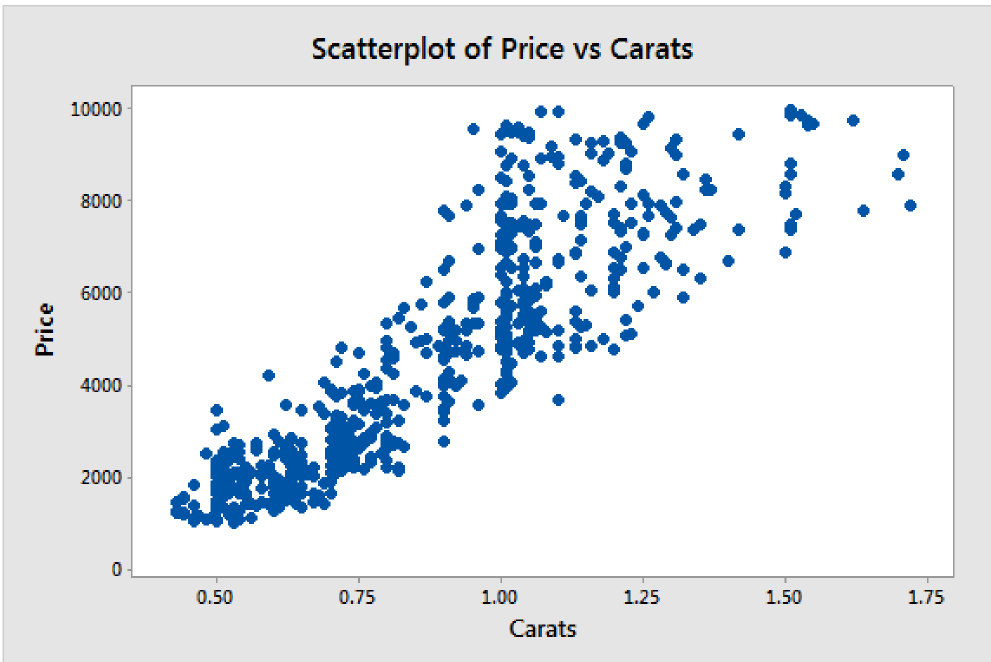
Homework #8 – Solutions
COR1-GB.1305 – Statistics and Data Analysis

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.*



Though not a perfect fit, it seems reasonable to use a linear regression model for 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`***

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
1234.71  75.96%     75.92%      75.80%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -2541      169   -15.04    0.000
Carats     8130      184    44.08    0.000  1.00


Regression Equation

Price = -2541 + 8130 Carats
```

(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.*

The fitted line is

$$\widehat{\text{Price}} = -2541 + 8130 \text{ Carats}.$$

The predicted price (in dollars) of a diamond right weighing 1.25 carats is

$$\widehat{\text{Price}}(1.25) = -2541 + 8130(1.25) = 7621.50.$$

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

Yes, there is a significant relationship. The p-value for testing the null hypothesis $H_0 : \beta_1 = 0$ is reported as 0.000, which means that it is less than 0.001, and certainly less than the threshold for statistical significance (0.05).

(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?*

The estimated slope in the fitted model is 8130. This means that in the fitted model, an increase in weight of 1 carat is associated with an increase in expected price of \$8130.

An approximate 95% confidence interval for the true slope coefficient is

$$\begin{aligned} \hat{\beta}_1 \pm 2\text{se}(\hat{\beta}_1) &= 8130 \pm 2(184) \\ &= 8130 \pm 368 \\ &= (7762, 8498). \end{aligned}$$

(A more exact interval would use $t_{.025,615} \approx 1.96$ instead of 2.)

The practical meaning of the true slope coefficient is that this is the *true* increase in expected price associated with a 1 carat increase in weight; that is, this is the increase in the population of all diamonds, not just the diamonds in the sample.

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

The coefficient of determination is $R^2 = 75.96\%$. The fitted regression model explains 75.96% of the variability in the data. This is a strong linear relationship.

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

No, this does not bother me. The true intercept, taken literally, is the expected price of a diamond with weight equal to zero. Since zero is outside the range of the data, this quantity is not directly interpretable.

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

The standard error of the regression is $s = 1234.71$. According to the model, approximately 95% of the data points are within $2s = 2469.42$ of the regression line.

(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.)*

Here is the Minitab output:

