)^2+(7-v)^2 = 7v^2-70v+205$$

TE_3 is read "the sum of the squares of the errors".

The graphs of these functions give some insight into how each of these "measures of central tendency" estimate the center of the data by minimizing the total error. The minimum value is indicated.

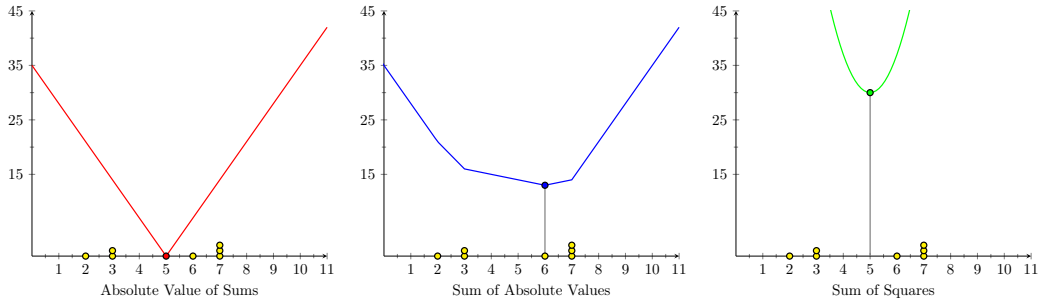

Figure 1.3.3: Three Metrics for Minimizing Error

We see that the mean appears to minimize the absolute value of the sum of errors, the median minimizes the sum of the absolute values of the errors, and the mean also minimizes the sum of the squares of the errors. Is this coincidence or will this always be true?

Derivations of Minimizing Properties of Median and Mean

Consider the set of n numbers $\{x_1, x_2, x_3, \cdots, x_n\}$. Assume for convenience, that n is odd, the number in the set are written in increasing order, and all n values are distinct. Let v be the value chosen to model this set of numbers.

Minimize the Absolute Value of the Sum of the Errors

$$\begin{aligned} TE_1 &= |(x_1 - v) + (x_2 - v) + (x_3 - v) + \cdots + (x_n - v)| \\ &= |(x_1 + x_2 + x_3 + \cdots + x_n) - (v + v + v + \cdots + v)| \\ &= \left| \sum_{i=1}^{n} x_i - nv \right| \end{aligned}$$

We know that the smallest the absolute value can be is zero, so if $\left| \sum_{i=1}^{n} x_i - nv \right| = 0$, then $\sum_{i=1}^{n} x_i = nv$, so $v = \frac{\sum_{i=1}^{n} x_i}{n}$. We recognize this as the formula for the mean. The mean is the one number that will minimize the absolute value of the sum of the errors. We can also see that the reason the absolute value is minimized is that the sum of the deviations from the mean is zero. This is an important attribute of the mean.

Minimize the Sum of the Absolute Values of the Errors

$$TE_2 = |x_1 - v| + |x_2 - v| + |x_3 - v| + \cdots + |x_n - v|$$

This function is more challenging to think about. Recall that the absolute value is

defined as $|a| = \begin{cases} a, & \text{if } a \geq 0 \\ -a, & \text{if } a < 0 \end{cases}$ Using this definition, we see that

$$|x_1 - v| = \begin{cases} x_1 - v, & \text{if } x_1 - v \geq 0 \\ v - x_1, & \text{if } x_1 - v < 0 \end{cases},$$

or equivalently,

$$|x_1 - v| = \begin{cases} x_1 - v, & \text{if } v \leq x_1 \\ v - x_1, & \text{if } v > x_1 \end{cases}.$$

Similarly, for

$$|x_2 - v| = \begin{cases} x_2 - v, & \text{if } v \leq x_2 \\ v - x_2, & \text{if } v > x_2 \end{cases},$$

and all the other absolute values are evaluated the same way.

So, as we observed in the graph above, we have a piece-wise defined function with $n+1$ pieces. For $v < x_1$, $TE_2 = (x_1 - v) + (x_2 - v) + (x_3 - v) + \cdots + (x_n - v)$ since all expressions in the absolute values are positive. This gives, $TE_2 = (x_1 + x_2 + x_3 \cdots + x_n) - nv$, which is a linear equation with a slope of $-n$ and an intercept that is the sum of all the numbers in the set. For $x_1 \leq v$, $TE_2 = (x_1 - v) + (x_2 - v) + (x_3 - v) + \cdots + (x_n - v)$ since the first expression was negative before applying the absolute value. This gives $TE_2 = (-x_1 + x_2 + x_3 \cdots + x_n) - (n-2)\, v$ and an intercept that is smaller than the previous piece. For $x_2 \leq v < x_3$, $TE_2 = (x_1 - v) + (x_2 - v) + (x_3 - v) + \cdots + (x_n - v)$ since the first two expressions were negative before applying the absolute value. This gives $TE_2 = (-x_1 - x_2 + x_3 \cdots + x_n) - (n-4)\, v$, which is a linear equation with a slope of $-(n-2)$ and an intercept that is smaller than the previous piece. We continue until the final two components are:

For $x_{n-1} \leq v \leq x_n$,

$$\begin{aligned} TE_2 &= (v - x_1) + (v - x_2) + (v - x_3) + \cdots + (v - x_{n-1}) + (x_n - v) \\ &= (-x_1 - x_2 - x_3 \cdots - x_{n-1} + x_n) + (n-2)\, v \end{aligned}$$

and for $v > x_n$,

$$\begin{aligned} TE_2 &= (v - x_1) + (v - x_2) + (v - x_3) + \cdots + (v - x_{n-1}) + (x_n - v) \\ &= (-x_1 - x_2 - x_3 \cdots - x_{n-1} + x_n) + nv \end{aligned}$$

$TE_2(v)$ is a function of v, and as v increases the expressions inside the absolute values change sign. For each piece of the function, we see that each time v increases past one of the data values, the expression inside one of the absolute values changes sign, and, as a consequence, the slope increases by 2 and the intercept becomes smaller. Since the slope for a set of n numbers began with a value of $-n$, and ends with a slope of n, we see that the slope marches through values of $-n, -n+2, -n+4, \cdots, n-2, n$. For $v < x_1$ and $v > x_n$, this function is the same as TE_1. But in between, the slopes of the line segments form a raggedy U-shape. For what value of v is this sum of absolute values the smallest? Notice that this happens when v is the median of the set of numbers. Can you explain why?

We have seen that, if we define the size of the total error as the absolute value of the sum of the differences between our chosen value and the actual data points, then the mean is the best choice for a single number to represent the data set. If we think the sum of the absolute values of the errors is a better way to measure total error, then we should be using the median. What methods we use to model often depends upon how we measure the quality of the model. In this example, we used an odd number of values in the set. In

the exercises, you will be asked to consider what happens when an even number are in the set.

Minimize the Sum of the Square of the Errors

The sum of quadratic equations is another quadratic equation, so we expand each squared error term and combine like terms.

$$\begin{aligned} TE_3(v) &= (x_1 - v)^2 + (x_2 - v)^2 + (x_3 - v)^2 + \cdots + (x_n - v)^2 \\ &= \left(x_1^2 - 2x_1v + v^2\right) + \left(x_2^2 - 2x_2v + v^2\right) + \left(x_3^2 - 2x_3v + v^2\right) + \cdots + \left(x_n^2 - 2x_nv + v^2\right) \\ &= \left(x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2\right) - 2(x_1 + x_2 + x_3 + \cdots + x_n) \cdot v + nv^2 \end{aligned}$$

The minimum value of a quadratic is always at the vertex. Recall if $y = ax^2 + bx + c$, the vertex is located at $x = -\frac{b}{2a}$. In this quadratic function, we have $a = n$, $b = -2(x_1 + x_2 + x_3 + \cdots + x_n)$, and $c = \left(x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2\right)$. So the vertex is $v = -\frac{2(x_1+x_1+x_3+\cdots+x_n)}{2n} = \frac{\sum_{i=1}^{n} x_i}{n}$. This is again the mean. The mean, which is commonly denoted $\bar{x}$ is the one number that will minimize the sum of the squares of the errors, that is, it is a *least squares* measure of typicality or center. Least squares is a very common method for minimizing error.

Why squares?

When first exposed to the models behind these classic measures of central tendency, many students favor the ideas behind the median. The absolute values seems to make more sense than squaring does, since squaring exaggerates the size of the errors and absolute values only affect sign. But almost everyone (certainly every teacher) uses the mean when "averaging" grades. Why? What's so special about the mean? The underlying idea behind all least squares methods in mathematics and statistics is simple: the Pythagorean Theorem. It is the way we measure distance. If you want to know how far apart the points $(1, 4)$ and $(5, -2)$ are, we use the distance formula. The distance, D, between the points in the plane $(1, 4)$ and $(5, -2)$ is

$$D = \sqrt{(1-5)^2 + (4-(-2))^2} = \sqrt{16 + 36} = \sqrt{52}$$

or $D \approx 7.211$ units. Notice we summed squares of differences in the calculation.

If we ask, "how far is the point (a, b) from the point $(1, 4)$?", the answer is given by the sum of squares expression $D = \sqrt{(1-a)^2 + (4-b)^2}$ If we want to determine the point closest to a line, that is, makes the perpendicular distance smallest, we will minimize this sum of squares.

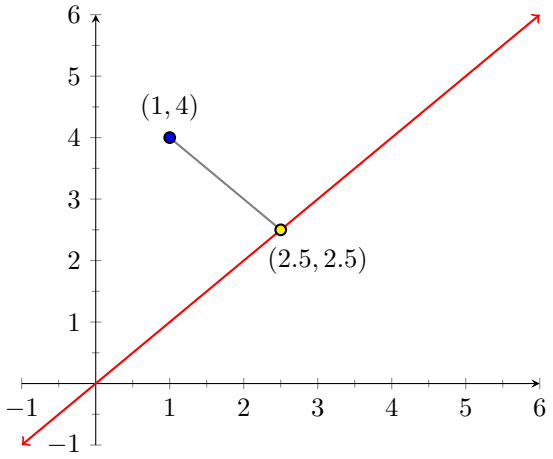

Figure 1.3.4: Point on $y = x$ closest to $(1, 4)$

The question we are asking about the mean is equivalent to finding a closest point to a line. When we want to use a single number to represent a set of numbers, we can ask, "what point (v, v) (which has both coordinates the same) is closest to the point $(1, 4)$?" The answer is, "whatever value of v makes $D = \sqrt{(1-v)^2 + (4-v)^2}$ the smallest." And we know the answer. The smaller a number is, the smaller its square root is. So, we only need to minimize $(1-v)^2 + (4-v)^2$. And we know that the mean of 1 and 4 gives the solution to this least squares question. So, the point in the plane with both coordinates the same (lies on the line $y = x$) closest to the point $(1, 4)$ is $(2.5, 2.5)$. We require our point to line on $y = x$ because we want to use only one value, v, to represent the numbers 1 and 4. Notice also, that the segment between $(2.5, 2.5)$ and $(1, 4)$ is perpendicular to the line $y = x$.

The smaller the value of $(2-v)^2 + (3-v)^2 + (7-v)^2$, the smaller the square root will be, so we don't need to consider the square root in finding v. We showed above that the mean is the value of v we want. The average of $\{2, 3, 7\}$ is 4. To see that this works for

$$D = \sqrt{(2-v)^2 + (3-v)^2 + (7-v)^2},$$

we can graph the function or expand the quadratics

$$D = \sqrt{(4 - 4v + v^2) + (9 - 6v + v^2) + (49 - 14v + v^2)} = \sqrt{62 - 24v + 3v^2}.$$

The vertex of the parabola $D = 62 - 24v + 3v^2$ is at $v = \frac{24}{6} = 4$, as expected. We can extend this computation into 3-space. What point in 3-space (v, v, v) is closest to $(2, 3, 7)$? In this case, we generalize the distance formula so that

$$D = \sqrt{(2-v)^2 + (3-v)^2 + (7-v)^2}.$$

The question we ask about what score should your teacher should use to represent your typical work on quizzes is actually a geometry question. Given the point in 7-space representing your seven quiz scores, $(2, 3, 3, 6, 7, 7, 7)$, what point using a single number, (v, v, v, v, v, v, v), is closest to it? We can't draw this, but the algebra and geometry of the Pythagorean Theorem is exactly the same as in the plane and in 3-space. The answer comes from the generalized distance formula,

$$D = \sqrt{(2-v)^2 + (3-v)^2 + (3-v)^2 + (6-v)^2 + (7-v)^2 + (7-v)^2 + (7-v)^2}.$$

The smaller the sum of squares inside the square root, the smaller the distance, so we just need to minimize $(2-v)^2 + (3-v)^2 + (3-v)^2 + (6-v)^2 + (7-v)^2 + (7-v)^2 + (7-v)^2$. And we know the mean is *always* the number that minimizes these sums.
The model that statisticians use is a geometric one, and the metric that determines whether one value is a better representative of a data set than another number is the geometric distance between the set you are given and the simplified set you are using. When we minimize the sum of the squares of the errors we are finding the shortest distance. The least in least squares refers to distance and the squares is just an application of the distance formula and the Pythagorean Theorem.

