HW 9

Before running the multiple regressions in Minitab, click on Results, and uncheck the box for Fits and Diagnostics, but make sure that the box for Analysis of Variance is checked. The analysis of variance output will contain several redundant rows and there is currently no option to suppress them automatically. In class handouts (and in exams) I manually delete the rows for the individual explanatory variables in the Analysis of Variance table before copying and pasting the regression output. Please feel free to do the same. All we need from the table is the rows for Source, Regression, Error and Total. I also delete the word “Adj” from “Adj SS” and “Adj MS” in the Source row, since there is nothing “adjusted” about the sums of squares and mean squares.

1) The file Magazine.MTP contains data on advertising costs and characteristics of magazines. The response variable is PageCost, which represents the cost of a full-page color ad in the magazine. Circ is the circulation of the magazine (in thousands), MedIncome is the median income of the readers, and %Male is the percentage of the readers who are male. The square root of the circulation is given in SqrtCirc.
   A) Run a multiple regression of PageCost on Circ, MedIncome and %Male. Before running it, click on Graphs, and check the box for Residuals versus Fits. Note that the plot shows structure: a generally upward-sloping pattern, with three outliers at the right dragging things down. Identify the three outliers (all of which have a very large circulation).
   B) To investigate further, generate a scatterplot of PageCost versus Circ. Note that the plot is “bunched up” at the left, and “stretched out” at the right, and also a bit curved. In what way do the points identified as outliers in A) deviate from the pattern in the plot here?
   C) To try to improve the linear relationship, let’s try working with the square root of Circulation (SqrtCirc) rather than the circulation itself. Plot PageCost versus SqrtCirc. Based on the plot, explain why it seems more appropriate to use SqrtCirc as an explanatory variable in a linear regression rather than Circ.
   D) Now, run a multiple regression of PageCost on SqrtCirc, MedIncome and %Male. Plot the residuals versus fitted values. Does it look better than in A)? Which coefficients in the regression are statistically significant? Based on the p-values for the regression coefficients, which variables seem to be useless for predicting PageCost?
   E) The F-Statistic and corresponding p-value in the Analysis of Variance part of the output provides a test of the null hypothesis that the regression is useless for predicting Y, i.e., that all regression parameters besides the intercept are zero. Based on the p-value, does the regression seem to be useful for predicting Y? Does this mean that all variables are useful? (Remember your answer to part D).
   F) Re-run the regression for PageCost, this time with just the two explanatory variables SqrtCirc and MedIncome. Are the coefficients of both variables statistically significant? Get a 95% confidence interval for the mean page cost of a magazine with a SqrtCirc of 100, and a median income of $40,000. To do this, after running the regression click on Stat → Regression →
Regression → Predict. Then enter 100 in the first line under SqrtCirc and enter 40000 in the first line under MedIncome. Did the R-square go down by much compared to the regression in D)? Is the F-Statistic still significant? What does this suggest about the deletion of the %Male variable?

G) Run a simple regression and scatterplot of PageCost versus MedIncome. Is the coefficient of MedIncome now statistically significant? Why is this puzzling in view of the regression output in F)? Remember, however, that the meaning and interpretation of the coefficient of a given variable depend on what other variables are included in the model.

2) Consider DiamondPrices.MTP, which you already studied in the context of simple regression. Clarity, Code and Color are categorical variables, which are numerically coded in Clarity Code, Color Code and Cut Code. (For example, Good, Very Good and Ideal cuts receive Cut Codes of 1,2,3, respectively.) Even though these are categorical variables, please enter them in Minitab as “Continuous Predictors”, since we will treat them as numerical variables.

A) Run a multiple regression of Price on Clarity Code and Cut Code. Based on the output, do these explanatory variables seem useful for predicting Price?

B) Run a multiple regression of Price on Clarity Code, Color Code and Cut Code. In what way do p-values for the individual coefficients and the F-statistic seem to provide contradictory evidence on whether any of these variables is helpful for predicting Price in the given model? (As we will discuss in class, the F-statistic is the best place to go to make a decision on this question.) Does the R-squared in this regression suggest that these three variables by themselves are very useful for predicting Price?

C) Next, run a multiple regression of Price on Carats, Clarity Code, Color Code and Cut Code. What happened to the variables that had insignificant coefficients in the regression in B)? And what happened to the R-squared?

D) For the regression in Part C), do the p-values for the individual coefficients and the F-statistic have any apparent contradiction?

E) Using the regression model in Part C), predict the retail price of a diamond weighing 0.6 Carats with a VS2 clarity (Clarity Code=3), a G color (Color Code=4) and a Very Good cut (Cut Code=2).

F) Run a simple regression of Price on Carats, and note the R-Squared. Comparing this with the results from C), do you think that the Clarity Code, Color Code and Cut Code (taken together) are worthwhile for predicting Price, above and beyond what can be obtained using Carats alone?

3) For the data in Market.MTP, the variables SMB and HML (Small minus Big, and High minus Low) are “factors” used in addition to the Market return by practitioners in finance for describing the returns on a given stock. Run three regressions: A) IBMRet on MarketRet, B) IBMRet on MarketRet and SMB, C) IBMRet on MarketRet, SMB and HML. Based on the R-squared values and p-values for these three regressions, do SMB and HML seem to add much above and beyond the Market Return for describing IBM returns?