Statistics with Excel

This introduction assumes that you already have a basic familiarity with using Excel. The screen shots are taken from Microsoft Excel 2010-2016. Any earlier versions will have similar properties. You will be learning how to use spreadsheets to perform statistical analysis. Statistics involves collecting and analyzing data. We are often interested in determining whether one variable depends on another and in what way (e.g. linear, quadratic, exponential, etc.).

Population Growth

We will use the U. S. population for the years 1790-1890 given in the chart below to explore statistical analysis using Excel.

Year	*Population (millions)*	*Population Growth*
1790	3.929	1.379
1800	5.308	
1810	7.240	
1820	9.638	
1830	12.866	
1840	17.069	
1850	23.192	
1860	31.443	
1870	38.558	
1880	50.156	
1890	62.948	

It seems reasonable to consider that the larger a population gets, the more it will grow. We would like to know if the population growth in each 10-year period depends linearly on the size of the population.

Enter the data for the Year and Population into a new Excel Spreadsheet. Enter the heading Population Growth in column C. We will get Excel to calculate how much the population grows for each 10-year period. Enter the formula = B3-B2 into cell C2 and make sure it calculates the correct value. Then fill this formula in column C. *Note: You will not have a value for Population Growth in the last row corresponding to the year 1890 since there is no data for 1900 from which to compute the growth. Hence, only select the data for years 1790-1880 for any plots you create.*

Plotting

Create a scatter plot of Population Growth vs. Population (y vs. x) to see if we can find a relationship between the Population Size and the Population Growth. *Since these are data points, the graph should just use symbols(markers) without connecting lines. Make sure the graph has an appropriate title and both axes have appropriate titles.*

Although all of the points do not lie on a single straight line, you should be able to see that the general trend looks linear. However, one point seems to be much lower than expected. Let's investigate whether it makes sense or not.

What 10-year time period does it correspond to? _____
What significant event occurred in U.S. History during that time period? _____
How could this event explain a smaller population growth? _____

Now that we think this point is okay and not a mistake, let's continue.

In cell D1 type the text $y = 0.3x-0.15$. Enter the formula =0.3*b2-0.15 into cell D2. Fill this formula into the cells D3-D11. Repeat the steps to add the data for the line $y = 0.27x+0.5$ in column E.

Create a new scatter plot with the same Population Data as the *x*-values and the Population Growth with the two additional lines as the *y*-values. *The data points should just be symbols without connecting lines and the two lines should be graphed as lines without symbols.*

You should see that both lines seem like reasonable fits to the data. We would like to know which line best fits the data points. We will use statistical analysis, namely linear regression, to determine the best-fit line.

Statistical Analysis

In order to determine a line which best fits the data, we will review a few terms that will help us measure how good the line fits the data. The first measure is called the **residual** which measures the difference between the actual (experimental) *y*-value and the predicted *y*-value given by the line $y_L = mx + b$, i.e. residual = $y - y_L$. You can think of the residual measuring how well the line matched the data at each point. The second measure is called the **residual sum of squares** SS_{res} which takes the residual at each data point, squares it, and then adds them all together. This SS_{res} essentially gives you one number that measures how well the line fit *all* the data points. So what we are looking for is a line that gives us the smallest possible value of *SSres*. Let's compute this for the two lines considered above.

Use the spreadsheet to compute the values we are interested in for each line. Clear the data in columns D and E. Type the headings yL = mx +b, res, and res^2 into cells D1-F1. Also type the text m= and b= into cells A14 and A15, respectively. Then enter the value 0.30 in cell B14 and –0.15 in cell B15. This is just a convenient place to store the values for the slope and the intercept of the line.

Entering Formulas

Enter the formula in D2 to compute the value for the line yL = mx +b. Be sure to use '$' when referring to the cells containing the values for m and b. Fill column D with this formula. If needed, change your graph so that it includes this line along with the data points.

Have Excel compute the residuals (e.g. C2-D2 and fill) and the residuals squared (E2^2 and fill). *Look at your graph. It is plotting the residuals because that used to be the column containing values for the second line. To remove these data points from the graph, right click on the graph and choose Select Data. In the window that opens, select res and remove it. Be sure not to remove the other series.*

Type the text SSres into cell E13 and then compute the sum of the residuals squared in cell F13.

You should have gotten the value shown in the table below. Enter the value for SS_{res} in the appropriate box in the table below. Change the values of *m* and *b* in cells B14 and B15 to be 0.27 and 0.5, respectively. Note that Excel automatically re-computes everything and changes the graph. Enter the value of SS_{res} for this line $y_L = 0.27x + 0.5$.

Line	SS_{res}
$y = 0.30x - 0.15$	13.5576
$y = 0.27x + 0.5$	

Make an educated guess for another pair of values *m* and *b* for the slope and intercept and enter them into cells B14 and B15. Enter the your equation for the line and the associated SS_{res} into the table above. Which of the three lines above is the best approximation (so far) to fitting the actual data? _____ Continue varying the values of *m* and *b* slightly to see if you find values that result in a smaller value of SS_{res}. Once you find one, enter the information into the table above.