Exercises

1. A large university offers a beginning statistics course in a variety of formats. Some students receive credit from individual projects, others take small seminar classes, "regular" classes, or large lectures given at night. Each of the 30 faculty in the department teaches one of these classes each year (they teach other classes as well, but we are focusing on this course). The data for the size of each class and the number of sections offered is given below:

Class Size	Number of Sections
1	5
15	8
25	10
50	5
150	2

Figure 1.3.5: Results of the Survey

The primary role for statistical thinking is to turn data into useful information.

(a) What is the average class size from the faculty's perspective?

(b) What is the average class size from the student's perspective?

(c) Which average class size do you think the university will advertise?

2. Suppose your next quiz grade is a 4, so your scores are now $\{2, 3, 3, 4, 6, 7, 7, 7\}$. Find the value v that minimizes the sum of the absolute values of the errors. What problem do you observe from the graph with finding a specific value to use?

3. We have seen in our examples that the median, m, appears to minimize the sum of the absolute values of the errors when there are an odd number of values in the set. Show that both $m + \varepsilon$ and $m - \varepsilon$ give larger sums of absolute values than m in the sum

$$TE_2 = |x_1 - m| + |x_2 - m| + |x_3 - m| + |x_4 - m| + |x_5 - m| .$$

Assume for simplicity that $x_1 < x_2 < x_3 < x_4 < x_5$.

4. Compare the mean and medians for test grades of $\{85, 90, 92, 92, \text{ and } 95\}$ with those for $\{15, 90, 92, 92, \text{ and } 95\}$.

(a) Describe the effect on the mean and the median of the outlier score of 15. The mean is said to be *sensitive* to outliers while the median is said to be *resistant* to outliners.

(b) Are the mode and midrange sensitive or resistant measures of typicality?

1.4 Using Scatterplots to Analyze Data

We will consider graphical models as a first step in analyzing data. These models provide information so we can answer questions like the following:

- How are heating bill costs related to average temperature?
- What is the relationship between a state's average ACT score and the proportion of high school seniors taking the exam?
- What might be the winning time marathon in the 2024 Olympics?
- Is there a relationship between the amount of time a student spends studying for a test and the grade on the test?

Though the questions posed above are different in many respects, each requires the collection, organization, and interpretation of data. To answer each question, we need to analyze the relationship between several variables. We will limit our attention to paired measurements, that is, the analysis of *two variables*. Sometimes one variable depends on the other. For example, we expect that blood pressure in adults of the same height in some way depends on their weight and that crop yield depends on amount of rainfall. Other times there is a relationship between the variables, but it is not one of cause and effect or dependence. For example, we could show that there is a relationship between points scored and personal fouls committed by college basketball players, but we would probably not consider one of these variables to be dependent on the other. Sometimes there is no relationship at all between the two variables. For example, we do not expect there to be a relationship between the distance a student lives from school and his or her height.

To determine whether there is a relationship between two variables, we must analyze data consisting of ordered pairs. Sometimes these data are gathered from a well-designed, carefully controlled scientific experiment. Other times we want to analyze data that exist in the world around us.

Example 1.4.1 Average ACT Scores.

State	%	Score	State	%	Score
Alabama	100	19.1	Montana	100	20.3
Alaska	53	20	Nebraska	88	21.4
Arizona	58	20.1	Nevada	100	17.7
Arkansas	96	20.2	New Hampshire	23	24.5
California	33	22.6	New Jersey	32	23.1
Colorado	100	20.6	New Mexico	70	19.9
Connecticut	34	24.5	New York	29	23.4
Delaware	21	23.6	North Carolina	100	19.1
District of Columbia	44	22.2	North Dakota	100	20.3
Florida	81	19.9	Ohio	73	22.0
Georgia	60	21.1	Oklahoma	82	20.4
Hawaii	94	18.7	Oregon	39	21.7
Idaho	39	22.7	Pennsylvania	23	23.1
Illinois	100	20.8	Rhode Island	29	23.3
Indiana	41	22.3	South Carolina	100	18.5
Iowa	68	22.1	South Dakota	76	21.9
Kansas	74	21.9	Tennessee	100	19.9
Kentucky	100	20	Texas	46	20.6
Louisiana	100	19.5	Utah	100	20.2
Maine	10	23.6	Vermont	29	23.4
Maryland	27	23.0	Virgina	31	23.3
Massachusetts	28	24.8	Washington	25	23.1
Michigan	100	20.3	West Virginia	67	20.7
Minnesota	78	22.7	Wisconsin	100	20.5
Mississippi	100	18.4	Wyoming	100	20.2
Missouri	100	20.2	National Average	64	20.8

Figure 1.4.2: Average ACT Composite Score vs % Taking Exam by State

The data in Figure 2, provided by the ACT website, gives the average Composite ACT scores for high school students 2016 and the percentage of graduates taking the exam from each state.

Do you think there is a relationship between the average SAT score and the percentage of graduates taking the exam? Study this list of data to determine whether or not you think there is a relationship.

How do we get information from the list of numbers in Figure 2? Did you actually read all of the data, or did you look for your state and then skip to this paragraph? Presented as just a table of numbers, the data are difficult to interpret. In analyzing data, we search for information hidden in the numbers. It should be obvious that we need to have some way of organizing and simplifying the data so that we can see its essential characteristics without getting lost in a jumble of numbers. Then we can decide for ourselves whether or not average ACT scores are related to the percentage of students taking the exam.

On the scatterplot shown in Figure 3, average Composite ACT scores are plotted on the vertical axis and the percentages of graduates taking the exam from each state are plotted on the horizontal axis. For example, the ordered pair representing North Carolina, $(100, 19.1)$, implies that all graduates in 2016 living in North Carolina took the ACT and their average composite score was 19.1, while the ordered pair for Massachusetts is $(28, 24.8)$, indicating that 28 of 2016 graduates took the ACT exam with an average composite score of 24.8.

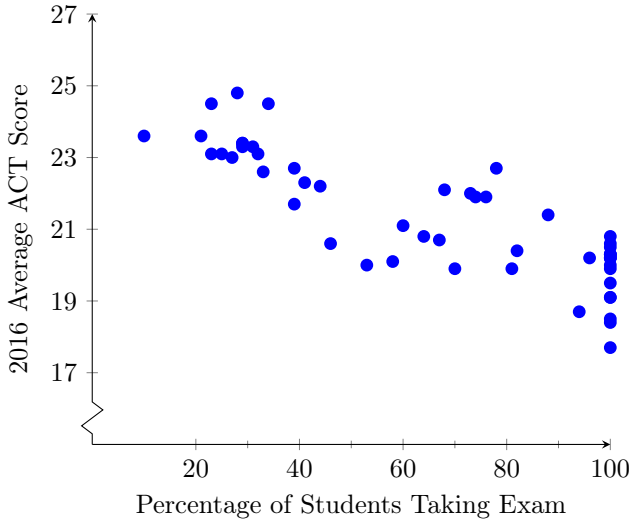

Figure 1.4.3: Average ACT scores versus Percentage of graduates taking the ACT in 2016

Looking at the scatterplot in Figure 3 should convince us that there is a relationship between these variables. The points tend to slope downward to the right, so we observe that the states with higher percentages of graduates taking the ACT have generally lower average ACT scores. When one variable decreases as the other increases, we say there is a negative association between the two variables. Notice also that the states with lower percentages of students taking the ACT have higher average composite scores. Can you give a plausible explanation for this? Which display do you find easier to interpret, the table or the graph?

A scatterplot is an effective tool for analyzing data. Special characteristics of data that may go unnoticed in a table are more obvious from a graph. If there is some relationship between the variables, a pattern or trend is usually apparent in the scatterplot. In this case, the scatterplot shows some spread in the data. Though the variables are related, we would have to consider the relationship somewhat loose, or weak, in the sense that knowing a value of one variable does not necessarily give us confidence in predicting a value for the other variable.

Notice that the data points appear to cluster with two groups with a gap between them.

In about half of the states less than 45 of the high school graduates took the ACT in 2016 with an average around 23. In the other half more than 60 took the ACT with an average around 21.

Example 1.4.4 The Leaning Tower of Pisa. The Leaning Tower of Pisa is a famous tower in Pisa, Italy. The tower was built on soft ground, and over time, it began to lean as one end sank into the soil. The amount that the tower leans is measured by comparing a point on the tower to where it would be if the tower were straight (see Figure 4.). Some measurements of the amount of the lean are shown in the data and scatterplot in Figure 3.3. We want to determine the relationship between the year and the amount the tower leans. You should note something striking in the data. The tower was closed to the public in 1990 and reopened in 2001 after the ground on which the tower had be built was stabilized and the tower tilted back to a more perpendicular position. It is often the case that we see need to consider sections of the data separately, since they may represent two different phenomena. In this case, from 1975 to 1987, we have the tower leaning more each year, but after it was closed and the tower secured, the amount of lean has remained constant (and should remain constant for the next century). So, it is only when the tower was leaning and in danger of falling that we are concerned about. Since, in this interval, the amount the tower leans increases every year, *lean*, measured in millimeters, is the dependent variable and the *year* is the independent variable.

Year	Lean (mm)	Year	Lean (mm)
1975	2964.2	1986	2974.2
1976	2964.4	1987	2975.7
1977	2965.6	2001	2144.7
1978	2966.7	2002	2144.7
1979	2967.3	2003	2144.7
1980	2968.8	2004	2144.7
1981	2969.6	2005	2144.7
1982	2969.8	2006	2144.7
1983	2971.3	2007	2144.7
1984	2971.7	2008	2144.7
1985	2972.5	2009	2144.7

Figure 1.4.5: Data from G. Geri and B. Palla, "Considerazioni sulle piu recenti osservasoioni ottiche alla Torre Pendente di Pisa"

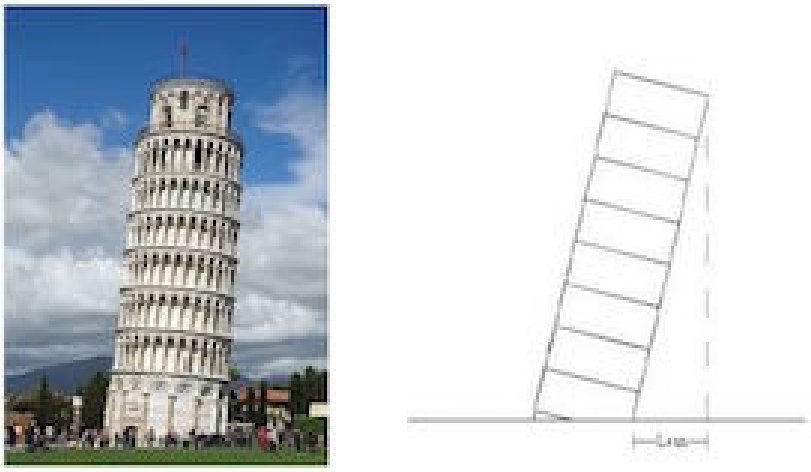

Figure 1.4.6: The Leaning Tower of Pisa and the measurement of the amount of lean

Looking at the scatterplot in Figure 4. should convince us that there is a tight or strong relationship between these variables. Notice that we are only interested in the interval before the tower was stabilized. The points tend to slope upward to the right, so we say in this case there is a positive association between the variables. That is, both variables increase together. There does not appear to be any obvious consistent curvature; rather, the points seem to be increasing fairly steadily, so we conclude that the shape is linear.

