Creating Scatter Diagrams (and Fitting a Line)
Excel 2010 Tutorial

Excel file for use with this tutorial	File Location
GraphTutorData.xlsx	http://faculty.ung.edu/kmelton/Data/GraphTutorData.xlsx

Introduction:

Purpose: Scatter Diagrams allow us to see if there is a relationship between two quantitative variables. A scatter diagram plots ordered pairs of observations. The vertical axis represents the dependent variable (the variable you would like to be able to understand/predict). The horizontal axis represents the independent variable (the variable that you believe might help you understand/predict the other variable).

Data for this example: We will use the data from the Enrollment sheet of the GraphTutorData file. The data on the Enrollment sheet relates to enrollment in the BBA program from Fall 2001 through 2012.

We would like to be able to estimate the number of credit hours of instruction when we are developing a course schedule. Unfortunately, this is very difficult to do. The question we will address here is:

It is easy to obtain a pretty good count for the number of students declared as business majors. Could we use this number to estimate the number of credit hours of instruction that will need to be delivered?

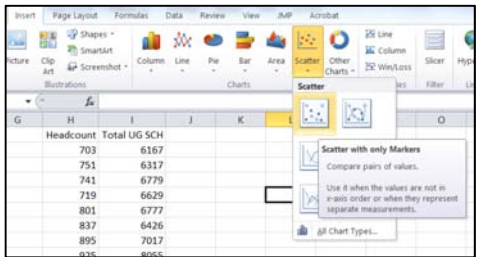
Creating the Graph:

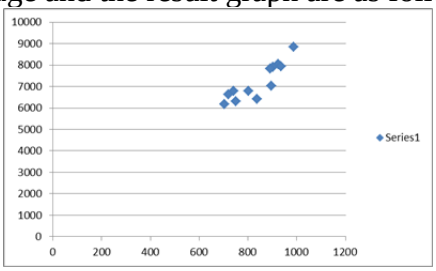
Step 1: Determine which variable you are trying to predict (the variable to go on the Y axis) and which variable you are going to use as a predictor (the variable to go on the X axis). For this example, we want to estimate the student credit hours (SCH), so this will go on the vertical axis. We think that might be able to estimate SCHs using the number of business majors (headcount), so this will go on the horizontal axis.

Step 2: Set up your spreadsheet so that the data for the variable on the X axis is in one column and the data for the variable on the Y axis in in the next column. For this example, the data have been copied into columns H and I. Then highlight the data (without highlighting the headings).

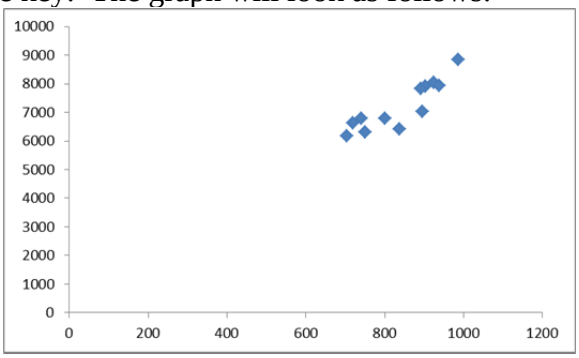
H	I
Headcount	Total UG SCH
703	6167
751	6317
741	6779
719	6629
801	6777
837	6426
895	7017
925	8055
904	7899
891	7815
937	7941
987	8843

Step 3: From the Insert tab at the top of the page, select Scatter and then the first sub-option (for Scatter with only Markers). Scatter diagrams, like histograms, ignore the time order of the data so we do not want to connect the dots. The screen image and the result graph are as follows:

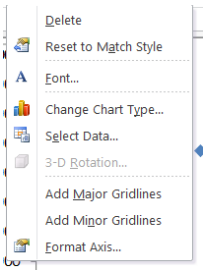


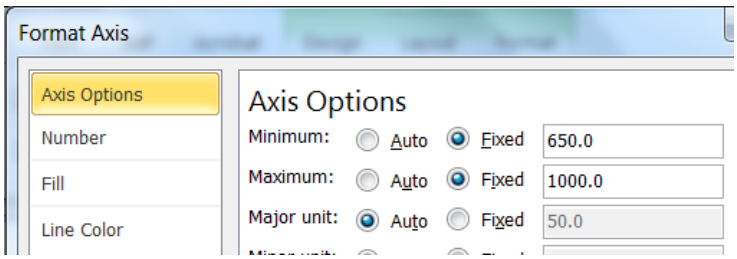


Step 4: Get rid of the chartjunk. For this graph, the legend and the gridlines are chartjunk. Click on the legend and hit the delete key; then click on one of the gridlines to highlight all of the gridlines, and hit the delete key. The graph will look as follows:



