Homework #9 – Due Monday, Dec. 15

COR1-GB.1305 – Statistics and Data Analysis

Problem 1

The file `Magazine.CSV` 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, an d check the box for Residuals plots: Four in one. Note that the residuals versus fit plot shows structure: a generally upward-sloping pattern, with three outliers at the right dragging things down. Identify the Magazines corresponding to 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* ⇒ *Preict*. Then, enter 100 in the first line under `SqrtCirc` and enter 40000 in the first line under `MedIncome`. Did the R^2 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 statisticall 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.

. . . .

Problem 2

Consider DiamondPrices.CSV, which you already studied in the context of simple regression. Clarity, Code and Color are categorical variables, which are numerically coded in ClarityCode, ColorCode and CutCode. (For example, Good, Very Good and Ideal cuts receive Cut Codes of 1, 2, 3, respectively.) Even though these are ordinal/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 ClarityCode and CutCode. Based on the output, do these explanatory variables seem useful for predicting Price?

(b) Run a multiple regression of Price on ClarityCode, ColorCode and CutCode. 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 discussed in class, the F-statistic is the best place to go to make a decision on this question.) Does the R^2 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, ClarityCode, ColorCode and CutCode. What happened to the variables that had insignificant coefficients in the regression in (b)? And what happened to the R^2?

(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 (ClarityCode = 3), a G color (ColorCode = 4) and a Very Good cut (CutCode = 2).

(f) Run a simple regression of Price on Carats, and note the R^2. Comparing this with the results from (c), do you think that the ClarityCode, ColorCode and CutCode (taken together) are worthwhile for predicting Price, above and beyond what can be obtained using Carats alone?

. . . .