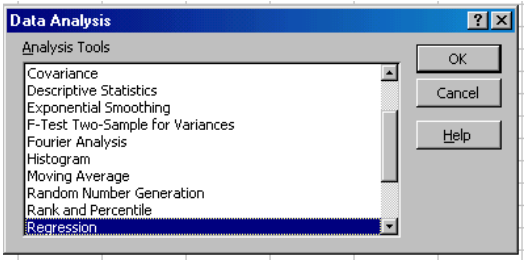
Linear regression

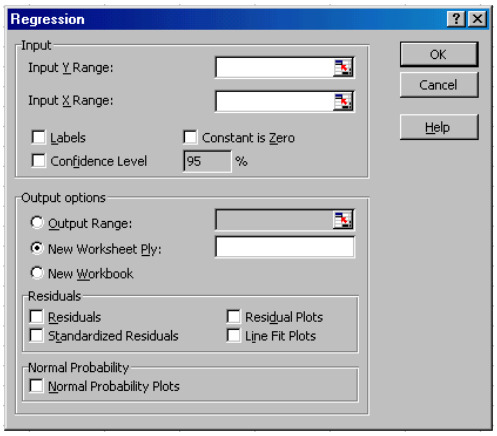
Although the spreadsheet can easily calculate these values, we still have infinitely many combinations of *m* and *b* that we could try to find the lowest value of SS_{res}. Fortunately, there are formulas based on Calculus that will give us the values of *m* and *b* that will result in the smallest possible SS_{res}. Even more fortunate is that these formulas are already programmed into Excel so that you don't have to do much work.

First, since we may want to refer to all of our previous work, save the current spreadsheet. Add a new sheet by clicking on the + button next to the Sheet 1 tab at the bottom. Select the first 3 columns in Sheet 1 containing the original data set and copy them. Paste this data into a Sheet 2.

Then go to the Data Tab and select the Analysis Grouping (if it appears). *If this grouping does not appear, click on File->Options. In the Excel Options window that opens, click on Add-Ins. Select Analysis ToolPak. At the bottom, it should say Manage: Excel Add-Ins. Click Go. Select Analysis ToolPak with a check mark and click OK. Once it has been installed, you should be able to find the Data Analysis option in the Analysis Grouping under the Data Tab.*

Select Data Analysis and the window on the left will appear.





Scroll down and select the Regression option and click OK. This option will find the values of coefficients that give the best fit of a line to the data. You should then see the window on the right. Enter the Y Range by clicking on the red arrow/grid and then highlighting the range of cells in the spreadsheet (i.e. highlight the Population Growth data). Close that popup window. Move your cursor to the Input X Range box and enter the Population data in the same way (*Remember to not include the population for the year 1890*). Also select the options Residuals and Line Fit Plots. In order to see the information in the same worksheet, select the Output Range option and click on cell E2 and it will be entered into this option. Click OK and you should see several new tables and a plot appearing on your worksheet. Autofit the column widths of this new data to view it easier. Resize and move the chart as desired.

The main information that we want from all of this new output is the slope, intercept, and SS_{res}. You will be able to find this information in the new table. About half-way down you will see a label for "Intercept". The value next to it under the heading "coefficient" is the value of the intercept *b*. Below

this is the coefficient for the X Variable 1 (i.e. the coefficient of *x)*, which is the slope *m*. Enter these values in the table below.

Slope *m*	Intercept *b*	SS_{res}

You can find the minimal SS_{res} for these values of slope and intercept in the table under the label ANOVA. In this table, there is a column labeled SS which stands for sum of squares. Now find the row labeled residuals and you will see a value for the sum of squares of the residuals. Enter this minimal value of SS_{res} in the table above.

The table in the spreadsheet labeled "Residual Output" gives the Predicted Y values and the Residuals. The Predicted Y are the values of Y for the line using these optimal values of *m* and *b*. You can see that the plot generated contains the original data and the Predicted Y values. *To make the graph look nicer, change the plot and axes titles to be something meaningful and accurate. Also, change the data to have only symbols with no connecting lines and the Predicted Y (i.e., best fit line) to be a solid line without symbols..*

Trendline

Another measure of how well the line fits the data set is to look at the *r*- and r^2- coefficients. The closer |r| and r^2 are to 1, the better the fit. An easy way to find the best-fit line and see the r^2- coefficient is to use the Add Trend Line option under the chart menu. This method is quick and easy, especially if you do not need all the detailed information generated by the Regression option.

Create a new Sheet 3 and copy and paste the first three columns into it. Create a new scatter plot of the Population Growth vs. Population data. Click on a data point in the graph. Right-click and select Add Trendline. In the window that opens, select linear. Check the boxes for Display Equation on Chart and Display R-squared value on chart. Your graph should now show the best-fit line and display the equation and r^2 value on the graph. If the r^2 value and equation are obstructing the view of the graph, you can move it around. Enter the values for r^2 and *r* below.

r	$\boldsymbol{r^2}$

What is the equation for the best-fit line? _____

Is it the same as the one you got using the Linear Regression method?_____

How well do you think the line describes the data?

[*Save your Excel spreadsheet as ExcelStats_YourNameHere.xlsx and email it to Dr. Crawford at crawford@elmhurst.edu.*]