INFO 561 Team Projects on Regression Model Building

Each Excel project file has two tabs – DATA and Variable INFO. The DATA worksheet contains a YELLOW column representing the numerical response variable for simple and multiple regression analysis. In addition, there is a BROWN column for the numerical predictor variable for simple (and then again for multiple) regression analysis and the BLUE column for the "dummy" categorical predictor variable in a multiple regression analysis. All other WHITE column numerical variables are potential predictor variables in the multiple regression analysis.

> *Note: For those wishing to practice logistic regression modeling after Module 7, the YELLOW-highlighted categorical variable column is the response variable while the BROWN-highlighted numerical variable column is joined with all other WHITE column numerical variables along with the BLUE column for the "dummy" categorical variable as the potential set of predictors.*

Developing the Team Project Report

To develop your team project report on regression modeling do the following:

1. Using the YELLOW-highlighted numerical variable column in your Excel worksheet as the response variable of interest, create an introductory "scenario" of just two to three sentences that describes the data file for your project and why you (the ?????? Corporation/Group) are performing a multiple regression analysis that will enable you to predict the outcomes of this numerical response variable based on the set of possible independent variables $X_1, X_{2,}..., X_k$. One of these independent variables should be a two-category "dummy" variable – the BLUE-highlighted categorical variable column in your Excel worksheet. Moreover, the BROWN-highlighted numerical variable column will be the predictor variable for a simple linear regression analysis.
2. Next, copy this Excel worksheet into a Minitab worksheet.
3. As you learned in Module 2, first develop a simple linear regression model using the numerical variable whose Excel worksheet column was highlighted in BROWN as the predictor of Y.
   A. Cut and paste into your report the scatter plot and the Minitab printout for this simple linear regression model.
   B. Write the sample regression equation.
   C. Interpret the meaning of the Y intercept and slope for your fitted model.
   D. Interpret the meaning of the coefficient of determination r^2.
   E. Interpret the meaning of the standard error of the estimate S_{YX}.
   F. Obtain the residual plots and cut and paste them into the report. Briefly comment on the appropriateness of your fitted model.
      (1) If the assumptions are met and the fitted model is appropriate continue to Step 3G.

(2) If any of the linearity, normality, or equality of variance assumptions are problematic state this but continue to Step 3G with caution. Note -- you do not need to check the assumption of independence in your project. (That assumption is automatically met because your project is not time-dependent).

G. Comment on the statistical significance of your fitted model

H. Select a value for your independent variable in its relevant range:
   (1) Predict $\hat{Y}$.
   (2) Determine the 95% confidence interval estimate of the average value of Y for all occasions when the independent variable has the particular value you selected.
   (3) Determine the 95% prediction interval estimate of Y for an individual occasion when the independent variable has the particular value you selected.

4. As you learned in Modules 3 and 4, you will be using the set of potentially meaningful numerical independent variables and the one selected "two-category" dummy variable in your study to develop a "best" multiple regression model for predicting your numerical dependent variable Y. Follow the "9-step modeling process" described in the PowerPoints at the end of Module 4.

   A. Start with a visual assessment of the possible relationships of your numerical dependent variable Y with each potential predictor variable by developing the scatterplot matrix and paste this into your report.
   B. Then fit a preliminary multiple regression model using these potential numerical predictor variables and, at most, one categorical dummy variable.
   C. Then assess collinearity until you are satisfied that you have a final set of possible predictors that are "independent," i.e., not unduly correlated with each other.
   D. Use both stepwise regression approaches and best subsets regression approaches to fit a multiple regression model with this set of potentially meaningful numerical independent variables (and, if appropriate, the one selected categorical dummy variable).
      (1) Based on the stepwise modeling criterion determine which independent variables should be included in your regression model.
      (2) Based on the forward selection modeling criterion determine which independent variables should be included in your regression model.
      (3) Based on the backward elimination modeling criterion determine which independent variables should be included in your regression model.
      (4) Based on the adjusted r^2 criterion determine which independent variables should be included in your regression model.
      (5) Based on Minitab's "predicted" r^2 criterion determine which independent variables should be included in your regression model.