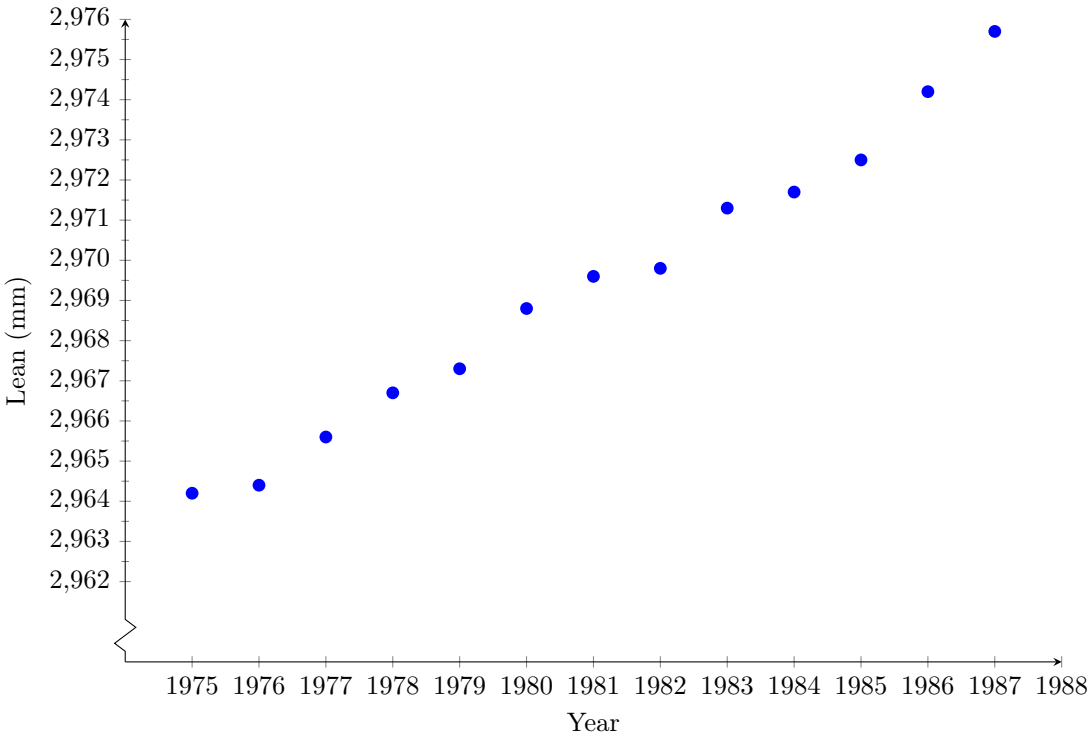

Figure 1.4.7: Lean versus Year

Sometimes when we examine scatterplots we observe points that appear to stand out from the rest. These points may follow the general pattern of the data but are far removed from other points. Other times there are points that are inconsistent with the general trend. Such points may indicate errors in measurement or in plotting that need to be corrected, or they may indicate the presence of some factor that deserves special attention. Whatever the cause, we should look for, and attempt to explain, odd points, called *outliers*, that do not appear to fit the general pattern of the scatterplot.

Exercises

1. Comment on the important characteristics of the scatterplots provided below. The legends describe the variables on each axis. Consider the shape (linear or curved), whether the data describes an increasing or decreasing (positive or negative) relationship, any gaps, clusters, or outliers apparent in the data. Write a sentence or two to explain the story the data telling about the relationship between the two variables.

