ADDRESSING CHALLENGING STATISTICAL TOPICS WITH MINITAB

Julie Belock
Salem State University
352 Lafayette Street, Salem, MA 01970
jbelock@salemstate.edu

Introduction

In over fifteen years of teaching statistics courses, I have found that the thorniest concepts for students to understand include the meaning of confidence intervals, interpretation of p-values and interpreting regression diagnostics, particularly residual plots. I have developed student projects and lab activities that attempt to help students understand and apply these concepts. I use Minitab software in these projects for its ease of use and production of excellent graphs, which aid the students in interpreting and presenting their work.

The activities I describe below exemplify the recommendations of the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report* [3], which include stressing conceptual understanding rather than mere procedural knowledge, and using technology for development of that conceptual understanding. The GAISE report also recommends using real data, which may include simulated data [3]. The activities designed to improve conceptual understanding of confidence intervals and p-values make use of simulations and an intuitive understanding of relative frequency as an approximation of probability. The assignment that addresses regression diagnostics uses an existing set of real data. Using simulations and using real data are commonly cited as important features of a modern statistics course.

For each of the three concepts: confidence intervals, p-values and residual plots, I describe an example of one student activity. In the activity the Minitab commands are detailed and in some cases I have used screen shots of dialog boxes to assist the reader who may not be familiar with the program. I also include output and what I am expecting the students to notice in each case.

Confidence intervals

Computing confidence intervals is simple; correct interpretation tends to be the problem. The paradox is that to correctly interpret an interval obtained from a single sample, the student must imagine many, many samples obtained under the same conditions. To help students understand the meaning of a confidence interval, I have designed an activity using simulated data. The activity may be used as either a lab activity or a homework assignment.

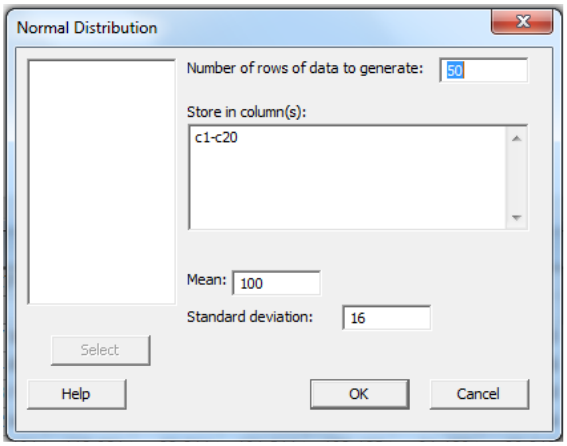
The key idea here is that the confidence level represents the chance of in interval capturing the true mean *before* the random sample is obtained and any computations are done. I stress to the students that once an interval is computed for a particular sample, it either contains the true mean or it does not; there is no longer anything random about it! (A clear explanation of the meaning of confidence intervals appears in [1], among many other sources.) So, to help them understand that the process of obtaining the confidence

interval is where probability comes into play, we use the idea of relative frequency as an approximation of probability.

Example: Scores on the Stanford-Binet IQ test are assumed to be normally distributed with a mean of 100 and a standard deviation of 16.

1. Use Minitab to generate 20 different samples of size n = 50 from this population. To do this, use the **Calc > Random Data > Normal** menu command to simulate 20 samples, each with 50 IQ scores. Store each sample in columns c1 through c20, generate 50 rows of data, and use mean = 100 and standard deviation = 16 (Figure 1).
2. Minitab computes confidence intervals based on the sample data. Find a 90% confidence interval for the population mean of all IQ scores (which we already know is 100) for *each* of your 20 samples. To do this, use the **Stat > Basic Statistics > 1-Sample Z** menu. You must input a standard deviation, which is assumed to be 16. Also, input c1 through c20 in the variables box. To select the confidence level, select **Options** and enter 90 for a 90% confidence interval. All intervals will be displayed in the session window.
3. How many of your confidence intervals would you *expect* to contain the true mean and why? How many of your 20 confidence intervals actually do contain the true mean of 100? Compare the proportion of intervals that contain the true mean to 90%.

Figure 1: Simulating random samples of 50 values each from N(100, 16).