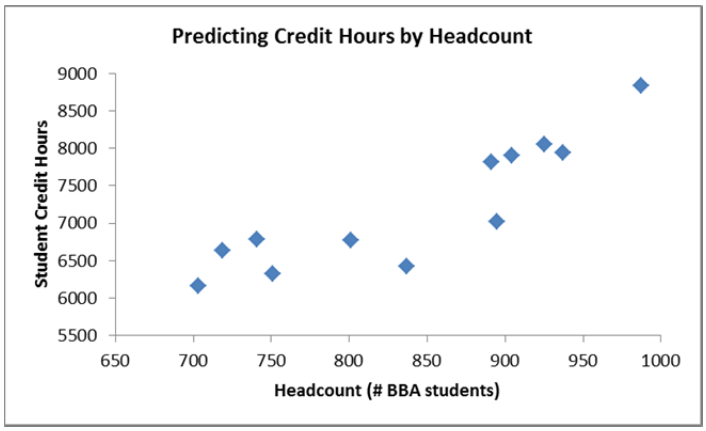
Step 5: Adjust the parts of the axes shown to focus on the data. To do this, we will eliminate the white space to the left and below the dots by changing the part of each axis that is visible. To change the horizontal axis, right click on one of the numbers on the horizontal axis to bring up the menu choices shown below; then select the Format Axis option. In the resulting dialogue box, change the minimum and maximum axis options from Auto to Fixed and select values just outside the observed values for X. The illustration below uses 650 to 1000



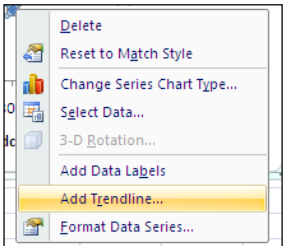


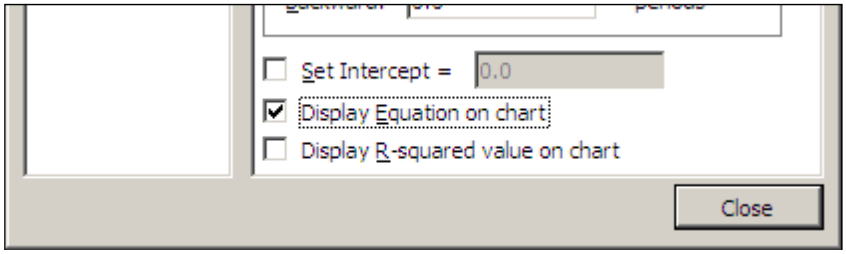
Repeat the process for the vertical axis to change these to show 5500 to 9000. Warning: If you select values that are within the range of the observations in the data set, these will not be seen on the chart—but will be included in any calculations related to the data.

Step 6: Add axis labels and a chart title. From the Layout tab, add a Chart Title and Axis Titles. The chart title should be chosen to clearly identify the predicted and predictor variables. Your resulting graph will look like the following:

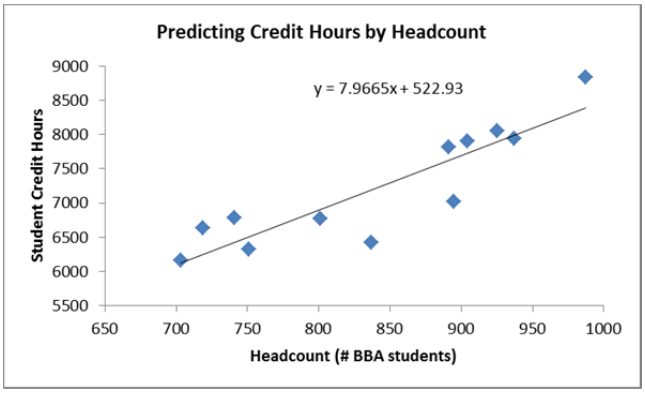


Step 7: Adding a trend line to the Scatter Diagram (Note: This is not really part of the scatter diagram, but is often added when the scatter diagram is used to show a linear relationship between the two variables.) To do this, right click on one of the dots and select Add Trendline… from the menu that appears. Then click on the box to the left of Display Equation on Chart near the bottom of the next dialogue box.





Your final graph will look as follows:



Communicating the Results:

From the scatter diagram (without the regression line), we can see that as the headcount increases, the number of student credit hours tends to increase. Headcount is not a perfect predictor of credit hours, but does appear to provide some insight. You could say the two variables appear to be positively correlated. From the regression equation, we see that for each additional student we would expect an additional 7.9665 credit hours. [Notes: 1) "Expect" to see relates to the statistical concept of expected value (or average). 2) The 7.9665 deals with a marginal change (for each additional student—not for each student). 3) The 522.93 is the vertical intercept (remember that extending the line to the vertical axis would include extending it through the white space we hid when we changed the values of the axes that we shows). Interpreting the 522.93 in the context of this problem would not be appropriate since we do not have any data near the situation where there are very few students.]

Cautions:
Statistics can show association but not causation—be careful not to phase responses as causal type statements.

Any "predictions" need to be limited to the range of X values included in the data set (and to conditions that match those in place when the data were collected. [For example, these data were collected prior to consolidation, and any predictions to post-consolidation would be suspect even if the predictions were still to business.]

Using the equation in a different setting even during the same period of time would not be appropriate. (For example: An equation based on number of students majoring in a field would not be a good indicator of credit hours of instruction in departments such as English, Math, or History where much of their instruction is to students in all majors—e.g., ENGL 1101, MATH 1111, HIST 1112.)

Checklist:

- Read the theory carefully to identify the dependent and independent variables
- Set up the data so that X is in one column and Y in the next column
- Remove the chartjunk
- Adjust the vertical and horizontal axes to focus on the part of the chart where the observations are plotted
- Provide a chart title that clearly communicates the Y and X variables (in the context of the problem)
- If appropriate, add a trend line showing the equation for the line
- Make sure that your conclusions are stated in terms of relationships rather than in language that implies cause-and-effect.