```
Regression Equation

Price = -2541 +8130Carats


Variable  Setting
Carats       1.25


    Fit   SE Fit         95% CI               95% PI
7621.51  85.0583  (7454.47, 7788.55)  (5191.00, 10052.0)
```

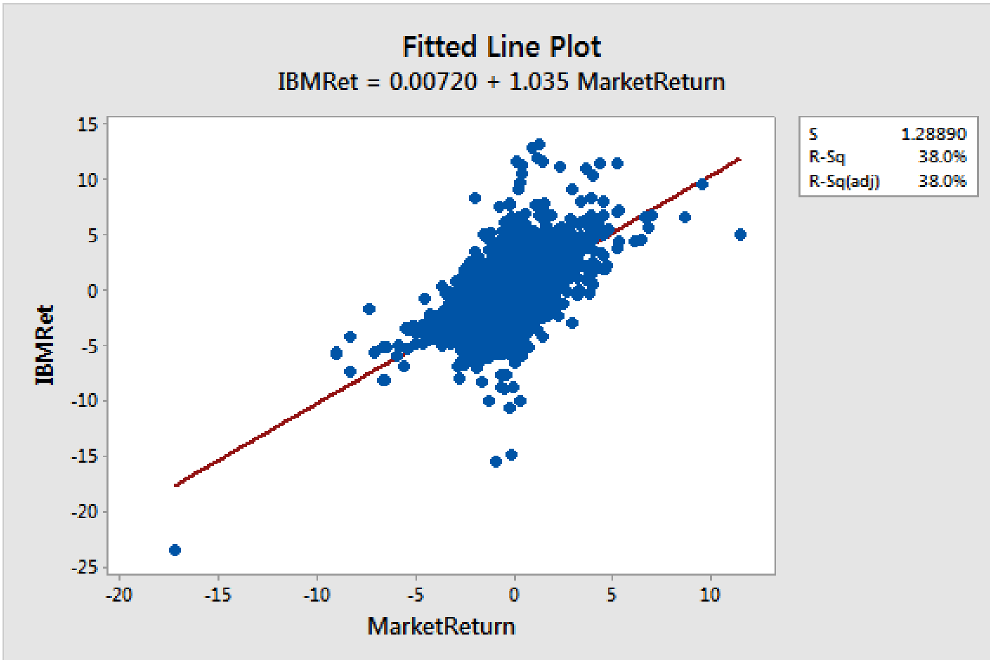
The 95% confidence interval for the expected weight is $(7454.47, 7788.55)$.

....

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?*



A linear relationship seems reasonable.

(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.*

The point is for October 19, 1987. This is "Black Monday" (see `http://en.wikipedia.org/wiki/Black_Monday_(1987)`).

(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.)*

Here is the Minitab output:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
1.28890  37.98%     37.97%      37.94%


Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant      0.0072   0.0120     0.60    0.549
MarketReturn  1.0352   0.0123    84.04    0.000  1.00


Regression Equation

IBMRet = 0.0072 + 1.0352 MarketReturn
```

The fitted model is

$$\widehat{\text{IBMRet}} = 0.0072 + 1.0352 \text{ MarketReturn}.$$

The slope of the fitted line is 1.0352.

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

Yes. The p-value for testing the null hypothesis $H_0 : \beta_1 = 0$ is reported as 0.000, which means that it is less than 0.001. There is very strong evidence of a linear relationship.

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

An approximate 95% confidence interval for β_1 is

$$\begin{aligned} \hat{\beta}_1 \pm 2\mathrm{se}(\hat{\beta}_1) &= 1.0352 \pm 2(0.0123) \\ &= 1.0352 \pm 0.0246 \\ &= (1.0106, 1.0598). \end{aligned}$$

The interval does not 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?*

The value of $\hat{\beta}_0$ is 0.0072. In the fitted model, this is the expected return for IBM when the return for the market is zero.

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

There is no evidence that the true β_0 for IBM is nonzero. The relevant p-value is 0.549. If the true β_0 were zero, then where would be a 54.9% chance of seeing data like observed.

. . . .