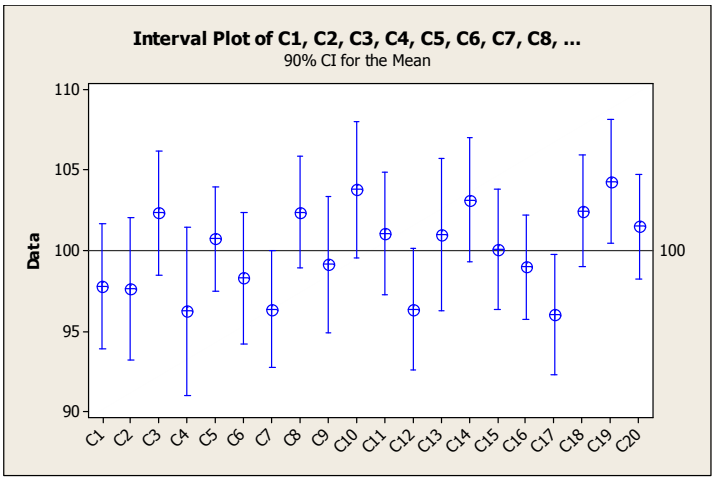
Students are asked to simply observe the proportion of intervals that actually contain the true mean, which is assumed to be 100 in this example. If this is being done as an in-class exercise, I have all students combine their findings; we find that the proportion of intervals containing the mean is usually quite close to the confidence level. In the simulated confidence intervals in Figure 2, 17 out of 20, or 85% of the intervals actually contain the population mean of 100 (the highlighted intervals are those **not** containing the population mean).

Figure 2: Twenty 90% confidence intervals for the population mean computed from simulated random samples.

Variable	90% CI
C1	(94.10, 101.54)
C2	(93.90, 101.35)
C3	(98.61, 106.05)
C4	(92.53, 99.98)
C5	(97.01, 104.46)
C6	(94.57, 102.01)
C7	(92.65, 100.09)
C8	(98.67, 106.11)
C9	(95.42, 102.87)
C10	(100.05, 107.49)
C11	(97.34, 104.78)
C12	(92.65, 100.09)
C13	(97.25, 104.69)
C14	(99.40, 106.85)
C15	(96.38, 103.83)
C16	(95.26, 102.71)
C17	(92.32, 99.77)
C18	(98.74, 106.18)
C19	(100.57, 108.02)
C20	(97.79, 105.23)

Students can also use Minitab to create a graph illustrating that some of the intervals contain the true mean of 100. Using menu selections **Graph>Interval Plot** gives a graphical display of the confidence intervals we have just computed. Note that the default confidence level on this graph is 95%; to change to 90%, first create the graph, and then right click on any of the interval bars, select **Edit Interval Bars** and **Options** to adjust the confidence level. For the graph in Figure 3, I also added a reference line at y = 100, which represents the population mean.

Figure 3: Interval plot of twenty 90% confidence intervals for the mean.



Students should observe that indeed three intervals are not crossing the reference line, indicating that they do not contain the population mean (although in this example, the confidence interval with upper bound 99.98 appears to end right on the line).

P-Values

The p-value for a hypothesis test is the probability of obtaining a test statistic at least as extreme as the one that was actually observed from the data, assuming that the null hypothesis is true. This is often poorly understood by students.

An exercise that can help students understand the concept of a p-value again involves simulation and an understanding that the relative frequency of an event is an approximation of its probability. We consider a test about a single population proportion in the following example.

Example: In the past, about 35% of college graduates went directly to graduate school after graduation. In recent years educational researchers believe this percentage may have increased. Suppose that a simple random sample of 215 college graduates is obtained and that 84 of them went immediately to graduate school. Does this sample provide evidence that the percentage of all college graduates going immediately to graduate school has increased?

1. Set up hypotheses and describe the hypotheses test (including assumptions) we should use to answer this question.
2. Use Minitab to simulate the results of the above situation 100 times. That is, we want to simulate taking 100 samples of 215 graduates under the assumption that 35% of the population of graduates go to graduate school. To do this, view each of the 215 graduates as an independent trial with a 35% chance of "success" (Binomial with n = 215 and p = 0.35).
   Generate random observations from a binomial distribution with probability of success p = 0.35 and n = 215 trials by selecting Calc > Random data > Binomial. Generate 100 rows of data, store in C1, enter 215 for the number of trials and 0.35 for probability of success.
3. Based on your findings in part 2, what would you say is the chance that an observation from a bimomial(215,0.35) model would be greater than or equal to 84? To make it easier to tell, sort the data in C1 in descending order.
4. Use Minitab to compute the p-value for the test and compare to your answer to part 3. Write a sentence or two interpreting your p-value in the context of this problem.