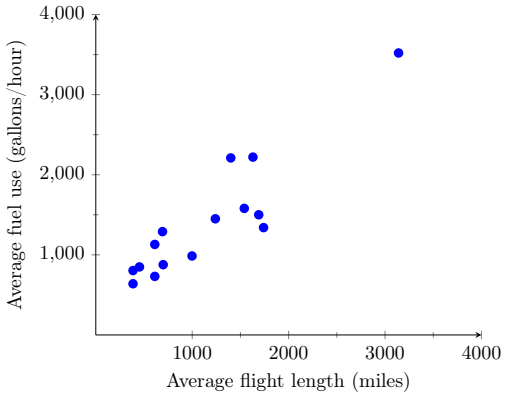

Figure 1.4.8: Plot 1

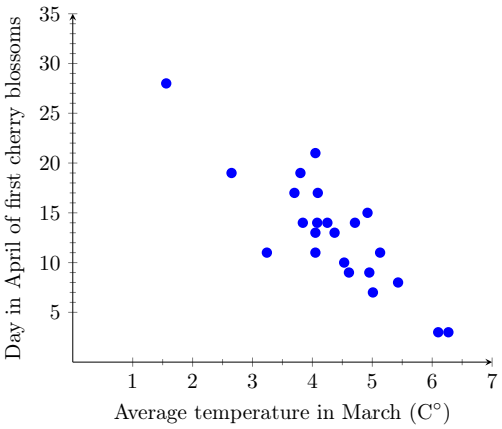

Figure 1.4.9: Plot 2

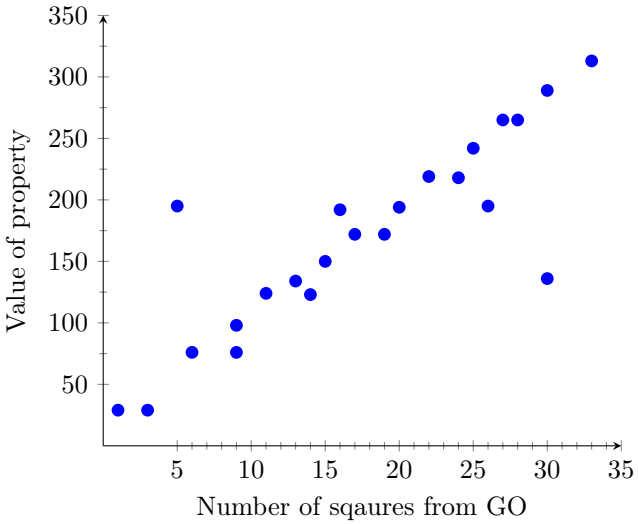

Figure 1.4.10: Plot 3

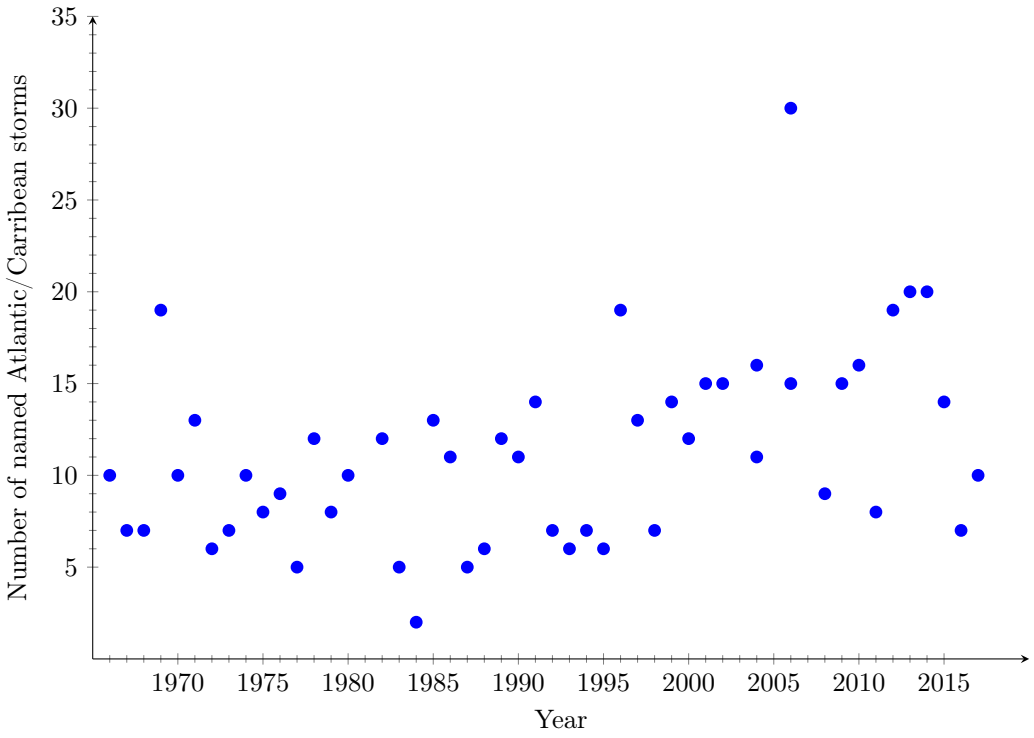

Figure 1.4.11: Plot 4

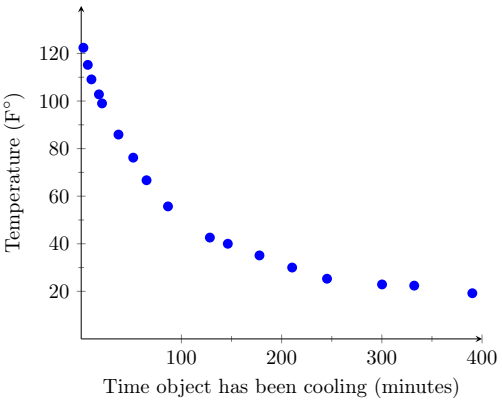

Figure 1.4.12: Plot 5

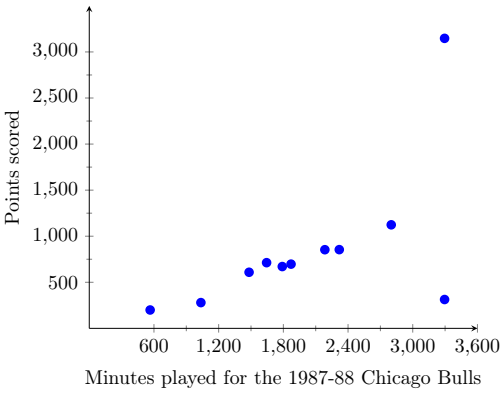

Figure 1.4.13: Plot 6

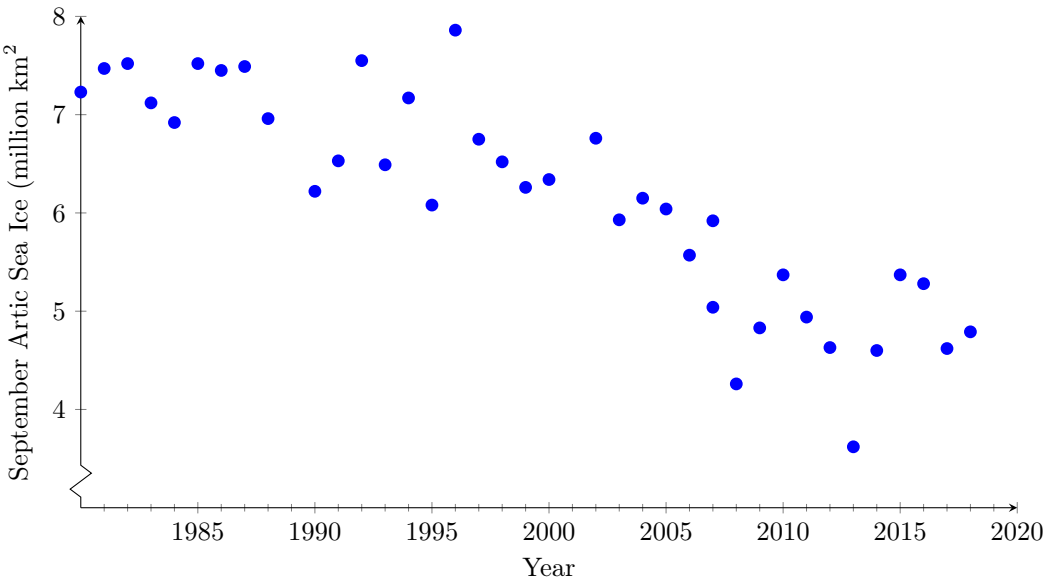

Figure 1.4.14: Plot 7

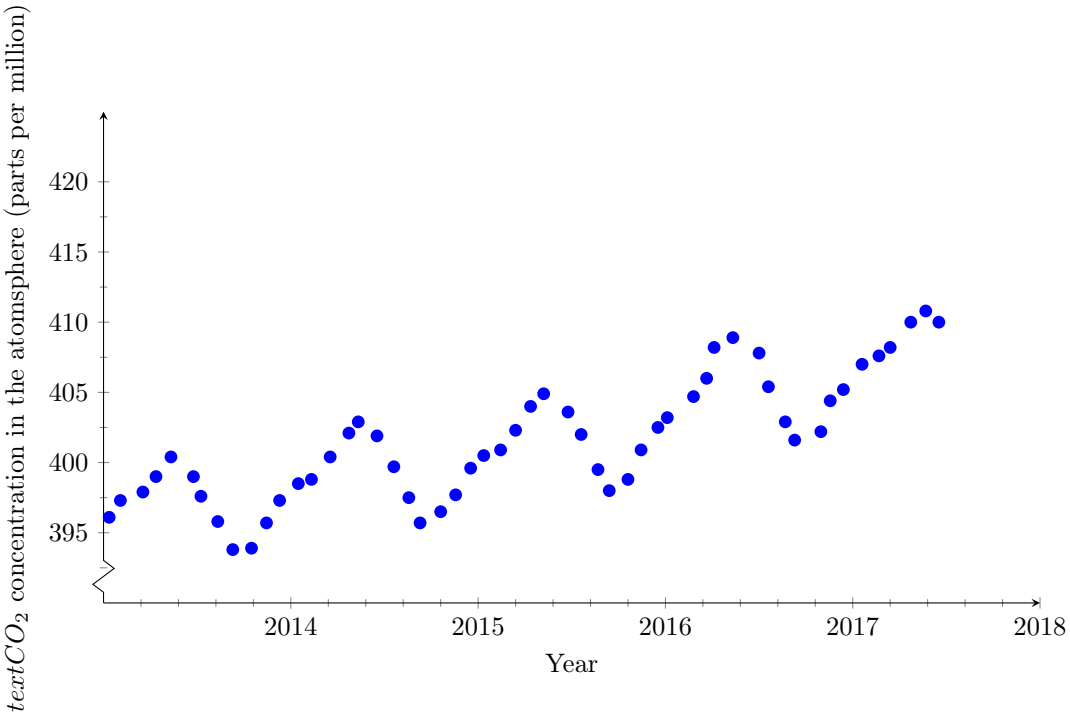

Figure 1.4.15: Plot 8

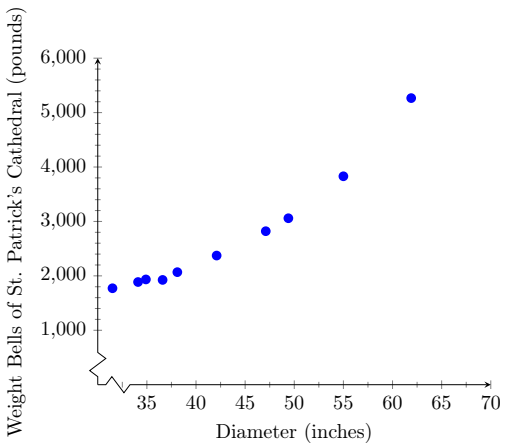

Figure 1.4.16: Plot 9

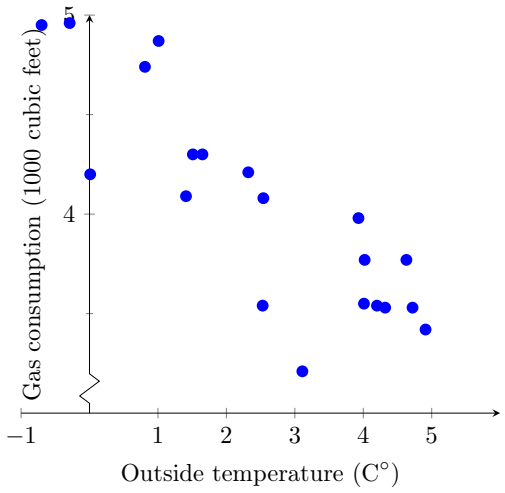

Figure 1.4.17: Plot 10

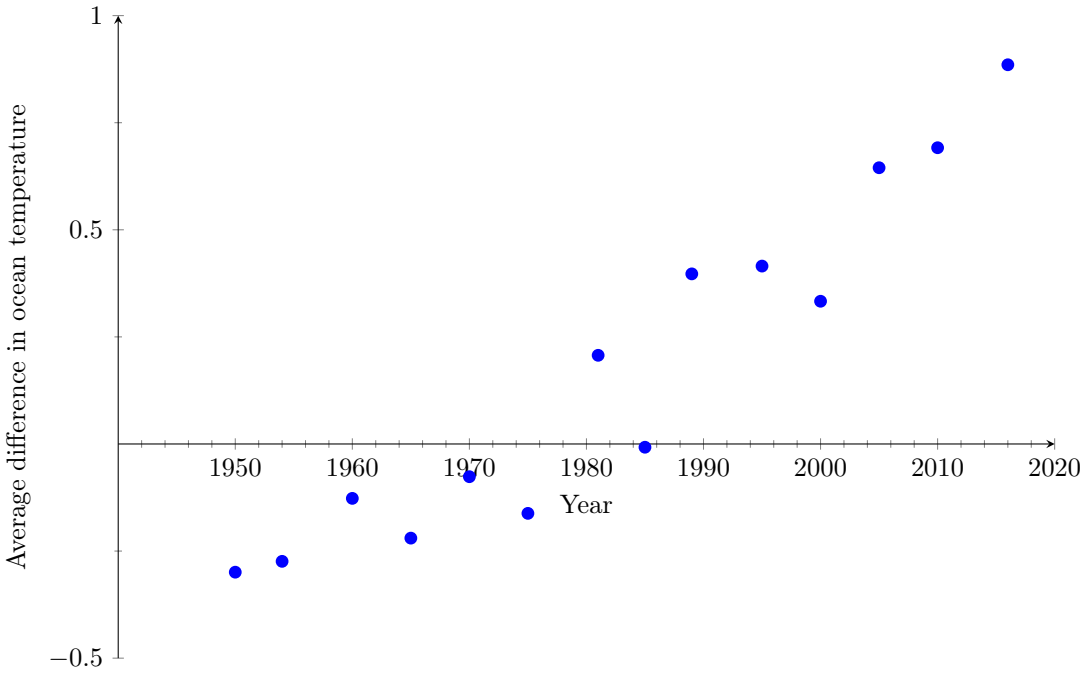

Figure 1.4.18: Plot 11

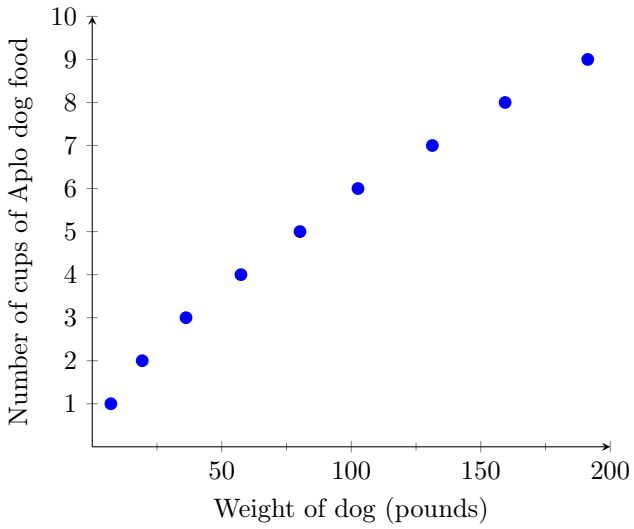

Figure 1.4.19: Plot 12

1.5 Linear Models

When you examined the scatterplots in the exercises in the last section, you should have noticed several graphs with points that stand apart from the others. In Figure 8, the point in the upper right corner represents an airplane with an exceptionally high number of seats and exceptionally high fuel consumption (Boing 747). Cover this points with your finger and use the other points to predict fuel consumption for a plane with 400 seats. Do you consider this point to follow the general trend of the other data points? In Figure 9, there is a point in the upper left corner that fits the general trend of the data but is noticeably removed from the other points. This point represents a year in which the average March temperature was unusually low and the cherry blossoms did not appear until late in April. These points differ from the outlier found in Figure 13. The outlier represents Michael Jordan, the legendary player who led the Chicago Bulls to six NBA Championships. Jordan had the highest playing time but his playing time is not inconsistent with that of the other players. This point is an outlier because of the very high number of points scored. Michael Jordan does not have statistics that follow the general pattern of the other players on the Bulls.

When a relationship is suggested by a scatterplot, we usually want to describe it mathematically by finding an equation that summarizes the way the two variables are related. Such an equation is another example of a mathematical model. When we discussed mathematical models at the beginning of the chapter, we pointed out that a good model simplifies the phenomenon it represents and gives us the ability to predict. If we can find the equation of a curve that closely "fits" a scatterplot, we can focus on the important characteristics of the relationship between the variables without the clutter of a scatterplot. We can also use this equation to predict the values of one variable for specific values of the other variable. Sometimes we use the model to *interpolate*, or estimate new values among data values and sometimes we use the model to *extrapolate*, or predict values outside the region of the data. To extrapolate, we must have good reason to believe that the pattern seen in the data continues.

To obtain an equation to model the Leaning Tower of Pisa data during the period in which the tower was leaning, you could sketch a line that passes through the data and follows the general path of the data. What is the equation of the line that best fits this data? The process of fitting a linear model to a set of data is an important aspect of data analysis. With the help of graphing calculators or computers, we can quickly fit a line to a given set of data. For the moment we will just estimate the location of the linear model to demonstrate how you can use this line. In Figure 1 we show a line through the data of our scatterplot.

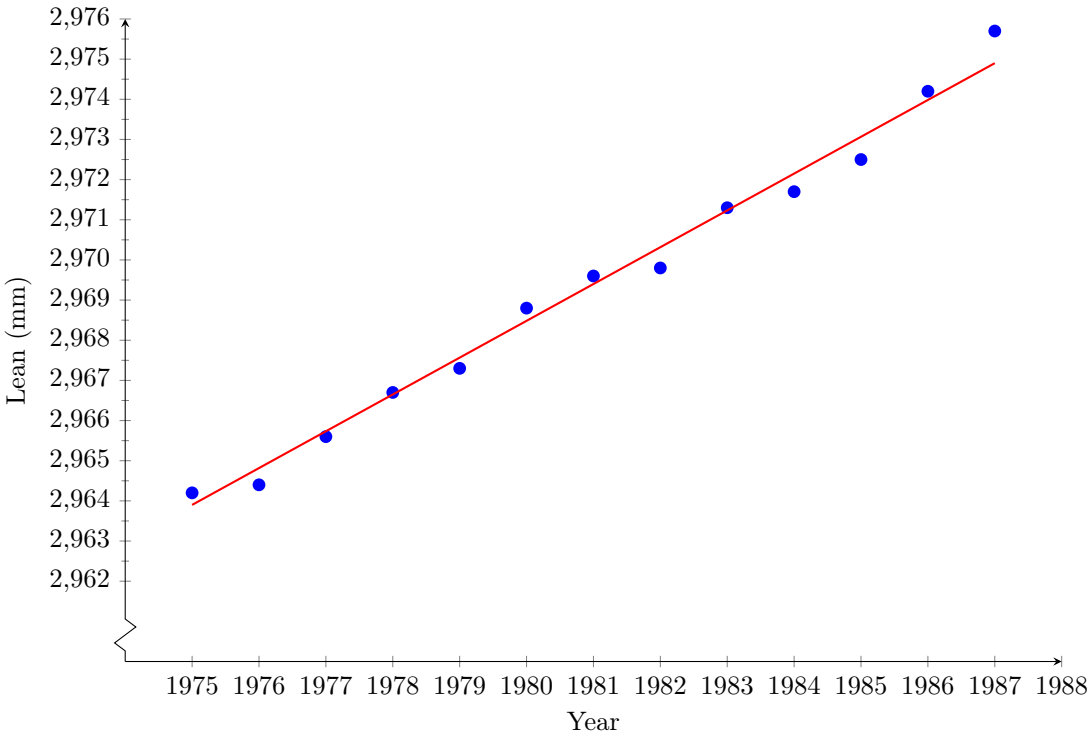

Figure 1.5.1: Leaning Tower of Pisa with linear model

Notice how the line follows the pattern of the points in this scatterplot. Some of the points are above the line, some are below, but they are all close to the line. Since the data points are closely following the path of our line, we can conclude that the relationship between these variables is strong and can feel very confident that our line does a good job of describing this particular phenomenon. We expect the same trend to continue into the near future, so we would also feel confident in using our model to extrapolate, or to predict the value of the lean in future years.

How close do the points have to be to consider the model good? Think back to Example 1 and re-examine the scatterplot in Figure 7. Try to sketch a line that follows the path of this data. Figure 1 shows one possible line that could be used to model from the data of Francis Galton, one of the founders of modern-day statistics. In the late 1800's, Galton compared the height, in inches, of 952 fathers with the height at maturity of his firstborn son. It was this study and others like it that led Galton to develop the method of linear regression and to define the standard deviation as a measure of spread. You will notice that most of the data points are not close to the line. This does not mean that a linear model is inappropriate. There is no evidence at all of any curvature in the data, so a linear equation is indeed appropriate. There is a great deal of variation in the heights of the sons, and so our model needs to acknowledge this in some way. The issue of closeness is relative and depends on the particular variables and the size of their values.

