Stat 100 Project 2

Purpose: To use a spreadsheet program to produce a scatter diagram, regression line, regression line equation, and correlation coefficient from sample data, and explore how these change with the data.

Reading: Section 3.4 on scatter diagrams, 3.5 on Pearson’s correlation coefficient, and 3.6 on the equation of the regression line.

Turn in: The spreadsheet window, including the regression plot, and the answers to the questions at the end of this assignment. A one or two page submission is all that is required.

Instructions: It is assumed that you have worked Project #1. If you have not, please do so now – at least through the computer part. It only takes a few minutes. Also remember that if you make a mistake, simply type ctrl-z to undo the last operation, and try again.

Before we start, let’s describe what we are doing. You may know that crickets chirp at a rate that depends on the temperature. If you go online, you can discover that a way to estimate the temperature (in degrees F) when you hear a cricket, is to count the number of chirps in 14 seconds and add 40.

We are told that crickets make the sound by rubbing their legs together. Sounds a bit strange perhaps, but they have bumps along their legs that help create the sound. The bumps create short bursts of sound as they over each other. A group of short bursts create the chirp. So the question is this: crickets rub their legs less frequently when it is colder, do they also rub them more slowly?

We have data to help us answer this question. Someone has measured the number of bursts per second for the ground striped cricket for various temperatures. We have loaded this data into a spreadsheet (much as you did in Project #1). You can find a link to this spreadsheet in the projects table. This spreadsheet can be opened in either Excel or Calc.

Once the spreadsheet is opened, you will see a two-column table of numbers. The first column is temperature and the other is bursts per second. Let’s make a scatter plot of this data so we can visualize it.

Click on the first number in the upper-left of the table, cell a6. Holding the left-mouse button, down drag down to the lower-rightmost number, cell b20, and let go of the mouse button. The two columns should be highlighted. Next click on “Insert” in the ribbon or menu and select “Scatter” under “Chart.” Select the chart that only has markers, no lines.

Move the chart by clicking just inside the upper left corner (not on the corner itself, and not in the body of the chart) and, while holding the mouse button, drag the chart to row 9, next to the two columns of data. (Calc users may have to click outside the chart first.)

Click on the legend in the chart (Calc users will need to double click on the chart first.) and press the delete key. We won’t need the legend and it takes up valuable space.
Right-click on the x-axis (on the 80.0, say) and click on “Add Major Grid Lines.” In Excel 2003, this function, and the titles, are accessed through the “Chart” menu, which is visible if the chart is selected.

Now let’s give the chart some titles so that someone else might be able to determine what the data are. Click on the “Layout” above the ribbon and then “Chart Titles.” Select “Above Chart” and type “Chirp Burst Rate for the Striped Ground Cricket -” (without the quotes) followed by your name, followed by the “enter” key. The title should appear in the graph. If not, click on the title and try again. (Calc users double click on the chart and use the insert menu.)

Next add the horizontal axis title: “Temperature, deg. F” and then a rotated vertical axis title: “Chirp Burst Rate, per Second”. You can click on the titles and move them around to improve the layout.

We can now clearly see that the Chirp Burst Rate does change with temperature, and maybe there is a straight line that would fit the data reasonably well. The spreadsheet provides several built-in functions for finding this line.

Click in cell e6 and type “Slope.” Then click in cell d6 and type “=slope(“ (without hitting enter just yet). Now click on the y-dataset. That is, click on cell b6 and drag down to the last cell in the column, highlighting the whole column. Release the mouse button and then type a comma. Next highlight the x-column the same way and type “)”. You should now see: “=SLOPE(B6:B20,A6:A20)” Press the enter key and this should change into a number – the slope of the linear regression line through the data. (If you leave out the “=” sign, you will just see SLOPE(B6:B20,A6:A20), and not the numerical result. Try again.) (Calc users will need to use a semicolon rather than a comma.)

