Homework #8 – Due Monday, Dec. 8
COR1-GB.1305 – Statistics and Data Analysis

Problem 1

The file `DiamondPrices.CSV` contains data on retail prices (in Dollars) for 617 round shaped diamonds. As the predictor variable, we will focus on Carats, a measure of weight. (One Carat = 200 mg.)

(a) Make a scatterplot of Price versus Carats, and comment on the reasonableness of fitting a linear regression model to this data.

(b) Run the regression of Price on Carats, using *Stat* ⇒ *Regression* ⇒ *Regression* ⇒ *Fit Regression Model*, set *Responses: Price*, and *Continuous Predictors: Carats.* Copy and paste the Minitab regression output for `Model Summary`, `Coefficients`, and `Regression Equation`

(c) What is the equation of the fitted line? Use this equation to predict the price of a diamond ring which weighs 1.25 carats.

(d) Is there evidence of a significant linear relationship between the price and the weight of the diamond? Justify your answer.

(e) Interpret the estimated slope of the fitted model, and construct a 95% confidence interval for the true slope coefficient. What is the practical meaning of the true slope coefficient?

(f) Discuss and give a practical interpretation of the coefficient of determination, R^2.

(g) Does the negative estimated intercept of the fitted model bother you? What is the interpretation of the true intercept?

(h) What is the estimate of the typical fluctuation of data points from the true regression line, measured in the vertical direction?

(i) Using Minitab, construct a 95% confidence interval for the expected price of a ring which weighs 1.25 Carats. (To do this , after running the regression click on *Stat* ⇒ *Regression* ⇒ *Regression* ⇒ *Predict.* Type in 1.25 in the first line under Carats.)

(j) Generate the "Four in one" residual plots for the model: run the command *Stat* ⇒ *Regression* ⇒ *Regression* ⇒ *Fit Regression Model*; then click on the "Graphs" button, then select "Four in one" under "Residual Plots". Based on these graphs, comment on the reasonableness of the four assumptions for the regression errors (mean zero, constant variance, normal, independent).

. . . .

Problem 2

Consider the data in MARKET.CSV.

(a) Construct the fitted line plot for IBMRet versus MarketReturn. Does this suggest a linear relationship between the two variables?

(b) Identify the outlier in the lower left-hand corner of the plot by resting the cursor over the point and then going to the spreadsheet to find the corresponding case.

(c) Run the regression of IBMRet versus MarketReturn. Write the equation for the fitted model. (In finance, this is called the market model.) What is the slope of the fitted line? (In finance, they call this the "beta" for IBM, but actually it's just an estimate of the true slope, β_1.)

(d) Is there strong evidence of a linear relationship between MarketReturn and IBMRet?

(e) Find a 95% confidence interval for the true slope. Does this interval contain the value 1?

(f) In finance, the performance of an investment compared to the market is often measured by the "alpha", which is equal to the estimated intercept, $\hat{\beta}_0$. What was the value of $\hat{\beta}_0$ for IBM? What is the interpretation of this value?

(g) Is there evidence that the true β_0 for IBM is nonzero? What is the relevant p-value? Interpret this p-value.

....

Problem 3

The file PriceToEarnings.CSV contains annual data on the S&P 500 index, returns on the index, defined as

$$\frac{\text{(This year's value)} - \text{(Last Year's Value)}}{\text{(Last Year's Value)}},$$

as well as several variables that may be useful for forecasting the index or its returns. We will focus on the LTM P/E Ratio (the Last Twelve Months Price to Earnings ratio), and the Dividend Yield. Since these variables cover the same time span as the S&P, they must be lagged before they can be considered as predictor variables. The lagged versions are in Lag PE and Lag Dividend Yield. These are the previous year's values. We start by trying to predict S&P 500.

(a) Construct a scatterplot of S&P 500 versus Lag PE. Does it suggest a linear relationship? If so, is it a positive or a negative relationship?

(b) Run the simple regression of S&P 500 (Response) versus Lag PE (Continuous Predictor). Before running it, click on "Results" and uncheck "Fits and diagnostics". Also, click on "Graphs" and select "Four in One". Copy and paste the regression output. Does the regression output (the table of estimated coefficients) suggest that Lag PE is a good predictor of S&P 500? Why or why not?

(c) Based on the residual plots, comment on the validity of the four regression model assumptions. (Note: "Order" in the Residuals versus Order plot refers to the order in which the data were listed. Since `S&P 500` is an annual time series, this plot provides the residuals versus year.)

(d) To try to fix the problems in (c), let's try predicting the `S&P Returns` instead of the index itself. Construct a scatterplot of `S&P Returns` versus `Lag PE`, run the regression and generate the four residual plots. How do the plots compare with the ones you generated earlier? Based on these plots, do you feel relatively comfortable with th e linear regression model for `S&P Returns` versus `Lag PE`?

(e) Construct the fitted line plot for `S&P Returns` versus `Lag PE`. Based on this and the regression output, does there seem to be a relationship between the two variables? Is the relationship statistically significant? What does the R^2 say about the strength of the linear relationship?

. . . .