BTRY 6020 Lab IV

February 27, 2017

Question 1: Multiple Linear Regression Example

In this example we look at how a fireplace is related to the selling price of a home. Specifically, can we quantify the monetary value of a fireplace as it results to the selling price of a home. The easy way to compare the value of a fireplace is to do a 2-sample t-test of the selling price of homes with and without fireplaces.

The data appears in the `Lab4q1Dat.xlsx` file. `Value` is the selling price of the house in thousands of dollars, and `Size` is the square footage of the apartment in thousands.

A) Plot the data, showing the relationship of `Size` and `Value` where `Firepl` data is also encoded. Be sure to include a legend with your plot by using the `legend` function.

B) Compute a 95% confidence interval for this difference *without controlling for the effect of size of home* using the `t.test` function while assuming equal variances. What is your 95% confidence interval for the difference in the selling price of homes with and without fireplaces.
Below we load the data and obtain the confidence interval.

C) Now do a simple linear regression using ONLY an indicator variable for FirePlace. Be sure to keep in mind the dummy coding of the categorical variable. What does the coefficient of Fireplace mean? (hint: the summary table of the linear regression can provide insight here) From this create a 95% confidence interval for the value of a fireplace WITHOUT controlling for hte effect of size of house.

D) Include the Size variable in the data in the linear model, and compute a 95% confidence interval. Note the discreprancy in parts B and C to the confidence interval computere here. What causes this to happen? You should answer this according to two items.

   i) look at the descriptive statistics for size of homes with and without fireaplces;

   ii) look at the estimated variance for these two different procedures (really, two different models). Why does this difference exist? How does this difference impact your confidence intervals?

E) In light of part D)i) above, explain again in a few sentences how multiple regression controls for the effect of one variable before evaluating the effects of another. Also, explain why adding significant controlling variables makes your estimates more precise in light of part D)ii) above.

Question 2: Interaction Example

A developer working in the Midwest and South is trying to predict selling price based on type of home (Single family (SF) or Townhoue (T)), the region built (South (S) or Midwest(M)), and the cost of the lot (which is pro-rated for the number of townhouses built on the lot). He randomly selects 167 homes from the 987 that he has built over the last 10 years, and adjusts the selling price for inflation. Data appears in `Lab4q2Dat.xlsx`. Unless otherwise specified, use $\alpha = 0.05$ or a 95% confidence interval.

A) Plot the data of Lot Cost against Selling Price. Include all relevant categories with a legend.

B) Run a multiple linear regrssion on selling price versus the three predictor variables Region, Type, and Lot Cost. For Region and Type, what are the baseline categories?

C) Look at diagnostic plots. What must be done before you proceed?

D) Fit a full interaction model with all first-order pairwise interactions. Check diagnostic plots including Cook's distance plot and determine if asumptions are met for the inference on this model. The builder would really like to simplify his model. Using this one multiple regression, test if all the interactions can be simultaneously dropped from the model. State hypotheses, test statistic, p-value, and conclusions.

E) Look at the significance of the interaction terms from your six-predictor model in Part C. Test the set of those which are not significant by themselves at $\alpha = 0.05$ with a simultaneous test by re-running the regression without these non-significant interaction terms and getting the test statistic by subtracting the SSRs from the two models.

F) Based on your results in parts D and E, determine an appropriate model (dropping sets of non-significant predictors) and use it to predict the selling price of a single family home in the south on a lot which cost \$42,500.

G) Based on the model you chose in E, does the region of the country have a fixed effect on selling price or does it depend on the other two variables? Explain in two sentences or less.

H) Based on the model you chose in E, does th type of house have a fixed effect on selling price or does it depend on the other two variables? Explain in two sentences of less.

I) Is the increase in selling price per dollar increase in lot cost greater in the Midwest than in the South? State hypotheses, test statistic, p-value, and conclusion.

J) What propotion of the variance in selling price (untransformed!) does the model you chose in E explain?