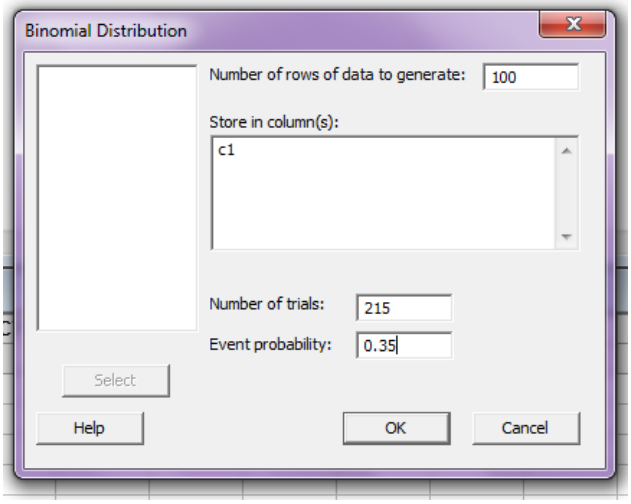
These are some of the ideas to highlight for the students before and during this activity: we are trying to recreate the conditions of the null hypothesis, which is a binomial situation with p = 0.35 and n = 215. We are using simulation to generate some outputs, say 100, and look at the proportion of results that exceed the number we found in the original sample. Note that each output represents the number of students who say "yes, I am going directly to grad school" in a random sample of 215 students.

The first step is to set up the hypotheses for this situation:

$$H_0 : p = 0.35$$
$$H_1 : p > 0.35$$

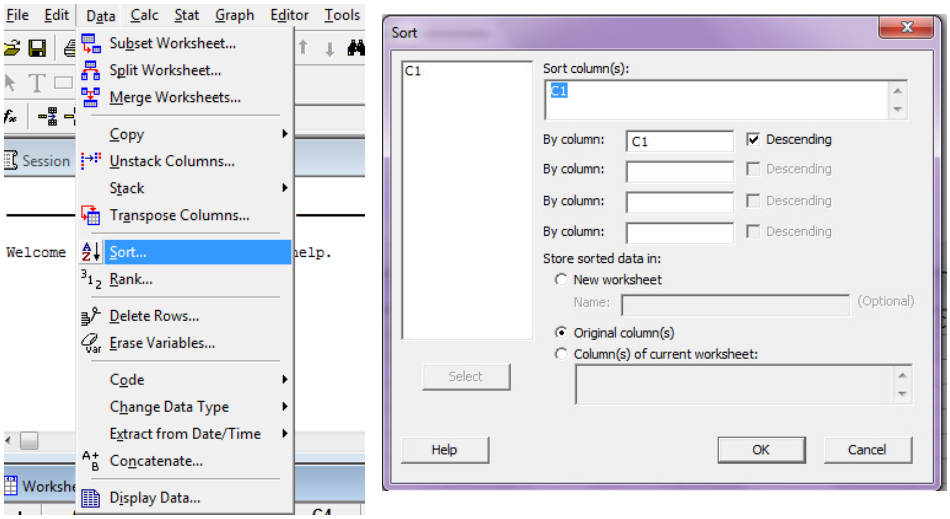
Next we simulate taking 100 random samples of size 215 from a binomial population with probability of success p = 0.35 using Minitab; the dialog box to do so is shown in Figure 4.

Figure 4: Simulating binomial data.



In the problem, the sample included 84 "successes." Since the p-value is the probability of obtaining 84 or more successes in subsequent random samples of size n = 215, we are interested in examining the proportion of our simulated data values that are 84 or greater. To make this easier, sort the data (Figure 5).

Figure 5: Sorting the data (descending order).



Below are the results from two sets of simulated data on Minitab:

- Run 1: X ≥ 84 in 11 of 100 outputs (11%)
- Run 2: X ≥ 84 in 10 of 100 outputs (10%)

Students should then compare proportions of values that are at least 84 in 100 simulated samples with the actual p-value. The Minitab dialog box for performing a 1-proportion Z-test using the summarized data from this problem is shown in Figure 6, and the output is provided in Figure 7. The p-value = 0.105, clearly very close to the proportion obtained in the simulations. (And if the two simulations are pooled, we happen to get 21 / 200 = 0.105, which equals the computed p-value.)

