1 Saratoga Real Estate

The R workspace *saratoga.Rdata* contains the data frame *saratoga* that contains the price (in dollars), size (Living.Area, in square feet), and if homes have a fireplace or not for homes is Saratoga NY. Here, we want to explore how the relationship between the price and size depends on having a fireplace or not. You will likely need the R scripts from the previous sessions and asynch materials.

1. Fit a regression model that uses the size to predict the price, denote this as model 1. Note the R^2 adjusted, and make sure you can interpret the model coefficients.
   (a) Make a plot of the price and size that includes the estimated regression model. You can use the base plotting functions or *qplot* in *ggplot2*.
   (b) Make a plot of the residuals and the fitted values (to check constant variance), but use different plotting characters and/or colors to represent the different fireplace levels. Can you visually detect any grouping in the residuals?
   (c) Use the function *tapply* to obtain the mean of the residuals for each of the fireplace levels. What do you find?
   (d) The estimated slope is a proxy for the price per square foot. Is there evidence the price per square foot is less than \$100? Even if the variance is constant, do the test using robust standard errors.
2. Fit a regression model that uses the size and the fireplace variable to predict the price, denote this as model 2.
   (a) Note the R^2 adjusted and make sure you can interpret the model coefficients. Does including the fireplace variable increase the R^2 adjusted?
   (b) Make a plot of the price and size that includes the estimated regression model. It is more challenging to plot the estimated regression lines for each of the fireplace levels, but if you can do this you should see parallel lines.
   (c) Use the function *tapply* to obtain the mean of the residuals for each of the fireplace levels. What do you find, are the means closer to or further from zero?
   (d) Is there evidence the fireplace variable is needed in the model? Is there evidence that model 3 is better for explaining the variation in the prices than model 1?
3. Fit a regression model that uses the size, the fireplace variable, and an interaction between them to predict the price, denote this as model 3.
   (a) Note the R^2 adjusted and make sure you can interpret the model coefficients. Does including the interaction increase the R^2 adjusted?
   (b) Make a plot of the price and size that includes the estimated regression model. What do you find, are the lines parallel still?
   (c) Is there evidence the interaction term is needed in the model?
4. Explain what an interaction between size and the fireplace variable means in the context of the problem.

Applied Regression and Time Series Analysis
Live Session 5

Paul Laskowski and Jeffrey Yau

September 27, 2015

Topics Covered in this Live Session

1. Hypothesis Testing
2. Interpretation of Regression Coefficients when either or both the dependent and independent variables take various functional transformation
3. Effects of Data Scaling and Functional Form Specification
4. The effect of violation of certain CLM assumptions on statistical inference
5. Standard Errors Re-visit: CLM SE vs Robust SE (if time permit)

In this live session, we will use one data set (twoyear.Rdata) to cover many important concepts when building a classical linear regression model and conducting statistical inference in practice. It is worth emphasizing that a classical linear regression model as well as all of the associated statistical inference depends on the validity of the underlying statistical assumptions. The violation of one or more of these assumptions will lead to incorrect statistical inference and estimators that do not have all of the desirable statistical properties. Therefore, when applying statistical model in practice, it is important to be very vigilant about the assumptions from which a statistical model is based.

Return to Junior College and University Education (Revisit)

1. Load the data (twoyear.Rdata)
2. Examine the data (and give a brief description of the data)
3. (Quickly) analyze the variables lwage, jc, univ, and exper.
   - Note: In practice, you should pay a lot of attention in the exploratory data analysis phrase. However, in live sessions, we do not have enough time for such an in-depth analysis, and that's why we only ask for a quick analysis to remind you of this step in building a regression model).
   - Please do not hesitate to use histogram, non-parametric kernel density, scatterplot, scatterplot matrix, tabulation, cross-tabulation when examining both the dependent and independent variables of interest.
4. Determine the sample you would like to use to estimate a linear regression model. That is, are there observations you would like to exclude (or alter) before proceeding?
5. Estimate the following linear regression mode

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

6. Interpret the coefficient estimates (both in terms of *wage* and *log wage*) associated with each of the explanatory variables, including the intercept? Does the intercept of this regression, given this data set, have a meaningful interpretation?
7. Does an increase in one year of university education increase with the return to working experience (where *return* is interpreted as increase in log wage)?
8. Test that the return to junior college education is zero, where return is measured in (an increase in) *log(wage)*. (Note: when we say test, we mean "hypothesis testing" or "statistically test".)
9. Test that the return to junior college education (measured in terms of percentage increase in wage for each year increase in junior college) is 5%.
10. Test that the return to junior college education is equal to the return to university education. Put it differently, test the hypothesis that whether one year at junior college is worth one year at a university.
11. Test that the return to university education is equal to the return to 1 year of working experience.
12. **Testing Exclusion Restriction:** Test that junior college and working experience has no effect on salary (measured by *log(wage)*), once university education is controlled for.
    - Note: The reason we keep writing the hypothesis testing in plain English term is that in practice, people who work with data scientists may not have training in statistics and they generally do not ask you "please test if $\beta_i = \beta_j = \beta_k = 0$ or something like that". One of the skill we would like to you master in this course is to translate back-and-forth mathematical formulation of a statistical question and a real world question.

- Review the **F-statistics** and make sure you understand every single term comprising the **F-statistics**.
- Compute the **p-value** associated with the **F-statistic** in this test

13. Test the overall significance of this regression

14. Including a square term of working experience, estimate the linear regression model again:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + \beta_4 exper^2 u$$

15. Test that working experience has no effect on salary.

16. Suppose that the homoskedasticity assumption does not hold. Under this regression above, how does it affect the testing of no effect of university education on salary change?