HW 8

1) The file DiamondPrices 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, Responses: Price, Continuous Predictors: Carats. Before running the regression, click on Graphs and un-check the box for Pareto. Next, click on Results, select Display of Results: Basic tables, and un-check the boxes for Analysis of Variance and for Fits and Diagnostics. After running the regression (by clicking OK), click on the down arrow in the upper right of the Minitab window to Copy and paste all of the Minitab regression output: the Regression Equation, the Coefficients and the Model Summary.

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% CI 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-Squared.

G) Does the negative estimated intercept of the fitted model bother you? What is the interpretation of the true intercept? Is there evidence at the 1% level of significance to suggest that the true intercept is negative?

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

I) At the 1% level of significance, can we reject the null hypothesis that the true slope is 8000 in favor of the alternative that it is not 8000?

J) 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.)

2) Consider the data in Market.

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

B) Identify the outlier on the far left 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, $\beta_1$.)
D) Is there strong evidence of a positive linear relationship between MarketReturn and IBMRet?

E) Is there evidence that the true beta for IBM is different from 1? Obtain the p-value for a null hypothesis that the true slope is 1. (You must do this by hand, since the null hypothesis in the Minitab table is that the true slope is zero.)

F) In finance, the performance of an investment compared to the market is often measured by the “alpha” (actually $\hat{\alpha}$). What was the value of $\hat{\alpha}$ for IBMRet? What is the interpretation of this value?

G) Is there evidence that the true $\alpha$ for IBMRet is negative? What is the relevant p-value? Interpret this p-value.

3) The file PriceToEarnings contains annual data on the S&P 500 index, returns on the index, defined as 
   \[(\text{This year’s value} - \text{Last Year’s Value})/ \text{Last Year’s Value}\], 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 Graphs, and check the boxes for Residuals versus Fits and Residuals versus Order. 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 two plots, identify two problems, both of which suggest that the basic assumptions for the linear regression model are violated, and therefore that the output (especially, the standard errors and p-values) is invalid. (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 plots of Residuals versus Fits and Residuals versus Order. How do the plots compare with the ones you generated earlier? Based on these plots, do you feel relatively comfortable with the 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? Which is most appropriate: a one-tailed, left-tailed or right-tailed alternative hypothesis? Why? What Does the $R^2$ say about the strength of the linear relationship?

F) Repeat parts D) and E) using Lag Dividend Yield as the predictor, instead of Lag PE.