Figure 6: Running a 1-proportion Z-test from summarized data

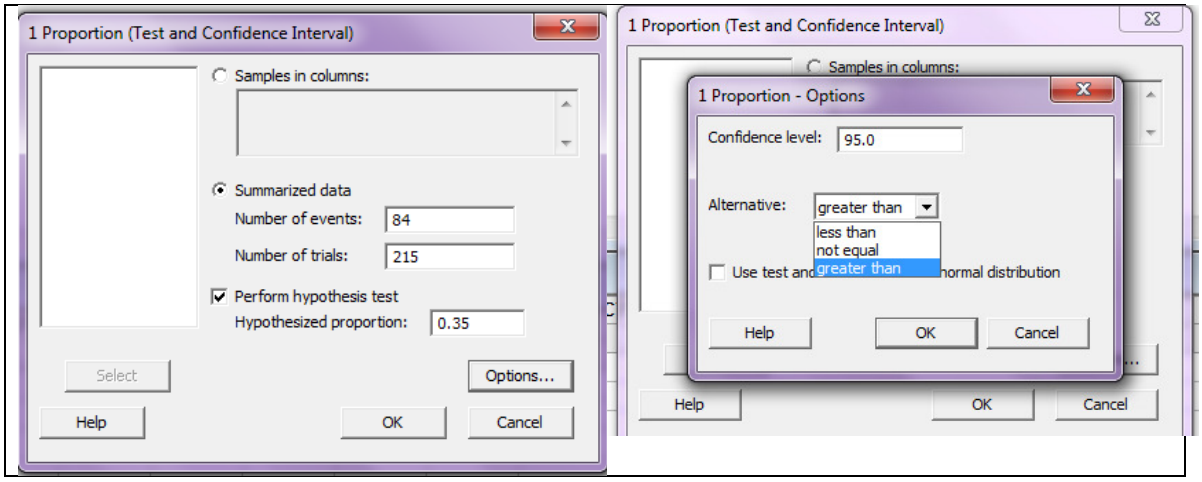


Figure 7: Minitab output for a 1-proportion Z-test

Test and CI for One Proportion

Test of p = 0.35 vs p > 0.35

Sample	X	N	Sample p	95% Lower Bound	Z-Value	P-Value
1	84	215	0.390698	0.335965	1.25	0.105

Regression diagnostics

When we explore with students the idea of an association between two quantitative variables, one of the first tools we use is a scatter plot. Students learn to describe the appearance of a scatter plot and see what that means for the association between the two variables: Is it positive or negative? Curved or linear? A clear pattern or more diffuse? Once we progress to trying to fit a model to the data in the form of regression we may examine a residual plot to assess the suitability of the chosen model.

Students often have trouble interpreting a residual plot. The main issue may be confusion between what a residual plot is "supposed" to show versus what a scatter plot is "supposed' to show:

- In a scatter plot it's "good" to see a pattern or association
- In a residual plot it's "good" to see NO pattern at all – this means the model describes the data well.

To help students interpret these graphs correctly and understand their different purposes, I have developed an assignment based on a relatively well-known set of real data. The data set, compiled by biologist Dr. Gary Alt, includes data for a number of measurements of black bears, including weight, length, neck girth, chest girth, head girth and head length. This data also appears as a sample data set in Minitab, and the following description of the data set is included under Minitab's Help menu:

Wild bears were anesthetized, and their bodies were measured and weighed. One goal of the study was to make a table (or perhaps a set of tables) for hunters, so they could estimate the weight of a bear based on other measurements. This would be used because in the forest it is easier to measure the length of a bear, for example, than it is to weigh it. This data set was supplied by Gary Alt. [2]

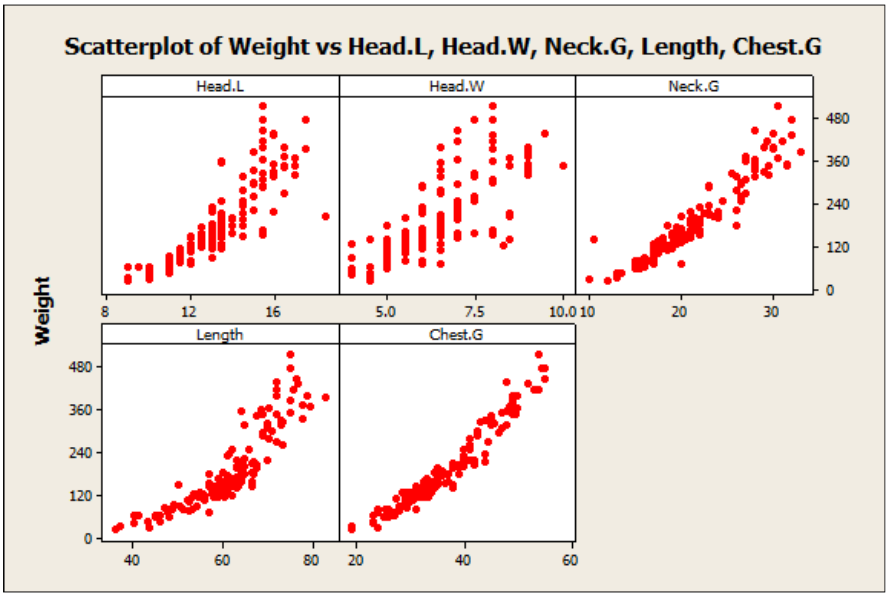
Since this activity was for an introductory statistics course, I only focused on simple (rather than multiple) regression, although I usually do briefly discuss nonlinear regression. In the assignment below, students are first asked to apply linear regression to a pair of variables with a distinctly nonlinear association, and this is to focus on how the residual plot behaves when you know you don't have a good fit. Then students are asked to find a more appropriate candidate for linear regression, analyze the residual plot and interpret the results.