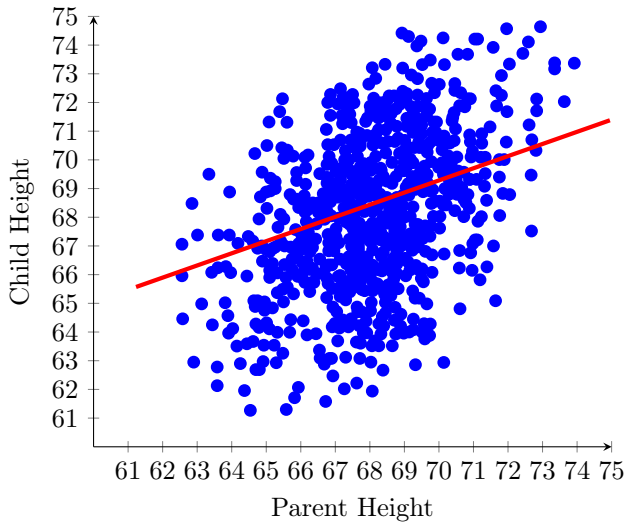

Figure 1.5.2: Galton's scatterplot with a linear model

Exercises

1. In Iowa City, Iowa, the monthly utility bill provides the customer with information about the daily cost of gas and electricity as well as the average temperature during the month. The following information has been taken from a household in Iowa City for the months of August through September. (Source: Kathleen M. Heid, Algebra in a Technological World, Curriculum and Evaluation Standards for School Mathematics Addenda Series, NCTM, Reston, VA, 1995.)

Month	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
Avg. Temp.	70	69	58	44	31	23	27	27
Avg. Daily Cost (gas)	0.35	0.38	0.78	1.41	1.86	1.94	1.97	1.76
Avg. Daily Cost (electricity)	0.98	0.78	0.82	0.77	0.86	0.65	0.80	0.73

Figure 1.5.3: Heating Bill Data

(a) Make two scatterplots of the data. One scatterplot should show the average cost of gas as the dependent variable and the average temperature as the independent variable. The other scatterplot should show average cost of electricity as the dependent variable and the average temperature as the independent variable.

(b) Describe the relationship between each pair of variables.

(c) Sketch a free-hand line through each set of data and find the equation of each line. Use the equations to estimate the gas and the electric bill if the average temperature for this February was 19 degrees.

(d) How confident are you in the predictions you made in part c? Explain your answer.

2. Comment on the important characteristics of the scatterplots provided below. The legends describe the variables on each axis. Consider the shape (linear or curved), whether the data describes an increasing or decreasing (positive or negative) relationship, any gaps, clusters, or outliers apparent in the data. Write a sentence or two to explain the story the data telling about the relationship between the two variables.

(a) The scatterplot below shows the length (in cm) of a pendulum and the period. Period is the time in seconds it takes to complete one oscillation, returning to the starting position. The ordered pairs are (length, time).

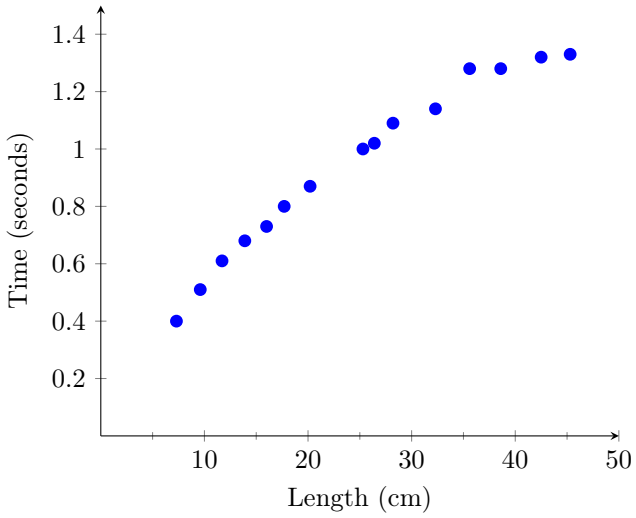

Figure 1.5.4

(b) The scatterplot below shows the horsepower of the engine for a variety of different cars and the number of miles per gallon an owner might expect to get driving. Source: StatCrunch.

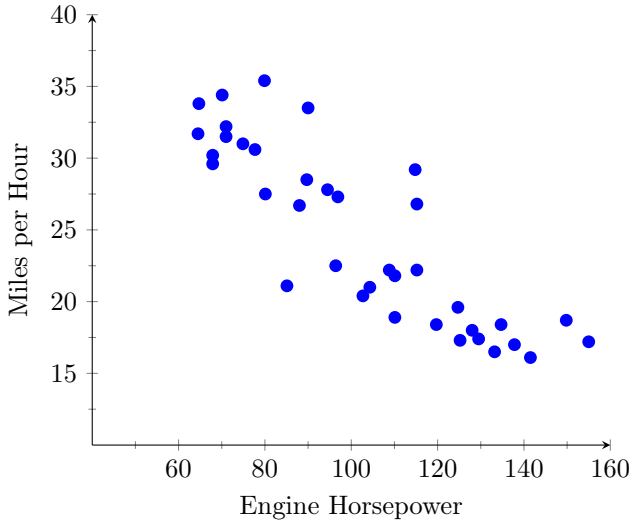

Figure 1.5.5

(c) The scatterplots below show the number of touchdowns and interceptions thrown vs the quarterback ratings for NFL quarterbacks during the 2016 season. Source: *The World Almanac and Book of Facts 2017.*

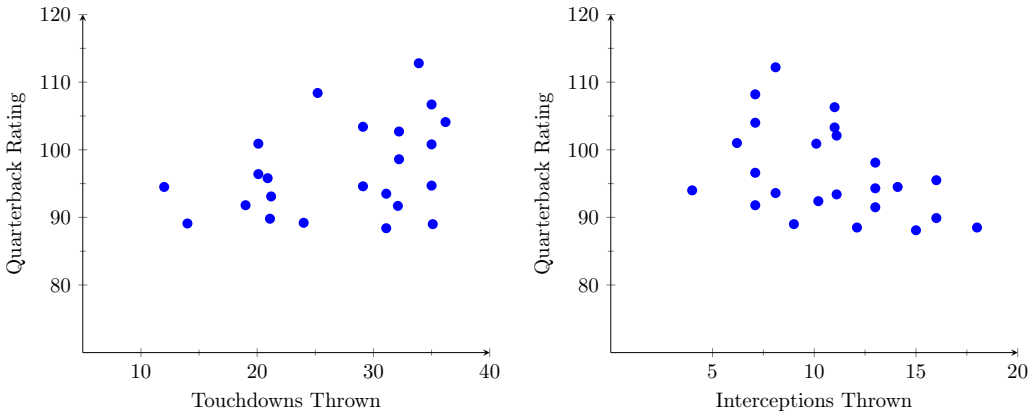

Figure 1.5.6

3. Characteristics of different Boeing aircraft flying in the US are given below.

(a) From the descriptions in the table, try to predict which variables have a linear relationship. Check two of your predictions by creating scatterplots of the data. Were your predictions reasonable?

(b) New data from old: Compute a data set that represents:

i. the average times of a flight by computing $\left(\frac{\text{Length}\,(m)}{\text{Speed}\left(\frac{m}{hr}\right)}\right)$

ii. the average cost per mile by computing $\left(\frac{\text{Cost}\left(\frac{\$}{hr}\right)}{\text{Speed}\left(\frac{m}{hr}\right)}\right)$

iii. the average cost per passenger by computing $\left(\frac{\text{Length}\,(m)}{\text{Speed}\left(\frac{m}{hr}\right)} \cdot \frac{\text{Cost}\left(\frac{\$}{hr}\right)}{\text{NumberofSeats}}\right)$

iv. the average cost per passenger-mile by computing $\left(\frac{\text{Cost}\left(\frac{\$}{hr}\right)}{\text{Speed}\left(\frac{m}{hr}\right) \cdot \text{NumberofSeats}}\right)$

(c) Is the relationship between Time of Flight and Cost per Passenger-Mile linear?

1.6 The Principles of Linear Regression – the Least Squares Line

1.7 How Good Is Our Fit?

1.8 Residual Analysis

1.8.1

When looking at a residual plot, what would you like to see? What makes a "good" residual plot?

A good residual plot shows a random scatter of residuals. The equation that we use for our model (currently we are only using linear equations) captures important information

about the situation or phenomenon being modeled. The equation tells us what we know about the setting. Statisticians often describe this as the *signal* given by the data. The residuals give us information about what we don't know about the setting. Statisticians often describe this as the noise given by the data. Our observations can be partitioned into signal and noise, so

$$Observations = Signal + Noise$$

When creating a model for some real-world situation, it is important to give as much information as you can about both the signal (what you know) and the noise (which puts bounds on what you don't know). A random scatter of residuals tells us that our model captures the important information, but no model can capture the random variation inherent in all processes. The spread of the residuals estimates the spread of this random variation, and serves as a bound on how far from our model actual data is likely to be found.

For example, in the Pearson Father-Son data, the residual plot shows a random scatter, so we accept our linear model as appropriate. The height of the son has a linear relationship with the height of the father.

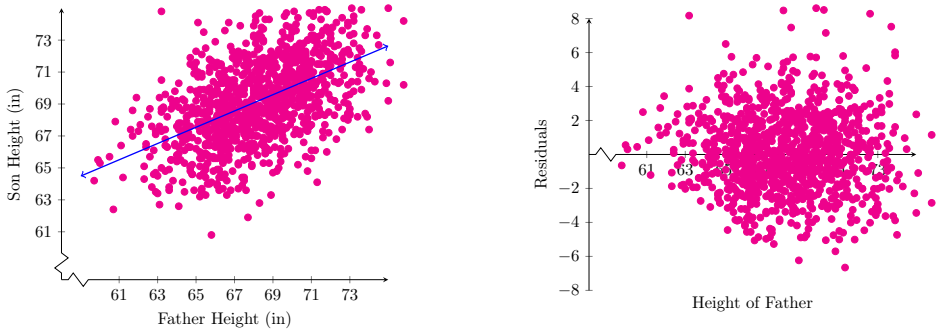

Figure 1.8.1: Pearson's Father-Son Data and Residual Plot

The equation of the line is $Son = 34.428 + 0.5095 Father$, so for each addition one inch of height for the father, the height of the son is expected to increase, on average, by about one-half inch. The residual plot just shows a scatter of points, so our linear model is appropriate. There is clearly a lot of variability in the data. In the next section we will learn how to estimate the size of this variability. For the moment, we are just interested in its shape.

We will often see patterns in the residual plots. A pattern in the residuals indicates that there is some aspect of the physical situation that is not being captured by our simple model. This does not mean our model is wrong, just that it is incomplete. For example, the CO_2 residuals clearly has a periodic pattern that describes the yearly fluctuations in CO_2 due to the seasonal growth of plants expelling CO_2 into the atmosphere. We can see from the residuals the yearly behavior introduces a change of about 6 units from winter to summer. Later in the course, we will find a model for this yearly fluctuation.

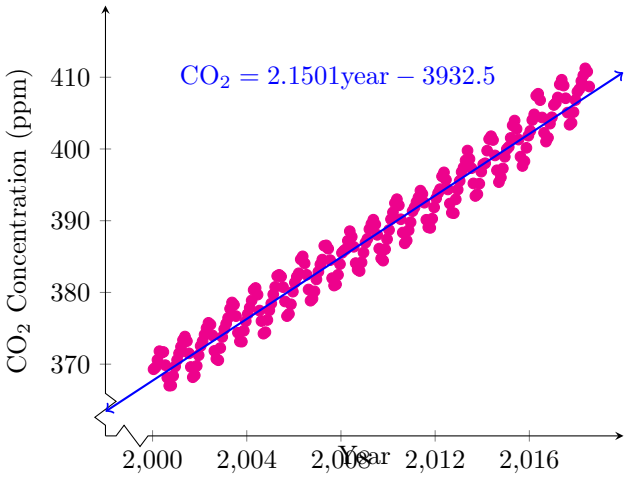

Figure 1.8.2: Residual Plot and Connected Residual Plot for CO_2 Data

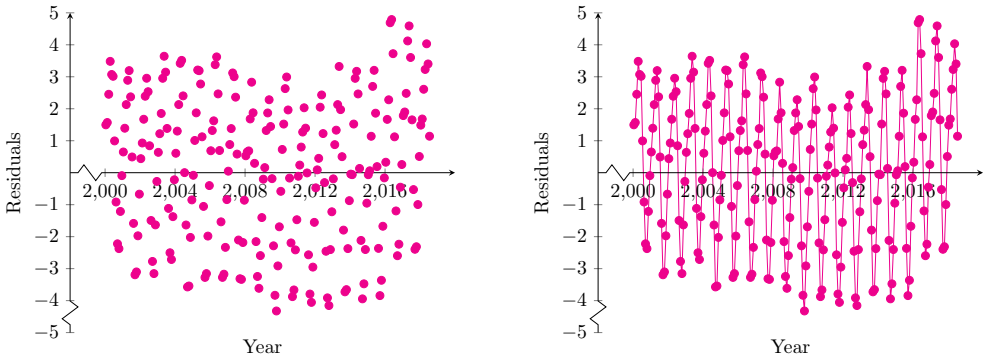

Figure 1.8.3: Residual Plot

Since the slope of the linear model is 2.15, we know that the mean yearly concentration of CO_2 in the atmosphere is increasing, on average, about 2.15 ppm each year. In any given year, this increase gets lost in the 6 to 8 ppm variation from winter to summer, but we can see the accumulated result in the 30+ ppm increase from 2000 to 2016. The periodic pattern in the residuals does not indicate that our linear model for the general increase is incorrect, only that it does not capture all of the important behavior in this phenomenon. In this case, there is a second signal to be found in the noise of our linear model.

