Project Module 3 – Due Thursday, Dec 11, 5:00 PM at KMC 8–63

STAT-UB.0103 – Statistics for Business Control and Regression Models

Please attach Module 3 at the end of your writeup for Modules 1 and 2.

Try to keep the writeup for this module to four pages.

In the following questions, "response variable" and "predictor variables" refer either to the original variables you collected, or to the logs, according to the decision you made in Module 1, part 5. So the decisions about taking logs (or not) have already been made.

Relate your findings to the things you said you hoped to learn in Module 1 as follows:

1. First, run simple regressions of your response variable against each of the individual explanatory variables. Interpret the slope coefficients. Determine the p-values for the slopes. Are the slopes statistically significant? Do the slopes agree with the scatterplots you made in Module 1?
2. Next, run a multiple regression, using all of your predictor variables. Are all of the coefficients significant? Which variables (if any) appear to be useless for predicting the response variable? Check the F-statistic. (Interpret, briefly). How is the R^2? Is it appreciably higher than what you got in the simple regressions?
3. Do you find any apparent inconsistencies in the coefficients you get in the full multiple regression model, compared with the coefficients for the corresponding variable in the simple regression? Did the coefficient values change appreciably from the simple model to the full model? Discuss, briefly.
4. For the full multiple regression model, get Cook's Distance and leverage, as well as residual vs. fits plot. Briefly discuss the results. (In multiple regression, the leverage is large if it exceeds $2(k+1)/n$, where k is the number of explanatory variables, and Cook's Distance is large if it exceeds 1). Identify any outliers or influential points, and discuss the meaning of these points , if possible. Do all of these points correspond to the ones found in the scatterplots and descriptive statistics graphs from Module 1? If not, discuss briefly. Overall, considering the R^2, the significance of the individual coefficients, and the Cook's Distance values, does the full model seem to fit well?
5. Based on the residuals vs. fits plot, is there evidence of nonconstant variance? Based on your results on normality of the response variable from Module 2, together with the evidence of the residual plots here, do you think that the Minitab output can be trusted?
6. Finally, we are going to use an "automatic" method for selecting the "best" predictor variables. For each of the models you have fitted in parts 1 and 2, you will use the residual sum of squares SSE to compute a number called AIC. The model with the smallest AIC is the "best." AIC is computed as AIC $= n\log(\text{SSE}/n) + 2(k+1)$, where "log" is the natural log (that is, "ln" on most calculators), n is the number of observations, and k is the number of

predictor variables in the model. If any of the AIC values are negative, then the most negative value is the “best”. Determine which of your possible models is “best” according to AIC. Are all of the coefficients in this model statistically significant? Interpret the coefficients of this “best” model, and say what it means in terms of the things you said you wanted to learn in Module 1, part 1. Please repeat this question utilizing the adjusted R-squares measure. Do your answers differ?