Click in cell e7 and type “Intercept.” Then click in cell d7 and type “=intercept(”. Proceed as before to get the y-intercept of the linear regression line.

Finally click in cell e8 and type “R^2” and then in cell d7 and type “=rsq(“. Continue as before to get the Pearson Correlation Coefficient, $R^2$.

Seeing the slope, intercept and correlation coefficient for our data is nice, but we would really like to visualize the regression line itself. Fortunately, the spreadsheet provides us an easy to do this. Click on the data points in the graph. (Calc users will need to double click the graph first.) Then right click on the data points and select “Add Trendline.” Go to “Forecast” in the dialog box that opens and enter “10” in the Forward and “10” in the Backward boxes. Also select “Display equation on chart” and “Display R-squared value on chart.” Finally, click on “Line Style” on the left and set the line width to 1.5.

If you hit enter and the dialog box closes before you are done, simply right-click on the data points again and select “Edit Trendline” to continue.

Let’s pretty-up the graph a bit at this point. Right-click on the data points in the graph again and select: “Format Data Series.” Then select “Marker Options.” Select “Built-in” under “Marker Type” and set the size to 4. Hit enter to close the dialog box.

Click on the equation in the graph and move it to a clear spot on the graph. We can make it a bit easier to read by setting the font size to 12 (under “Home”) and the “Shape Fill” under “Format” to white. Let’s also lock down the scales on the graph. Right click on the x-axis and select “Format axis.” Under “Axis Options” select “Fixed” under “Maximum and “Minimum.” (The
numbers in the boxes to the right should be 0.0 and 120.0, respectively.) Repeat for the y-axis, making sure that the minimum is set to 0.0 and the maximum is set to 25.0. Hit the enter key to close the dialog.

You can compare the slope, intercept and $R^2$ that you computed with that shown on the graph. They should be essentially the same. If one or more parameters differ, you need to find the error. A common one is swapping the columns in the slope, intercept, or rsq functions.

**Printout:** At this point you should save your work and print the worksheet. There should only be a single page of printout containing all of the data and the plot. (Check print preview first to verify this. You may have to click on the right edge of the graph and drag it to the left a little to get everything to fit.)

Now let’s experiment with the regression, and see what happens. Right click on the regression line and select “Format Trendline.” Try out each type of regression offered to see how the trend line changes. In Excel, you can select “Polynomial” to see how increasing the order of the trend line affects things. You may want to uncheck “Display equation on chart” and “Display R-squared value in chart” to better see what is happening.

Notice what happens when you increase the polynomial order. (Calc users will have to skip this part.) The points are more closely fit, but the extrapolation to the right and the left can go quite wild. This explains (to a large degree) why the simple linear fit is the most popular.

The exponential, logarithmic and power regression types are useful if you have an underlying, theoretical model of the data, and wish to exploit that knowledge. You can also force the y-intercept if you have some reason to believe that another value might be more appropriate than the one calculated from the data.

Return the regression type to linear and display the equation and $R^2$. Close the dialog box. Now experiment with introducing errors into the data. Select the first y-value and change it from 20.0 to 10.0. Notice what happens to the graph and how sensitive the regression line to this change. Try making it 0.0. The erroneous point itself stands out as an outlier – well away from the trend line – suggesting the data point was in error. Imagine trying to discover this error by just looking at the table of numbers! (This technique does not always work, but when it does it can save you a lot of trouble.) Experiment with changing a few other numbers.

**Questions:** Answer the following questions, being concise and clear:

1. Using the equation for the linear regression line or just by reading from the graph, what would be your estimate for the chirp burst rate at 80 degrees F?

2. Use the formula for the regression line to estimate the chirp burst rate at 40 degrees F.

3. Do you think this last estimate would be a valid conclusion? What might be wrong?

4. What is the value of the correlation coefficient $R$? (Note we are given $R^2$, not $R$.) Is $R$ positive or negative, and why?

5. What does $R$ tell us about how well the regression line fits the data (qualitatively)?