There are many characteristic patterns that can be observed in residuals. Interpreting them all takes experience and knowledge of the different processes being modeling.

1.8.2 Residual Analysis vs. Correlation

The correlation, denoted as r, is often part of the output of a linear regression using technology. It is often misinterpreted as a measure of linearity. It is incorrect reasoning to assume that having a high value of r indicates a linear model is appropriate.

Consider the following two data sets. The first set of ordered pairs describes the diameter in inches and the weight in pounds of the bells of St. Patrick's Cathedral in Dublin, Ireland.

Diameter (in)	29.5	31.5	34.0	35.0	36.5	38.0	42.0	47.0	49.5	55.0	62.0
Weight (lb)	801	925	1050	1116	1109	1253	1638	2122	2467	3339	5091

Figure 1.8.4: The Bells of St. Patrick's Cathedral

The second set of data describes the average monthly temperature in degrees Celsius and the average amount of gas (1000 cubic feet) used to heat a house.

Temp (C)	−0.8	−0.7	0.3	2.5	2.9	3.2	3.6	3.9	4.2	4.3	5.5	6.0	6.0
Volume (1000 ft^3)	6.9	6.5	5.8	5.4	5.3	5.1	4.8	3.6	5.1	4.2	3.8	4.0	3.2
Temp (C)	6.0	6.3	6.4	7.0	7.0	7.5	7.5	7.5	7.6	8.0	8.5	9.1	10.2
Volume (1000 ft^3)	3.1	3.3	3.6	2.2	2.6	3.0	2.9	2.3	2.0	2.6	2.2	1.2	0.6

Figure 1.8.5: Outside Temperature and Gas Consumption

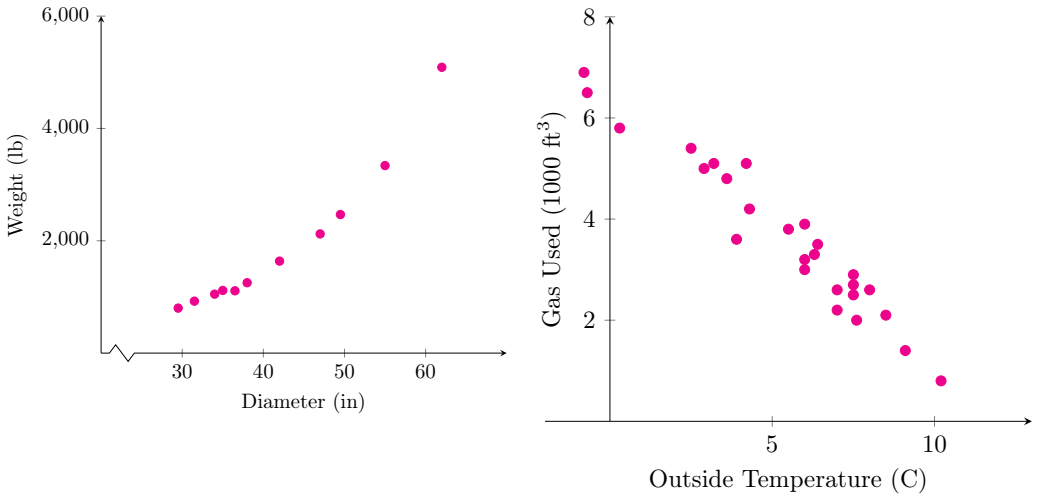

Figure 1.8.6: Graphs of the Bells of St. Patrick's Cathedral and Gas Consumption

Which of these two sets of data would best be modeled by a linear function and which do you think is best modeled by a non-linear function? It should be clear that the relationship between the data describing gas consumption could be well described by a linear function, while the relationship between the data describing the bells is clearly non-linear. In the scatterplots in Figure XX, we compare the residuals for the regression line for each of the two data sets. In particular, notice the signature U-shaped residual plot for the bells, indicating non-linearity in the data that requires a non-linear function for its model.

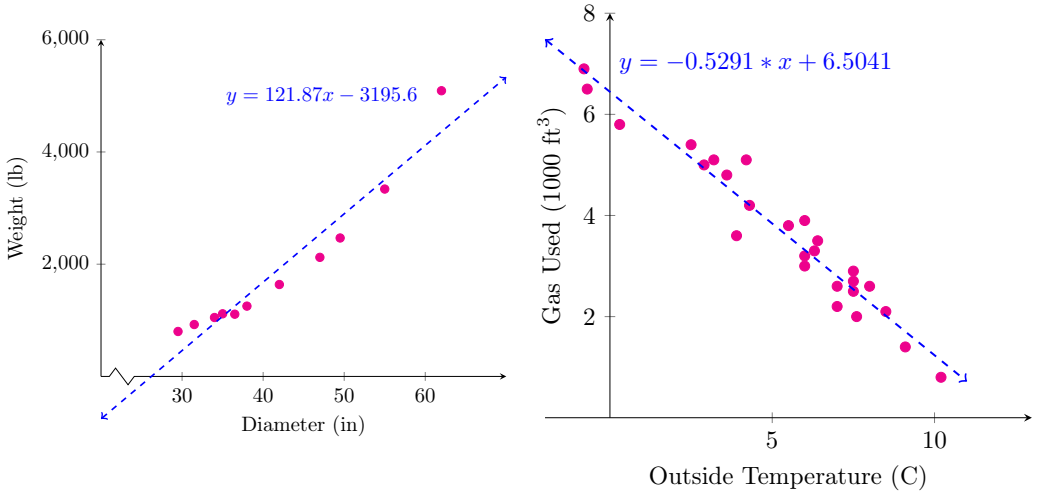

Figure 1.8.7: Linear Regression Models for Bells and Gas Consumption

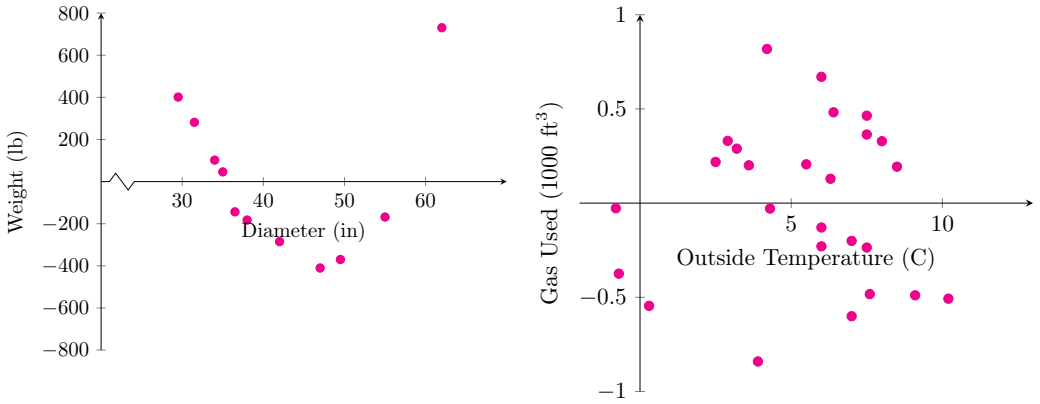

Figure 1.8.8: Residual Plots

The correlation for the data in Figure XX (the bells) is $r = 0.963$ and for Figure YY (gas consumption) is $r = -0.963$. Having a high correlation is not an indication that the data is best described by a linear model. For this reason, we always use the residual plot as the arbiter of linearity.

We look for this curvilinear pattern in the residuals to indicate some non-linearity in the data that has not been accounted for in the model we have chosen. We accept a linear model as appropriate for the relationship between Temperature and Gas Consumption, but reject a linear model as appropriate to describe the relationship between the Diameter of a bell and its Weight.

As a second example, consider the Galton father-son data and the water flowing from an urn data set. The linear models and the residual plots are shown below.

INSERT FIGURES HERE

It should be clear from the residual plots that for the Galton data, a linear equation is appropriate and for the urn data, a linear model fails. The curved shape of the data becomes apparent as the characteristic U-shaped residual plot indicates curvature in the data. So, some non-linear function should be used instead of a line (we will determine the

appropriate model in a later chapter).

If we compute the correlation for the urn data, we find that r=-0.9935. This is only 65-thousanths from a perfect negative correlation, but a linear model is not appropriate. Compare this to the correlation of the Galton data set, which is modeled by a line. For Galton, r=0.4209. So, what good is a "measure of linearity" that varies from 0 to 1 in absolute value, with the closer to 1 being "more linear" if a correlation of 0.4209 is linear but 0.9935 is not? The moral here is that we cannot use the correlation r as a means to determine if a set of data is best modeled by a line. That's a job for the residuals! Correlation plays a different role in regression analysis.

1.9 Standard Deviation

We see in FIGURE XX in Section 3 that $(2.5, 2.5)$ is the point on the line $y = x$ closest to the point $(1, 4)$. A natural question to ask is, "what is that shortest distance?" To answer this question, we again use the distance formula. This shortest distance found by the least squares criterion is

$$d = \sqrt{(1 - 2.5)^2 + (4 - 2.5)^2} = \sqrt{4.5} \approx 2.121$$

This distance between the mean point and the data point is a measure of how close the two points are, and since the mean point using all the same values for each coordinate, it also measures how close to the mean the individual components of the data are. That is, it contributes information about the spread, or variability, of the data. We saw that the mean is a least-squares measure of the "center" of data, and we interpret the value of the mean as the size of a "typical" value. To describe the spread or variability of the data, we continue with a the least-squares (distance formula) approach and use the standard deviation, s, as our measure of spread.

The computational formula for the standard deviation is $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$. We will not be using this equation to make computations, all of our computations will be done automatically on the calculator or computer, but the equation gives us important information about what the standard deviation is measuring. Notice that the numerator, $\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, is the length of the line segment from the point defined by the data to the point $(\bar{x}, \bar{x}, \bar{x}, ..., \bar{x},)$.

We saw that for the data $\{2, 3, 7\}$ the mean of 4 minimizes the distance between the points $(2, 3, 7)$ and the line $x = y = z$. That is, the point $(4, 4, 4)$ is the closest to $(2, 3, 7)$. The distance between $(4, 4, 4)$ and $(2, 3, 7)$ is $d_1 = \sqrt{(2-4)^2 + (3-4)^2 + (7-4)^2} = \sqrt{14} \approx 3.74$.

Compare the data $\{2, 3, 7\}$ and $\{1, 3, 8\}$ to $\{2, 3, 7\}$. All three sets have means of 4, so they would all be represented by $(4, 4, 4)$. However, $(4, 4, 4)$ is farther from $(1, 3, 8)$ and closer to $(3, 3, 6)$ than it is to $(2, 3, 7)$. We see that $d_2 = \sqrt{(1-4)^2 + (3-4)^2 + (8-4)^2} = \sqrt{26} \approx 5.099$ while $d_3 = \sqrt{(3-4)^2 + (3-4)^2 + (6-4)^2} = \sqrt{6} \approx 2.45$.

Since $(3, 3, 6)$ is closer to $(4, 4, 4)$, than $(2, 3, 7)$ we also know that $\{3, 3, 6\}$ is less variable than $\{2, 3, 7\}$; the point are closer to $(4, 4, 4)$ because the coordinates are closer together. In general, the length of the segment between (x_1, x_2, x_3) and $(\bar{x}, \bar{x}, \bar{x})$ is a "natural" measure of how spread out is the data, and this length is the numerator in the expression for the standard deviation of the data set $\{x_1, x_2, x_3\}$.

But what about the denominator $\sqrt{n-1}$ in the definition of the standard deviation? What's it doing there? While the length of the segment between (x_1, x_2, x_3) and $(\bar{x}, \bar{x}, \bar{x})$ is a good way to compare sets where both sets have 3 data points, it suffers when comparing sets with an unequal number of data points. Remember, the distance formula sums squares, which are non-negative numbers. The more numbers being added, the larger the sum is expected to be. So, knowing the distance is, for example 50 units, doesn't tell us all we need to know. If three squares add to 50, they should be fairly large, but if 40 squares add to 50, they are each likely quite small. The standard deviation tries to estimate the typical size of a summand in the distance formula.