Example: Look in the Minitab Sample Data folder for the file named BEARS.mtw.

1. Create scatter plots for weight vs. each of the following variables: head length, head width, neck girth, length and chest width. Describe the association, if any, apparent in each plot.
2. Choose a scatterplot that shows a ***nonlinear*** association. Use **Stat > Regression > Fitted Line Plot** to run a linear regression anyway. Select Graphs and choose to look at residuals vs. the explanatory variable. Describe the residual plot.
3. Do any of the associations appear linear? If so, choose one and use **Stat > Regression > Fitted Line Plot** to find the linear model for the data.
4. Examine the plot of residuals vs. the explanatory variable. What does the residual plot tell you about the model? How could you adjust the model to improve it?

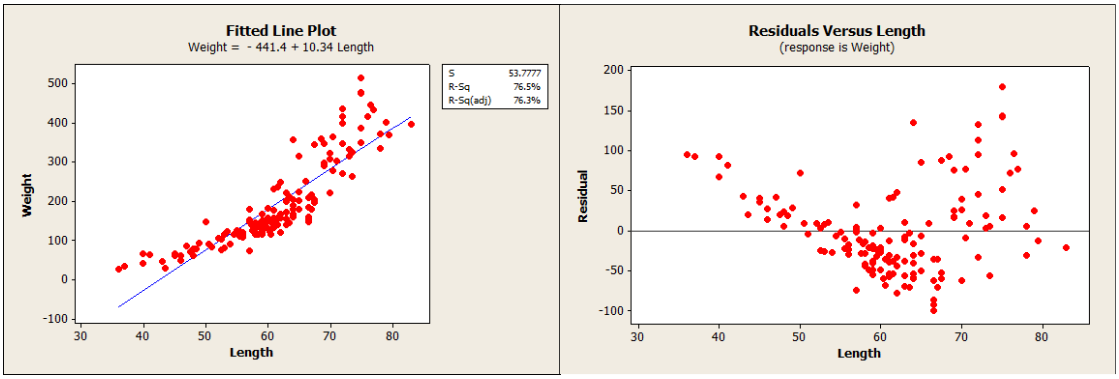
In Figure 8, you can see the five different scatterplots students are asked to create;

Figure 8: Scatterplots of weight vs. five different variables



It is clear that several of the associations are nonlinear. Students were then asked to do a linear regression on a clearly nonlinear plot and examine the residuals. The scatter plot in the lower left corner of Figure 8 shows that one example of a nonlinear association is the pair of variables Weight and Length. In Figure 9, a regression line is shown on the scatter plot and the resulting residual plot is shown. I want students to observe that the curved pattern on the scatter plot results in a curved residual plot (when trying to fit a line). In my introductory courses, I generally have students plot residuals versus the explanatory variable because it makes it easier to see the connection between the residual plot and the scatter plot.

