Contemporary Pre-Calculus Through Applications

Contemporary Pre-Calculus Through Applications

Mathematics Department
North Carolina School of Science and Mathematics

July 27, 2018

Contents

Chapter 1

Data

Introduction to this chapter

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. 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).

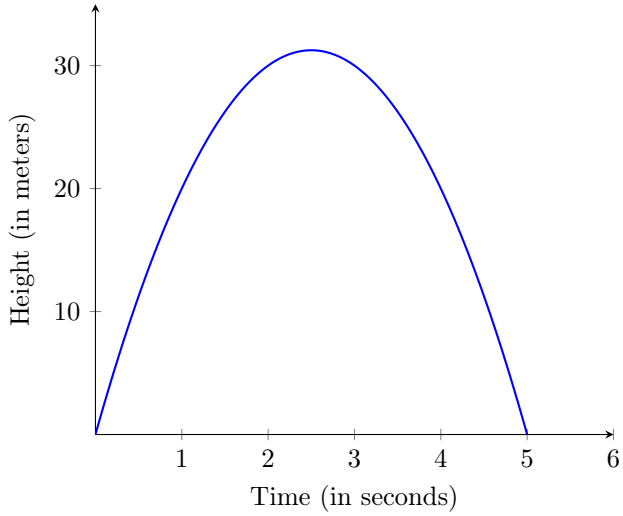

Figure 1.1.1: Height of a rock over time

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 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.

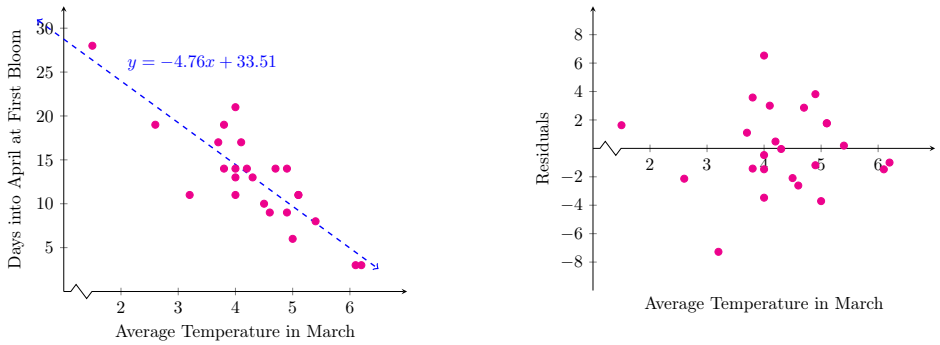

Figure 1.2.1: Least-squares line with residuals indicated

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 1.2.1, 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.

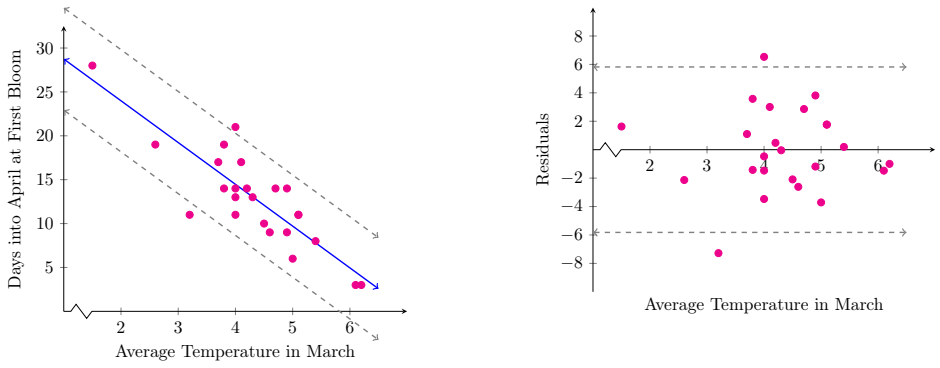

Figure 1.2.2: Least-squares line with residuals indicated

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 Galton's Father-Son data, the equation of the line is $\text{Son} = 34.428x + 0.5095\text{Father}$. 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.

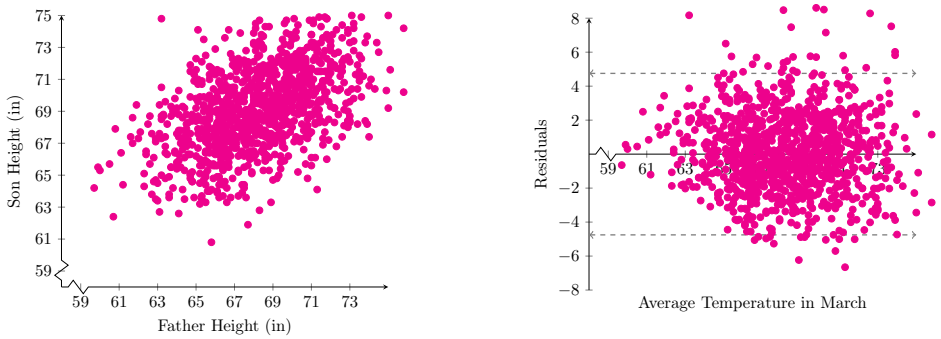

Figure 1.2.3: Least-squares line with residuals indicated