We saw that the mean of $\{2, 3, 7\}$ was 4. Now, consider the data set $\{2, 3, 4, 4, 4, 4, 4, 4, 4, 7\}$. This set also has a mean of 4, and the distances between the points $(2, 3, 7)$ and $(4, 4, 4)$ and the points $(2, 3, 4, 4, 4, 4, 4, 4, 4, 7)$ and $(4, 4, 4, 4, 4, 4, 4, 4, 4, 4)$ are both $\sqrt{14}$. But the

individual components making up the sum of squares are, on average, much smaller for the larger data set, so we consider $\{2, 3, 4, 4, 4, 4, 4, 4, 4, 7\}$ to be less variable (or more consistent). In one case we compare 3 values and get a total length of $\sqrt{14}$, while in the other case, we have 10 values and get a total length of $\sqrt{14}$. By dividing by $\sqrt{n-1}$, we take into account the number of data values used in the sums of squares. Just why we divide by $\sqrt{n-1}$ instead of $\sqrt{n}$ or some other function of n requires some knowledge of statistics and linear algebra, which are well beyond this course. But, rest assured, there are good reasons for this and the applications of the standard deviation abound in higher mathematics.

Exercises

1. In Section XX (1.3) we considered the average ACT scores by state for the 50 states and the District of Columbia. The mean for this data set is $\bar{x} = 21.32$ and the standard deviation is $s = 1.75$. In the dotplot below, discuss the proportion of scores observed within one standard deviation of the mean and within two standard deviations of the mean.

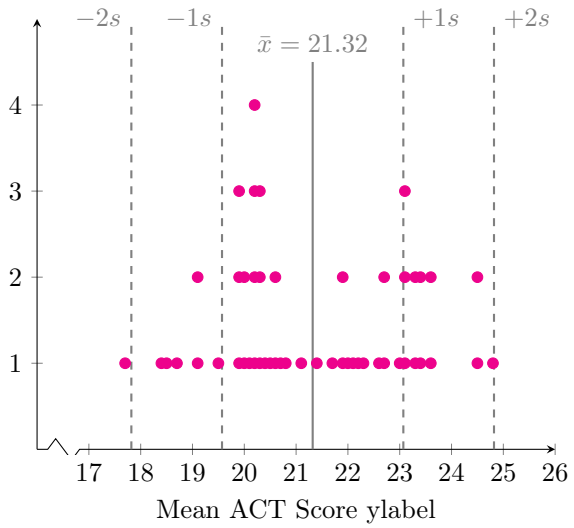

Figure 1.9.1: Mean ACT Scores by State / District

2. The average temperature in Tokyo during the month of March in degrees Celcius was collected over a 24-year period. The mean temperature was 4.32 degrees and the standard deviation was 1.02 degrees. How well does the mean measure the center and spread of the data?

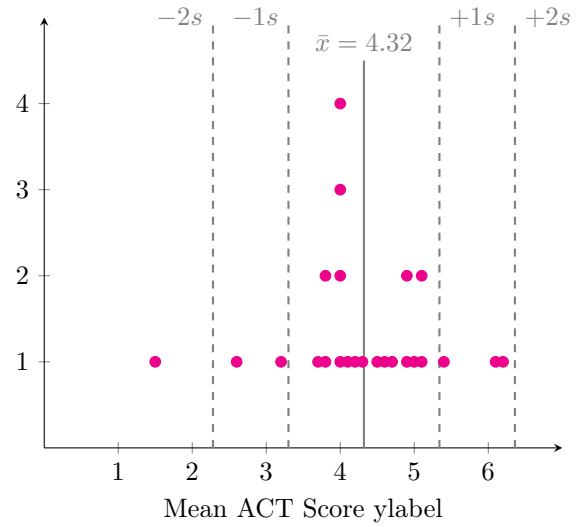

Figure 1.9.2: Average Temperature in Tokyo in March

3. The Volume of Artic Ice is estimated every year. The dotplot below shows the estimated values for 17 recent years. The mean is 5.37 and the standard deviation is 0.83

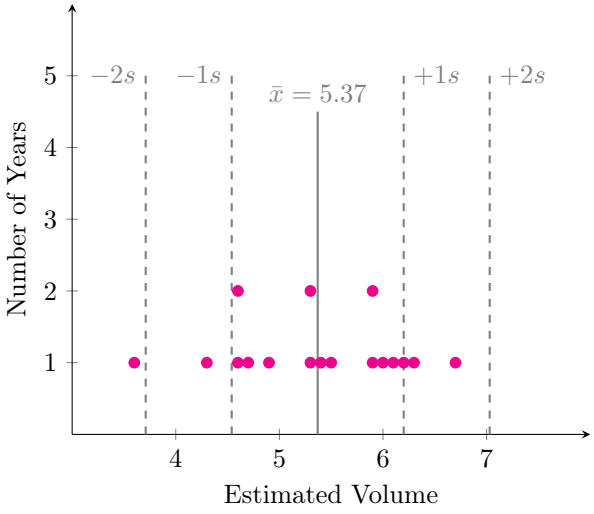

Figure 1.9.3: Volume of Artice Ice

If we pair the Arctic Ice Volume with the year the measurement was taken, we get a different impression. Describe how these two graphs tell a similar story and how they tell a different story.pg

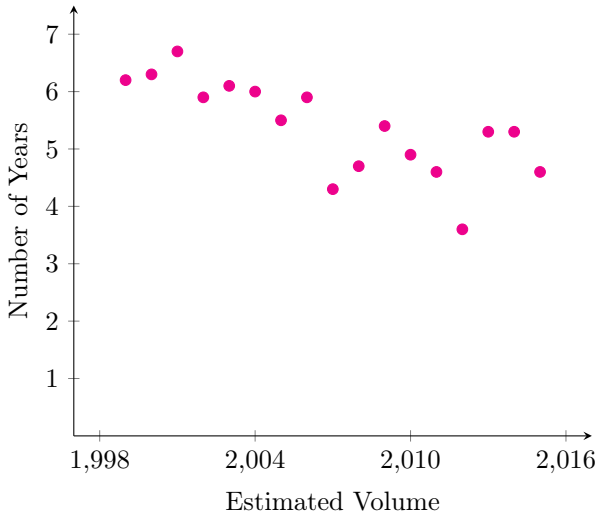

Figure 1.9.4: Volume of Artice Ice

1.10 Error Bounds and the Accuracy of a Prediction

Consider Hiro's interest in predicting the date on which the first cherry blossoms appear from SECTION XX. Using a larger set of data than the simple example used earlier, we can find the least squares line relating the average temperature in March and the number of days in April before the cherry blossoms appear is

$$f(x) = -4.76x + 33.51$$

If the average temperature in March this year was 3.5 °C, Hiro expects the blossoms to appear on a date close to the 17th if he uses the least squares line as his model, since $f(3.5) = 16.85$. But, remember the models we create from data using regression capture the important features of the process being considered, but cannot give exact predictions. Taken literally, 16.85 would be at 8:24 pm on April 16. Clearly, we do not believe our model could possibly be this precise. In fact, given the obvious variability, at best we can say that we expect to see the blossoms appear somewhere *around* the 17th of April. Maybe within a day of April 17th if we're feeling confident, or within the week of April 17th if we feel less confident. Notice that there were four years in which the average temperature was 4 °C, and the cherry blossoms appeared on the 14th, 21st, 13th, and 11th of April. On average, around the 15th (14.75) and our model precicts 14.47. The goal of the least-squares regression line is to estimate the *average* y-value for any given x-value. This goal acknowledges the inherent variability in all real-world processes.

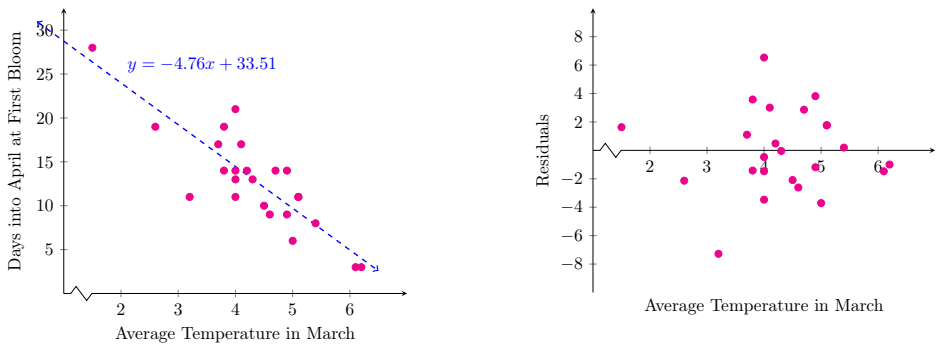

Figure 1.10.1: Least-squares line with residuals indicated

The regression equation gives us the signal coming from the relationship between temperature and the date of blooming. What about the noise? We know that the value given by the regression equation is unlikely to be correct, the actual date is a result of signal plus the random variation about the signal ever-present in the world that we call the noise. So we need an estimate of the size of the noise in the model.

Since the residuals provide information about how the data varies from the model, we can use the residuals to determine a range of plausible values from the model. Looking at the residual plot in Figure 1.10.1, we see that all but two of the residuals have magnitude less than 4. So it is quite likely that the blossoms this year will occur between April 13 and April 21, that is 17 ± 4. This interval is not guaranteed, but the information from previous years makes us feel fairly confident in an interval of 4 days in either direction.

The method we have just used to determine our interval estimate is quite subjective. As you might expect, statisticians have objective ways to produce intervals associated with predictions from linear models. You will learn some of these methods if you take a statistics course, but the mathematics behind the methods are beyond what we can do in this course. Nevertheless, we can create some approximate rough-and-easy bounds. One way would be to use the value of the residual with largest magnitude, like we did above when we added and subtracted 7 from the predicted value. Another method would be to use the average value of the residuals. This sounds good, but as noted earlier in SECTION XX(3?), the average residual value will always be zero, since the positive residuals balance out the negative ones. To avoid this cancellation, we could first take the absolute value of the residuals and then calculate the mean or median. If this reminds you of the discussion in SECTION XX(8?), it should!. It is exactly the same discussion. We have a set of numbers, in this case the residuals, and we want one number to represent a typical value. The discussion ends just as it did earlier, with the least-squares criterion based on the distance formula being the choice of statisticians.

As we noted when we first discussed the standard deviation in SECTION XX(8?), the majority of values in the data set will fall within 2 standard deviations of the mean, and almost all within 3 standard deviations from the mean regardless of the shape of the data. The standard error of the estimate is a useful and simple measure of the degree of concentration of the observations around the regression line. The standard error of the estimate is most easily approximated by the standard deviation of the residuals.

If a least squares line is fit to a linear data set, then more than 75% (most often much more than 75%) of all the residual values will fall within two standard deviations of the average of the residuals. Since the average of the residuals is always zero for the least squares line, most of the data will fall within 2 standard deviations of the residuals from the values predicted by our linear model. In this example the standard deviation of the residuals is $s = 2.91$. If we add and subtract 5.82 days from the predictions associated with this model, we should have a reliable estimate of when to expect the blossoms. The standard deviation of the residuals is easy to compute since this is a built-in feature of most calculators and computer software that help us analyze data. Statistical software can

compute a more precise estimate using higher level mathematics, but our two standard deviation approximation works well and we will use it in the remainder of this course.

Once you choose a technique for calculating intervals to place on the estimates, you can produce error bounds for your model. *Error bounds* are models that predict the upper and lower bounds you expect your predictions to fall between. The model we developed for Hiros data is $f(x) = -4.76x + 33.51$. If we decide to determine an interval for our predictions by adding and subtracting twice the standard deviation, 5.82 days, we are fairly certain that the actual day the flowers will bloom falls between the two linear models $f(x) = (-4.76x + 33.51) - 5.82$ and $f(x) = (-4.76x + 33.51) + 5.82$. These equations simplify to $f(x) = -4.76x + 27.69$ and $f(x) = -4.76x + 39.33$. These error bounds are shown with the least squares line and data in Figure 2. All except two data points are within these bounds, so they appear to do a good job of capturing the variation in the original data.