(6) Based on the smallest S_{YX} criterion determine which independent variables should be included in your regression model.

(7) Based on Mallows' C_p criterion determine which independent variables should be included in your regression model.

E. Comment on the consistency of your findings in Step 4D (1)-(7).

F. Cut and paste the Minitab printouts from Step 4D into your report.

G. Based on Step 4D (along with the principle of parsimony if necessary) select a "best" multiple regression model.

H. Using the predictor variables from your selected "best" multiple regression model, rerun the multiple regression model in order to assess its assumptions.

I. Look at the set of residual plots, cut and pasted them into the report, and briefly comment on the appropriateness of your fitted model.

(1) If the assumptions are met and the fitted model is appropriate continue to Step 4J.

(2) If the normality assumption is problematic state this but continue to Step 4J with caution because your sample size is large enough for the central limit theorem to enable the use of classical inferential methods. Note: You do not need to check the assumption of independence in your project. That assumption is met because your project is not time-dependent.

(3) If the either the linearity or equality of variance assumption is violated in each plot of Y with the individual predictors $X_1, X_2, ..., X_k$ then transform the dependent variable Y (likely to log Y) and rerun the multiple regression model as in Step 4H.

(4) If either the linearity or equality of variance assumption is violated in one or two scatter plots of Y with individual predictors then transform the particular independent variables involved following Tukey's "ladder of powers" and rerun the multiple regression model as in Step 4H.

J. Assess the significance of the overall fitted model.

K. Assess the contribution of each predictor variable.

L. If the dummy variable is not a significant predictor go to Step 5; however, if the dummy variable is a significant predictor, develop an interaction term for it in combination with every other significant predictor and then rerun the multiple regression model to determine whether any interaction term significantly belongs in the final model and comment on your findings.

5. Write the sample multiple regression equation for the "final best" model you have developed.

A. Interpret the meaning of the Y intercept and interpret the meaning of all the slopes for your fitted model (but do this in whatever units you used for *Y* to build this model).

B. Interpret the meaning of the coefficient of multiple determination r^2.

C. Very briefly comment on how much r^2 has changed from the simple regression model in Step 3D to the "final" multiple regression model in Step 5B.
D. Interpret the meaning of the standard error of the estimate S_{YX} (in the units you used to build this model).
E. Select one value for each of your independent variables in their respective relevant ranges:
   (1) Predict $\hat{Y}$. (If you used log *Y* take the antilog of the predicted value so you are back in units of *Y*).
   (2) Determine the 95% confidence interval estimate of the average value of Y for all occasions when the independent variables have the particular values you selected. (If your lower and upper boundaries are in units of log Y convert back to Y by taking the antilogs).
   (3) Determine the 95% prediction interval estimate of Y for an individual occasion when the independent variables have the particular values you selected. (If your lower and upper boundaries are in units of log Y convert back to Y by taking the antilogs).
F. For your "final best" model, as per Module 1, prepare a brief descriptive analysis highlighting the key measures of central tendency, variation, and shape for your dependent variable Y and for each of the predictor variables. Show the individual histograms and boxplots for these variables. If a dummy variable was included as a predictor in your "final best" model show its summary table and bar chart.

Specific instructions for the written team project report follows.

Writing the Team Report

Each team report has a title page followed by an Introduction section describing the study "scenario" and mentioning the possible predictor variables and the dependent variable. A section on the Simple Linear Regression Model is then followed by a section on the "final best" Multiple Regression Model. The final section of the report is a Discussion section assessing the gains (if any) by using the "best" multiple regression model in lieu of the simple linear regression model. In addition, a short descriptive analysis of the response variable and the predictor variables included in the "best" multiple regression model should be provided in the Discussion section with tables and charts relegated to the Appendix of the report. Note: All discussed Minitab printouts pertaining to regression modeling should be "cut and pasted" into the report. These should be placed either in the body of the report or in an Appendix to the report. If the latter approach is taken, be sure to number and reference these printouts when discussing them in the body of the report.