

## CE311S Lab 9

### Regression Tutorial #1

#### ***Purpose***

To learn the basic, underlying principles of least-squares regression and significance; and to learn how to run simple regressions using Microsoft Excel.

#### ***Introduction***

Least-squares regression offers a way to model a certain dependant variable as a linear combination of a series of independent variables. One way to view this is that regression creates a model with as many dimensions as there are independent variables, and in each dimension, it gives the function of a line which best approximates the relationship between the dependant variable and the independent variable specific to that dimension. The term “least-squares” refers to the fact that regression minimizes the square of the error between the actual data and the predictions generated by the model. The reason the square is used is two-fold: (1) Error can be either positive or negative, and to accurately minimize it, one needs a form of absolute value. Squaring the error provides this without the analytic complexities introduced by the discontinuities of the absolute value function. (2) By squaring the error, a “penalty function” is being used which exponentially penalizes a line (linear function) as it gets farther from data points.

#### ***Procedure/Tutorial***

In this tutorial, you will run a simple regression of the form:

$$I = \alpha_0 + \alpha_{YE}YE + \alpha_A A + \alpha_H H + \varepsilon$$

where  $I$  = income (in thousands of \$s)

$YE$  = years of education

$A$  = age (in years)

$H$  = height (in feet)

Before beginning this tutorial, you should think about how you would expect income to be affected by the three variables. Do you think some of the variables affect yearly income more than others? Do you think that some variables will have no effect at all? With regression, we can try to answer some of these questions.

1) Investigate the data using descriptive statistics (note, the colored areas in the worksheet will be used later in the lab, so *do not* place results in these areas). Does this data seem realistic? In what ways may it be biased?

2) Run a regression on the data. To do this, click Tools→Data Analysis→Regression. Your dependant (Y) variable is “true income,” and all of the data should be selected. The independent (X) variables are “years edu,” “age,” and “height” and the three data columns should be selected together (select the data by hand – do not click on the top of the column, this will produce an error). Be sure to include the first row of each column and check the “Labels” box in the regression window. Select the box “Output Range” and choose the cell M1 for the output of the data (this is the upper left corner of the light

blue box). This last step is necessary to get the automatically generated charts to display right. Click OK.

Look at the coefficients for the three variables. Do they make sense? Write about how they interact, especially the age and years of education variables. These latter two variables both involve time and imply that a year of education is worth more than a year of age (why do they imply this?) One cannot go to school forever, however, and one will always get older. That means that we can intuitively say that early on in a person's life, education is more "valuable" (with respect to income) than just working, but as one gets older, age begins to have a greater effect on one's level of income than education. You should find how old a person with only a high-school education (12 years edu) would have to be to have their age make a larger difference on their income level (height can be considered constant).

3) The  $t$ -stat for the height variable is  $< 1.645$  which means that, with a 95% confidence level, we can say that the coefficient for this variable is equal to zero. To see what this implies, predictions based on the regression results have been calculated. These calculations are shown in the purple section of the worksheet. These predictions were created by using the coefficients from the regression in the formulation of the original model (see above).

There are three charts included in the worksheet. Each maps the original data and some of the predictions in the dimension of one of the independent variables. Look at the three charts and think about what they show. What is important to note is how the predictions with all of the variables map a kind of rough "line" through the "center" of the original data. This is the essence of regression. The purple lines map the predictions of the chart's variable (the mean of the other variables has been used to place the line in a correct spot); these purple lines give a good one-dimensional representation of how least-squares regression approximates linear functions to fit the original data.

Now look at how the predictions are affected if one takes off the height variable. Does it look as though they are affected that much? The fact that the predictions are nearly the same is a graphical representation of what is meant by the results of the  $t$ -test earlier (where we found the coefficient on height to be statistically equivalent to zero). Look at the chart showing the "height" dimension. The distribution of the original data seems much more random than that of the other two charts (you may say they look almost the same, but you should see a linear trend in the other two that doesn't appear in the height chart). This is because there is not a strong relationship between it and the dependent variable, income. Regression and hypothesis tests thus allow one a way of testing how, *or if*, certain variables can be used to model others in a linear fashion.

4) As a final exercise, think about what information this regression seems to be missing. That is, are there some variables which you feel should affect income significantly which weren't included in this tutorial. How big of an effect do you think these variables would have. Would they be positive or negative? Why? Also, look at the  $R^2$  value. This value is equal to one if your model has a perfect fit, and is zero if your model is essentially random when compared with the original data. What does the  $R^2$  value tell you about the "goodness-of-fit" of this model? That is, would you use this model to plan out your life?