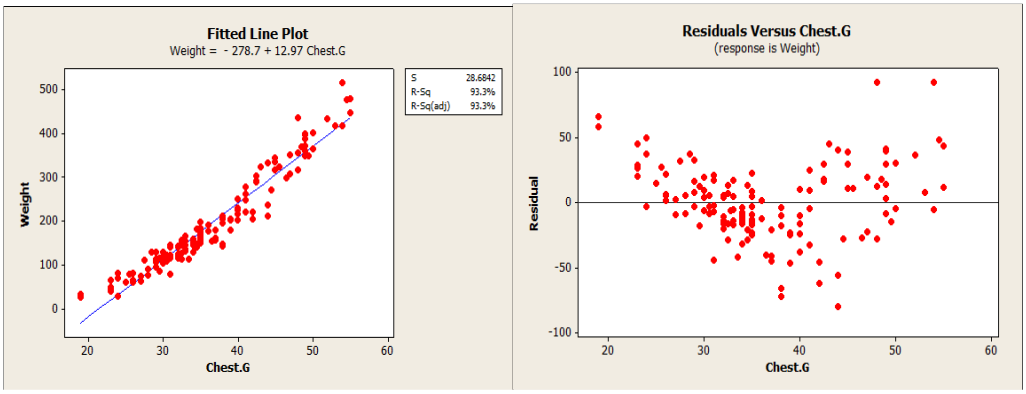
Figure 9: Regression line and residual plot for Weight vs. Length



Next the students are to try a linear regression on a pair of variables that appear to have a linear association. Of the five pairings, chest girth and weight look as though their association may be linear. Computing the correlation coefficient yields a value of 0.966,

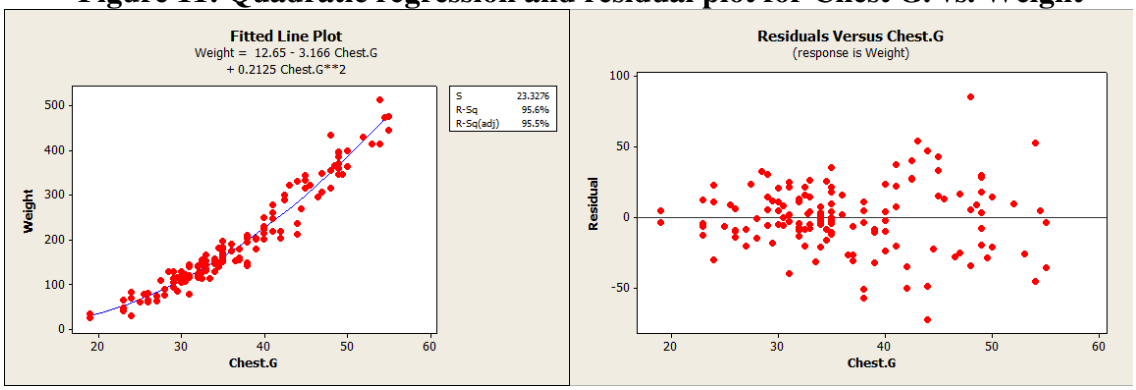
indicating a strong positive correlation. However, in Figure 10, the residual plot reveals that there is actually a subtle curve to the plot. Students should then conclude that linear regression is not appropriate to model this relationship.

Figure 10: Linear regression line and residual plot for Chest G. vs. Weight



Since it appears that there is some kind of strong association between the two variables, students are asked to try to improve the model. Since there is a bit of a curve, the next attempt should be quadratic. The fitted quadratic curve and the resulting residual plot are shown in Figure 11.

Figure 11: Quadratic regression and residual plot for Chest G. vs. Weight



The fitted quadratic (also found using Fitted Line Plot on Minitab) appears to fit the scatterplot very well. Students should now observe that the residual plot appears scattered with no apparent pattern, which indicates that the model is a good fit. Unfortunately, there is also a bit of "fanning" as the residuals on the right side of the graph appear farther on average from those on the left, which may violate the constant variance assumption for using regression, but it does not appear pronounced in this case. I would expect the students at this point to conclude that a quadratic model seems to be a good fit.

Conclusion

I have used versions of these activities in a variety of statistics courses, from the standard introductory course to the calculus-based mathematical statistics course to a graduate course designed for middle school teacher. No matter what their level, I found that students struggling with concepts such as the meaning of a confidence interval, the interpretation of a p-value or the uses of a residual plot can benefit from the activities I have described above. In working through the assignments, whether as a homework assignment or an in-class lab session, the students are active learners using technology to enhance their conceptual understanding. While other technologies may be used to achieve similar goals, my experience has been that Minitab works better than others for these particular activities due to several factors, including ease of use, clear graphics and appropriate options, such as the ability to generate and display multiple confidence intervals simultaneously.

References

[1] DeVeaux, Velleman, Bock. *Stats: Data and Models*, 3rd ed., Pearson Addison Wesley, Boston, 2012.
[2] Minitab® Statistical Software Release 14
[3] Zieffler, Andrew and Karl, Stacy, ed. (2010), *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report*, American Statistical Association, 2012.