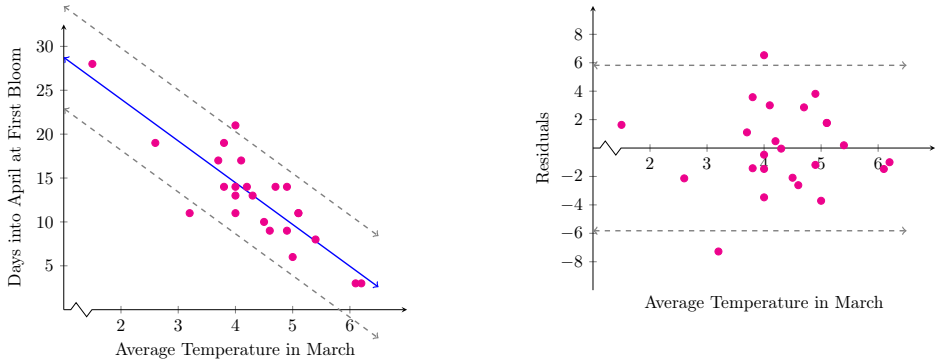

Figure 1.10.2: Least-squares line with residuals indicated

By including error bounds in his model, Hiro gives additional information about the accuracy of predictions from the model. Using two times the standard deviation of the residuals, Hiro can use the average temperature in March to predict the date on which the first blossoms will appear to within approximately 6 days. His best guess is still the prediction from his least-squares line, which is the 17th of the month. But, he should be more certain in predicting that the first blossom will appear somewhere between the 11th and the 23rd of April. The way to interpret these error bounds is to say, we would not be surprised if the blossoms appear sometime between the 11th and 23rd. If someone were to tell us that the blossoms first appeared on the 25th of April, we would be surprised and ask, "Are you sure? That is not what I expected." The error bounds give the interval in which we are not surprised.

For Pearson's Father-Son data, the equation of the line is Son $= 34.428x + 0.5095$Father. For a father that is 67 inches tall, our prediction would be that his first-born son would be about 68.3 inches tall. The standard deviation of the residuals is $s = 2.434$ inches, so we would find it unsurprising to find a son whose father is 67 inches tall to be somewhere between 63.4 and 73.2 inches tall. Just where in this interval the son falls is due, naturally, to the characteristics of the mother, the level of nutrition through childhood, and many other variables not taken into account in our model.

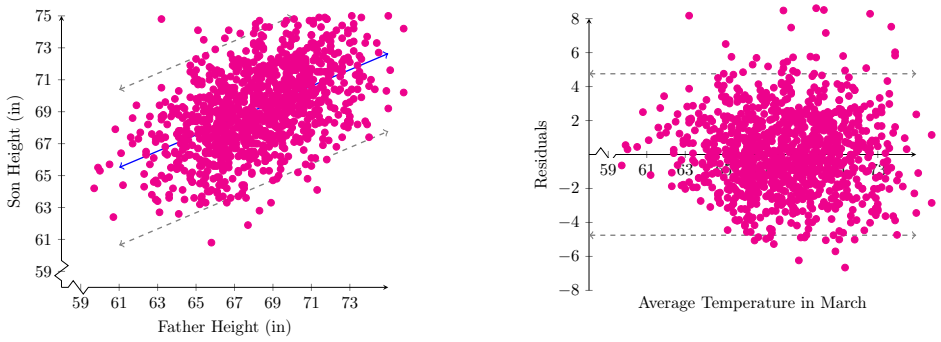

Figure 1.10.3: Least-squares line with residuals indicated

As a final example, in Figure 4, the standard error of the estimate is used to indicate with shading the region within which the majority of the observations fall. The estimated volume of Arctic Sea Ice for the years 1979 – 2016 are shown below [1] with one and two standard deviations of the residuals highlighted. The values shown in the graph are the differences in the estimated volume and the average over the time interval and the one and two standard deviation lines are shaded. Notice that almost all of the data (except for a few large deviations) fall within the two standard deviation region.

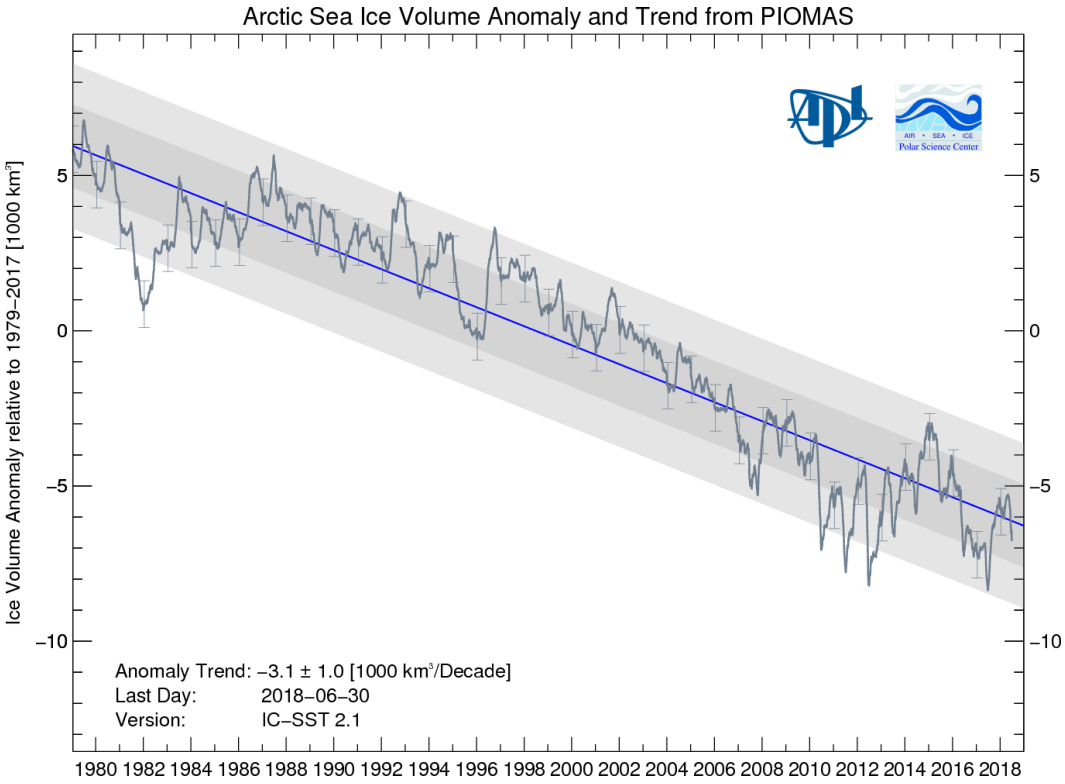

Figure 1.10.4: Lines parallel to the regression line indicate Error Bound on the Observations

Exercises

1. As we will see in CHAPTER XX(5?), the error bands can be used with non-linear models as well. Curves "parallel" to a function can be created by vertical shifts of the model

[1] Source: Polar Science Center, Applied Physics Laboratory, University of Washington

up and down two standard deviations of the residuals. In the figure below, one and two standard deviations are shaded. Explain why the bands look wider in September than in June.

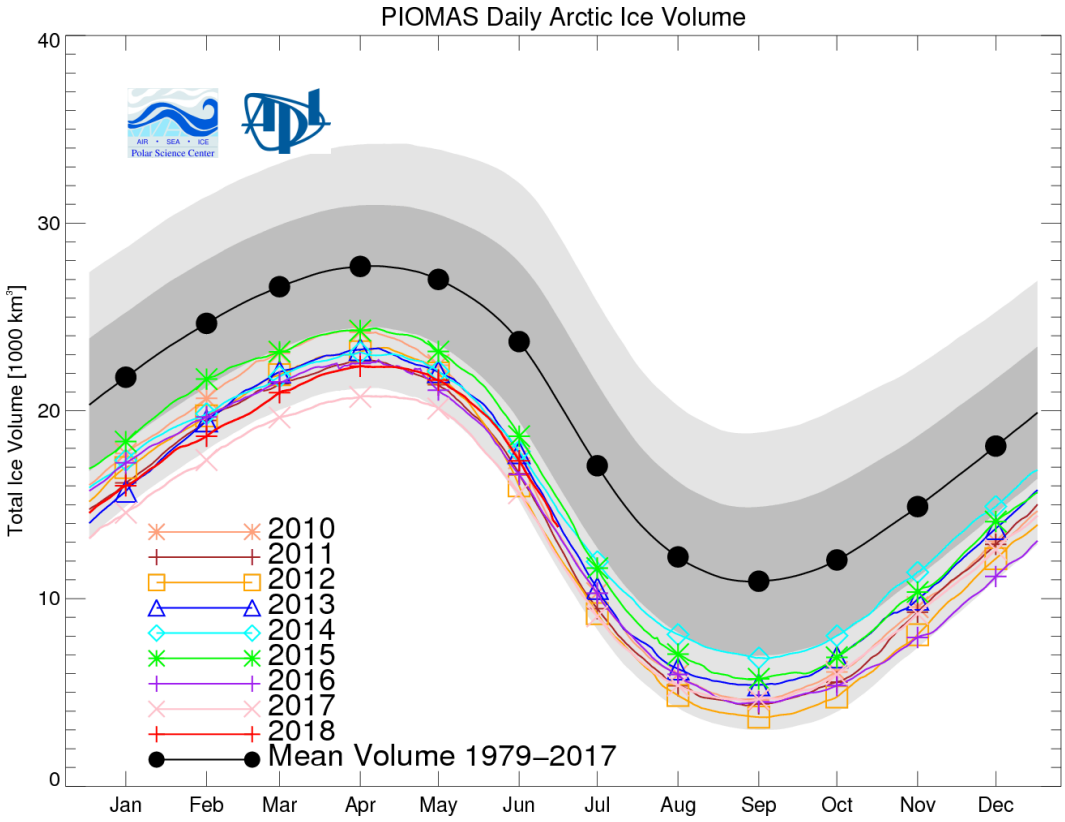

Figure 1.10.5: Yearly fluctuations in Arctic Ice Volume with 1 and 2 standard deviation curves highlighted.

2. Calculate the error bounds for the Leaning Tower of Pisa data. Explain how you determined your error bounds.

3. A lumberjack has been studying the trees on the mountain beside his house. Last summer he measured the diameter of the trees at a height of 1.2 meters above the ground. He then used a tool to measure the height of the trees. The data he collected is given below:

Diameter (cm)	16	18	19	20	21	21	23
Height (m)	12.9	13.2	12.8	15.0	15.7	14.9	14.7
Diameter (cm)	24	25	25	26	28	30	31
Height (m)	17.3	19.2	15.2	18.6	17.9	16.4	19.5

Figure 1.10.6: Tree Data

(a) Make a scatterplot of the data, then find an equation of a line to fit this data.

(b) If a tree has a diameter of 36 centimeters, what height does your equation predict? Find error bounds. How close should you expect your prediction to be?

(c) If you know the diameter of a tree, based on your model above, is it likely that your prediction of the height of the tree will be within 2 meters of the actual height?

(d) Use the linear equation that models the relationship between the diameter of the tree at 1.2 meters and the height of the tree to determine the equation of the linear model

if the circumference had been used rather than the diameter. Which do you think the lumberjack actually measured, the diameter or the circumference? Is one measurement better to use than the other? Explain your reasoning.

4. Fit a line to the data you gathered in the experiments in SECTION XX. Describe the residual plot. For those data sets that appear to be linear, find the error bounds. How confident are you in your model?

5. As we age, our vision deteriorates. A sample of 30 drivers had their vision tested to determine the distance they could identify a roadside sign. The equation of the regression line is $y = -2.953x + 573.438$.

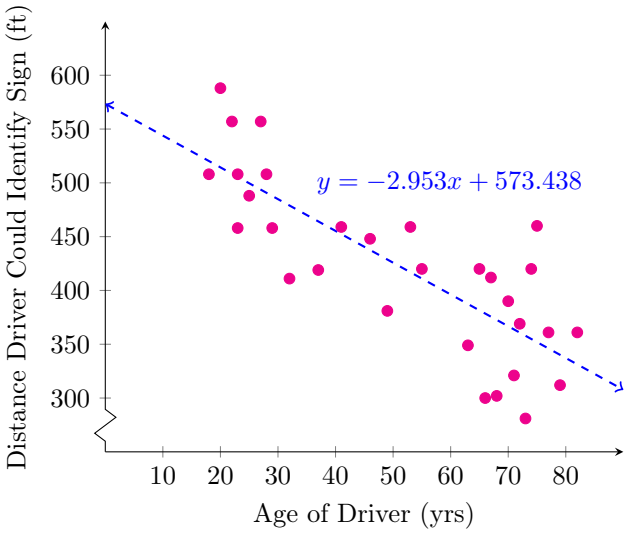

Figure 1.10.7: Age (yrs) vs Distance (ft) at which driver could identify road sign

(a) In what interval would it be unsurprising for a 30 year-old driver to recognize the sign?

(b) A 60 year-old driver claims to be able to recognize the sign at 500 feet. Is this claim credible? Explain why or why not.