YOU SEEM SO DISTANT

Lab 4: Using Regression for Prediction

Figure 4-1: *What is the distance from the first fencepost to the tree at the right-center of the photo?*

INTRODUCTION

One of the main reasons we study the relationship between two quantitative variables is to try to determine if the value of one variable (called the response variable) can be reliably predicted from the value of the other variable (called the predictor variable). For example, college admissions offices use an applicant's SAT score to predict their first-year college GPA. A biologist might use the concentration of chlorophyll in water to predict primary production of aquatic plants (the rate at which the plants are producing new carbon).

Before we can determine if values of one variable can be used to predict values for the other, we must carefully examine the nature and strength of the relationship between the two variables. We do this by conducting a **regression experiment**, where values for both variables are obtained for a number of trials. If a reasonably strong relationship exists in the data, we develop a statistical model to predict new values of the response variable using the predictor variable.

SETTING

In this lab, we will determine the relationship between the distance you guess between two visible objects and the true distance between the objects. The purpose is to find out if we can effectively use guessed distances (easy measurements) to predict the true distances (more difficult measurements).

NOTE:

You will not be graded on how accurately you guess; we are interested in determining if a relationship exists between your guessed distances and the true distances. Often people do not guess accurately, but inaccurately guess in a consistent manner. Consider a person whose guess x is always about half as large as the true distance y. A good prediction formula for y can then be formed, namely $y \approx 2x$.

MATERIALS

Surveyor's measuring wheel for each group of four students

GROUP SIZE

Groups of four students will record guessed distances individually and measure true distances. Students work in groups of two for the computer data analysis.

METHODS

Weather permitting, your lab instructor will take you outside to a field or similar place and identify a reference point, called the base. He/she will then point out 11 objects (e.g. trees, fence posts, etc.) in the field one at a time. For each object, you should silently guess its distance to the base in feet. Record these guessed distances in Table 4.1. Do not talk about your guess with others in the group; this could influence their guesses.

Figure 4-2: *Surveyor's Wheel*

Once all groups have recorded their guessed distances, your group will use a surveyor's measuring wheel to measure the actual distances between the base and your group's set of assigned objects. Each distance will be measured by several groups. Record measured distances for your group in Table 4.1. Note that the 11th object's distance will not be measured at this time; it will be predicted later using a statistical model.

DATA ANALYSIS

Upon returning to the lab computer room, the instructor will lead the class in determining the median measured distance for each object. From this point on, use this median distance as the "true" distance for each object. Students will now work in pairs using statistical software to explore the relationship between their guessed distances and the true distances to determine a statistical model for prediction.

For preliminary assessment, repeat steps 1-3 below for each of your computer-group members:

1. Use statistical software to create a scatter plot of your data set. Use "Guessed" as the x-variable name and "True" as the y-variable name. Answer Discussion Question 2 before going further.
2. Use software to superimpose the line $y = x$ over your scatter plot. (Note that this is *not* a line of best fit. If you were a perfect guesser, then your guesses would fall exactly on the line $y = x$. For most guessers this will not happen.)
3. Print this graph. You will use it to answer Discussion Question 3.

Determining the Line and Curve of Best Fit

If the point-cloud in your scatter plot has a fairly linear shape, then you may be able to use your guessed distances to predict true distances using a *linear* regression model. Use software to find the equation of the line of best fit for the scatter plot. Print the plot with the fitted line as well as the software's output summary table for the fitted line. Do this for each member in the group.

If your scatter plot shows more of a curved shape than a linear shape, it may be better to fit a curve to your data than a simple line. Use software to fit your scatter plot with a quadratic regression curve. Print the plot of the fitted quadratic curve and also the software's output summary table for the fitted quadratic. Do this for each member in the group.

Prediction Error

Predictions using real data are almost never perfect. A statistical model acknowledges this by providing a way to generate predictions as well as to measure precision for these predictions. Every fitted statistical model includes a simple measure of precision called the *error standard deviation* s. This quantity is essentially the standard deviation of the prediction errors (also called *residuals*)

$$(\text{true } y \text{ value}) - (\hat{y})$$

under the model. (Note: some software programs don't label the error standard deviation as "s" on the output—your instructor will help you find the right value). If the model is a good one, the empirical rule for standard deviation can be applied to the error standard deviation: approximately 95% of future prediction errors should lie in the interval

$$(\text{average prediction error}) \pm 2s.$$

And since (if the model is a good one) the average prediction error is 0, we can simplify the above: approximately 95% of future prediction errors should lie in the interval $-2s$ to $2s$. We call the quantity $2s$ the "95% prediction error." For each of your fitted models, write the value of the error standard deviation s and the 95% prediction error $2s$ in Table 4.2. Note that the error standard deviation s is a very natural measure of the strength of the relationship between the two variables: if s is small, then values of x can be used to very precisely predict y, so there must be a strong relationship between the two.

CHOOSING A PREDICTION MODEL

We have now fit two models to your collected data: linear and quadratic. You must choose which of these two models to use for predicting true distances using only guessed distances.

1. We will first consider the visual quality of the fit in each case. Which seems to follow the plotted points more closely—the fitted line or the fitted quadratic curve? Provide your answer in Table 4.2 (Note: it is possible that both models provide good visual fits; it is also possible that neither model does).

2. Which model has the smallest value of 2*s*, the 95% prediction error?
3. If the visual quality of fit (a) seems to be acceptable and similar for the two models, and the prediction error (b) is also similar, usually the simpler model is chosen for use—in this case, the linear model is simpler.

PREDICTING THE MYSTERY 11TH DISTANCE

We have now fit two models to your scatter plot data: linear and quadratic. Give your model choice in Table 4.3, with your reasons. In the third row of Table 4.3, write the prediction equation for your chosen model. For example, if the linear model was chosen, the equation might look like

$$\hat{y} = 10.1 + 2.2x$$

except with different numeric values. If the quadratic model was chosen, the prediction equation might look like

$$\hat{y} = 2.3 + 3.8x + 2.5x^2$$

Now, in the fourth row of Table 4.3, fill in the value for your guessed distance to the 11th object and then use this value for x in your prediction equation to calculate a predicted distance $\hat{y}$. Show your work.

Your instructor will then supply you with the true distance y to the 11th object; write this value in the fifth row of Table 4.3 and use it to calculate the prediction error (true y) - $\hat{y}$, to be written in the 6th row of Table 4.3. Finally, in row 7, answer yes or no: Does your prediction error lie within the 95% prediction bounds -2*s* to 2*s*?

Reality Check

The Glomerular Filtration Rate (GFR) is a measure of how efficiently a person's kidneys filter waste from the blood. GFR can be measured directly but this process is expensive, time-consuming, and invasive (NIH, 2009). For example, at the very least it requires an injection into the bloodstream and a urine sample collected 24 hours later. An easier measurement is the concentration of creatinine in the patient's blood sample. Creatinine is strongly associated with GFR and hence a good predictor of kidney function. (The prediction equation also depends on the patient's age, gender, and race.)