STAT 463 - Assignment 5

Due Date: Wednesday, April 19th, 2023

Data Description and Background:
This assignment will focus on just two of these variables, the diamond price (measured in Singapore Dollars) and the clarity rating. A diamond's clarity rating is an ordinal variable that indicates the pristineness of the carbon lattice structure for that diamond.
The diamond clarity ratings that appear in this dataset are (in order from best to work):
IF–Internally Flawless
VVS1–Very Very Slightly Included 1
VVS2–Very Very Slightly Included 2
VS1–Very Slightly Included 1
VS2–Very Slightly Included 2

1) Fit a linear model where the response variable is the diamond price and the explanatory variable is the clarity rating. Write out the model equation for the diamond price. – 3 points

```
> diamond.lm<-lm(diamond$PRICE ~ diamond$CLARITY)
> summary(diamond.lm)

Call:
lm(formula = diamond$PRICE ~ diamond$CLARITY)

Residuals:
    Min      1Q  Median      3Q     Max 
-5220.2 -1940.0  -990.9  2063.5 11218.2 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2694.8      494.2   5.453 1.03e-07 ***
diamond$CLARITYVS1     2362.3      613.9   3.848 0.000145 ***
diamond$CLARITYVS2     3163.4      668.5   4.732 3.42e-06 ***
diamond$CLARITYVVS1    2872.9      671.4   4.279 2.52e-05 ***
diamond$CLARITYVVS2    2661.8      618.0   4.307 2.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3278 on 303 degrees of freedom
Multiple R-squared:  0.08428,	Adjusted R-squared:  0.0722 
F-statistic: 6.972 on 4 and 303 DF,  p-value: 2.216e-05
```

Diamond Price = 2694.8 + 2362.3VS1 + 3163.4VS2 + 2872.9VVS1 + 2661.8VVS2

2) Provide an estimate of the mean price of a diamond within each clarity rating level.

```
> mprice<-with(diamond, tapply(diamond$PRICE, diamond$CLARITY, mean))
> mprice
      IF       VS1       VS2      VVS1      VVS2
2694.773  5057.037  5858.170  5567.635  5356.551
```

** It was confusing why the IF diamonds had a smaller mean price than the other clarities but after looking at the data for more context I saw that the weights of the IF diamonds were smaller so it makes sense that by carat IF diamonds are more expensive but by price they are cheaper because IF diamonds are smaller.

3) It is desired to know whether there is a statistically significant relationship between diamond price and clarity rating. What is the method one should use to answer this question? What conclusion would one come to? – 1.5 points

We perform ANOVA to answer whether there is a statistically significant relationship between diamond price and clarity rating. We accomplish this by using the anova() on the lm object.

```
> d.anova<-anova(diamond.lm)
> d.anova
Analysis of Variance Table
Response: diamond$PRICE
                  Df     Sum Sq   Mean Sq  F value      Pr(>F)
diamond$CLARITY    4  299668847  74917212   6.9722  2.216e-05 ***
Residuals        303 3255758499  10745078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0 : there is no difference in means therefore there is no statistically significant relationship between diamond price and clarity rating.

H_a : the means are not all equal therefore there is a statistically significant relationship between diamond price and clarity rating.

The F statistic is 6.9722 and the corresponding P value is 2.216e-05 making it smaller than the alpha level of 0.05 leading us to reject the null hypothesis and conclude that yes, there is a statistically significant relationship between diamond price and clarity rating.

4)determine for which clarity ratings the difference in mean price is statistically different from one another (for example, is the difference in mean price of IF rated diamonds and mean price of VVS1 rated diamonds statistically significant?). Determine the previous in a manner that ensures the family wise error rate is at most 0.1. – 4 points

Inorder to determine which clarity ratings the difference in mean price is statistically different from one another, we must conduct a **pairwise t test**

To ensure the family wise error rate $(1 - (1-\alpha)^N)$ is at most 0.1, we set alpha level to 0.01 because it satisfies the inequality $0.1>1 - (1-0.01)^5 =0.049$

Where N is the number of comparisons and we are making 5 comparisons

```
> pairwise.t.test(diamond$PRICE, diamond$CLARITY, p.adjust.method = "none",
conf.level=0.99)

        Pairwise comparisons using t tests with pooled SD

data:  diamond$PRICE and diamond$CLARITY

     IF      VS1     VS2     VVS1
VS1  0.00015 -       -       -
VS2  3.4e-06 0.16758 -       -
VVS1 2.5e-05 0.38141 0.65009 -
VVS2 2.2e-05 0.56506 0.39067 0.71933

