

# CE311S Lab 10

## Regression Tutorial #2

### *Purpose*

To learn how to calibrate a simple regression model and run hypothesis tests using its results.

### *Introduction*

Estimating construction costs for a job is an extremely important task for a construction firm. Informed prediction of how much a certain job will cost, given the attributes of the job, allows a firm to make a realistic bid for the build contract. In this lab, you will calibrate a regression model which can then be used to determine the probable construction cost of a large building. To develop this model, you are given data on 999 jobs that have been performed in the past. This data includes six variables:

**Labor** – The average number of workers per day on the job

**Project Length** – The length of job, in days

**Square Feet of Floor Space** – The floor space (in feet<sup>2</sup>) of the building

**Building Height** – The height of the building (in feet)

**Number of Rainy Days** – How many days it rained during the project

**Construction Cost** – How much the construction of the building actually cost (\$)

You will go through a series of steps in order to calibrate a model based on this data.

### *Procedure*

- 1) Download the Excel file titled “regress2.xls” from the Lab webpage.
- 2) Worksheet #1 gives the data that is given to you initially. Run descriptive statistics on the data to get a feel for what it contains. Does it seem realistic? What do the statistics for building height and square feet of floor space tell you about the building size?
- 3) Using the data in Worksheet #1, run a regression (Tools → Data Analysis → Regression) on the data. Include the first row of the data (the labels) and check the “Labels” box. Select “Output Range” and pick a cell someplace on Worksheet #1 to put the results of the regression.
- 4) Examine the coefficients on the regression results. What do they represent? Do they make sense? Does the coefficient (I should say coefficient estimator, but for brevity, I am shortening this to coefficient) on the “Days of Rain” variable seem realistic? Look at the t-statistics for the variables. A ( $\alpha = 0.10$ ) hypothesis test for whether a coefficient is zero or not with 90% confidence requires a t-statistic of 1.645 or greater. That is,

$$H_0 : \hat{\beta} = 0 \quad \text{accept if } t_{.05, \infty} \geq \left| \frac{\hat{\beta} - 0}{S_{\hat{\beta}}} \right|$$

$$H_A : \hat{\beta} \neq 0 \quad \text{accept if } t_{.05, \infty} < \left| \frac{\hat{\beta} - 0}{S_{\hat{\beta}}} \right|$$

(Why do we use infinity here?) So, any regression coefficient with a t-statistic less than 1.645 should be eliminated. However, it is inappropriate to eliminate all variables at once (unless you wish to run a more complicated test called an F test). So, we would like to first eliminate the variable which has the least statistical significance – that is “Days of Rain.” By eliminating this variable from our model, what are we actually saying about the relation between construction cost and this variable?

5) Worksheet #2 has the same variables without the “Days of Rain” variable. Run a regression on the variables and determine which ones, if any, are insignificant via a t-test. Eliminate the least significant variable (you can just highlight the column, right click, and select “Delete”), and run a regression on the rest. Continue until all variables are statistically significant.

6) You should have eliminated “Labor” and “Project Length” from your model. However, it would seem that the cost of a project would definitely depend on these two variables. Perhaps the form of these variables is not capturing their effect on the cost of a construction project. Think about what the variables actually say and of ways they might be interacted in order to more correctly account for their relationship to construction cost.

7) One way to relate the “Labor” and “Project Length” variables is to multiply them together. What does this new variable represent? Should it be a more accurate representation of how these variable affect construction cost? Worksheet #3 has these variable on set up. Run a regression on the variables and interpret the results. Are all the variables significant? Do they make sense?

8) Based on comparisons of competing firms, the generally accepted marginal cost per square foot of building is \$8 (marginal implies this is value is determined after controlling for other variables). Your model gives you a higher value per square foot than this. In order to be more competitive, your firm would like to replace the results of the regression with a value of \$8 per square foot. To test whether this is supported by the data, run a hypothesis test to see if a marginal cost of \$8 per square foot is statistically consistent with your model with a 90% confidence level - use a one-tailed test with

$$H_a : \hat{\beta}_{\text{sqr. foot}} > 8$$

What is your conclusion? What are the consequences of this result, in terms of the “competitiveness” of your firm?