P value adjustment method: none
```

The differences in mean price between IF and each of the other clarity ratings (VS1, VS2, VVS1, and VVS2) are statistically significant, with p-values less than the alpha value of 0.01.

5) Construct two-sided confidence intervals for the difference in mean diamond price for each possible pair of clarity ratings at confidence level of 99%. How many of these intervals contain 0? Do you notice any similarities between the intervals that cover 0 and the results of part 4?

Explanation of the following commands and how to interpret results:
The linear model we fit in problem 1 used IF as the reference so when we take the confidence interval of that linear model, we are only getting the two-sided confidence intervals for the difference in mean diamond price between IF and each of the other four clarity ratings (VS1, VS2, VVS1, and VVS2). To construct the two-sided confidence intervals for the difference in mean diamond price for each possible pair of clarity ratings we have to **change the reference** when we run the lm() function. We ignore the intercept value and look to the first column as the labels for the pairs. For example, the row labeled IF..VS1 holds the values of the two-sided C.I. for the difference in mean diamond price between IF and VS1 rating.

The intervals highlighted blue contain 0 while the ones highlighted yellow do not contain 0.

```
> IF..<-relevel(as.factor(diamond$CLARITY), ref = "IF")
> changeref = lm(diamond$PRICE ~ IF..)
> confint(changeref, level = 0.99)

                0.5 %   99.5 %
(Intercept) 1413.8028 3975.743
IF..VS1      770.9666 3953.562
IF..VS2     1430.4437 4896.350
IF..VVS1    1132.3665 4613.357
IF..VVS2    1059.7452 4263.812
```

```
> VS1..<-relevel(as.factor(diamond$CLARITY), ref = "VS1")
> changeref = lm(diamond$PRICE ~ VS1..)
> confint(changeref, level = 0.99)

                  0.5 %    99.5 %
(Intercept)  4112.9267 6001.1474
VS1..IF     -3953.5620 -770.9666
VS1..VS2     -700.0626 2302.3282
VS1..VVS1    -999.2980 2020.4932
VS1..VVS2   -1048.4365 1647.4650

> VS2..<-relevel(as.factor(diamond$CLARITY), ref = "VS2")
> changeref = lm(diamond$PRICE ~ VS2..)
> confint(changeref, level = 0.99)

                 0.5 %     99.5 %
(Intercept)  4691.018  7025.3216
VS2..IF     -4896.350 -1430.4437
VS2..VS1    -2302.328   700.0626
VS2..VVS1   -1949.054  1367.9833
VS2..VVS2   -2014.189  1010.9522

> VVS1..<-relevel(as.factor(diamond$CLARITY), ref = "VVS1")
> changeref = lm(diamond$PRICE ~ VVS1..)
> confint(changeref, level = 0.99)

                 0.5 %    99.5 %
(Intercept)  4389.314  6745.956
VVS1..IF    -4613.357 -1132.366
VVS1..VS1   -2020.493   999.298
VVS1..VS2   -1367.983  1949.054
VVS1..VVS2  -1732.289  1310.122

> VVS2..<-relevel(as.factor(diamond$CLARITY), ref = "VVS2")
> changeref = lm(diamond$PRICE ~ VVS2..)
> confint(changeref, level = 0.99)

                 0.5 %    99.5 %
(Intercept)  4394.456  6318.646
VVS2..IF    -4263.812 -1059.745
VVS2..VS1   -1647.465  1048.437
VVS2..VS2   -1010.952  2014.189
VVS2..VVS1  -1310.122  1732.289
```

As seen from my highlights above, **6** confidence intervals contain 0. Although 12 are highlighted, the pairings become redundant. So **6** pairs of clarity ratings whose C.I. for difference in means were constructed have intervals that contain 0 and are listed below: VS1..VS2, VS1..VVS1, VS1..VVS2, VS2..VVS1, VS2..VVS2, VVS1..VVS2

Through the same deduction, **4** pairs of clarity ratings have confidence intervals for difference in means that do not contain 0 and are listed as follows: IF..VS1, IF..VS2, IF..VVS1, IF..VVS2

The pairs with confidence intervals that do not contain 0 are identical to the results of part 4 where we determined for which clarity ratings the difference in mean price is statistically different from one another. We found that the differences in mean price between IF and each of the other clarity ratings (VS1, VS2, VVS1, and VVS2) are statistically significant which are the same as pairs with intervals that do not contain 0 listed above.