Lab 10

Exercise 1

We will continue using diamonds dataset

```
diamonds <- read.csv("diamonds.csv")

diamonds$price = as.numeric(diamonds$price)

## Select Numeric Columns
diamonds = diamonds[, sapply(diamonds, class) == "numeric"]
head(diamonds)
```

```
##   carat depth table price length.in.mm width.of.mm depth.in.mm
## 1  0.23  61.5    55   326         3.95        3.98        2.43
## 2  0.21  59.8    61   326         3.89        3.84        2.31
## 3  0.23  56.9    65   327         4.05        4.07        2.31
## 4  0.29  62.4    58   334         4.20        4.23        2.63
## 5  0.31  63.3    58   335         4.34        4.35        2.75
## 6  0.24  62.8    57   336         3.94        3.96        2.48
```

```
### Fitting a linear model
fit <- lm(price ~ ., data = diamonds)
summary(fit)
```

```
##
## Call:
## lm(formula = price ~ ., data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23878.2   -615.0    -50.7    347.9  12759.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20849.316    447.562  46.584  < 2e-16 ***
## carat        10686.309     63.201 169.085  < 2e-16 ***
## depth         -203.154      5.504 -36.910  < 2e-16 ***
## table         -102.446      3.084 -33.216  < 2e-16 ***
## length.in.mm -1315.668     43.070 -30.547  < 2e-16 ***
## width.of.mm     66.322     25.523   2.599  0.00937 **
## depth.in.mm     41.628     44.305   0.940  0.34744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1497 on 53933 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592
## F-statistic: 5.486e+04 on 6 and 53933 DF,  p-value: < 2.2e-16
```

```
### Getting fitted values
fitted.vals = fit$fitted.values
```

(a)

Calculate the residuals, the residual sum of squares (RSS), and the total sum of squares (TSS) using the `fitted.value()`

```
# Insert you code here
# RSS <-
# TSS <-
```

(b)

Calculate the R-square (R^2) using RSS and TSS. What is the interpretation of R^2?

```
# Insert you code here, save your results as `Rsq`
# Rsq <-
```

(c)

Fit another multivariate model (fit.restricted), but this time, drop `length.in.mm`, `width.of.mm` and `depth.in.mm`. Plot the residuals from this model against the variables added as covariates.

```
# fit.restricted <- lm(price ~)
#summary(fit.restricted)
```

Regression dianosis

Red wine dataset

Reload the data.

```
wine<- read.csv("winequality-red.csv", sep = ";")
wine$quality <- wine$quality + rnorm(length(wine$quality))
```

Fit the model.

```
wine.fit <- lm(quality~volatile.acidity+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+pH+sulphates+
summary(wine.fit)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7716 -0.7499  0.0148  0.7936  5.2557
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.948234   0.744742   6.644 4.17e-11 ***
## volatile.acidity     -0.632463   0.186396  -3.393 0.000708 ***
## chlorides            -1.849943   0.734807  -2.518 0.011914 *
## free.sulfur.dioxide   0.005567   0.003929   1.417 0.156693
## total.sulfur.dioxide -0.004167   0.001269  -3.283 0.001050 **
```

```
## pH                   -0.629553   0.217292  -2.897 0.003816 **
## sulphates             0.984574   0.203152   4.846 1.38e-06 ***
## alcohol               0.261186   0.031045   8.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 1591 degrees of freedom
## Multiple R-squared:  0.1167, Adjusted R-squared:  0.1128
## F-statistic: 30.03 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Exercise 2

(a) Do regression diagnostics using the `plot` function.

```
# insert your code here to do regression diagnostics.
```

(b) Answer the following TRUE/FALSE questions based on the diagnostics plot. Uncomment your answer.

```
### I. The plot indicates heteroscedasticity.
# TRUE
# FALSE
### II. There are non-linearity between the explantory variable and response variable.
# TRUE
# FALSE
### III. The normal assumtion holds for this model.
# TRUE
# FALSE
```

(c) Identify at least two outliers from the data.

I think the sample ??? and ??? are outliers.

Multiple regression with continuous and categorical variables

Exercise 3

(a) Fit a linear regression model with explanatory variable `carat`, `depth`, `table`, `clarity`, `color` and `cut`.

```
diamonds <- read.csv("diamonds.csv")
head(diamonds,2)
```

```
##   carat     cut color clarity depth table price length.in.mm width.of.mm
## 1  0.23   Ideal     E     SI2  61.5    55   326         3.95        3.98
## 2  0.21 Premium     E     SI1  59.8    61   326         3.89        3.84
##   depth.in.mm
## 1        2.43
## 2        2.31
```

```
# Insert you code here, save your model as `fit.categorical`
# fit.categorical <- lm(price ~ .......
```

(b) Write the equation when

i. Clarity is VS2, color is H, and cut is Premium. Replace ??? with numerical values.

$$\text{price} = ??? + ??? \cdot \text{carat} + ??? \cdot \text{depth} + ??? \cdot \text{table}$$

ii. clarity is I1, color is D and cut is Fair. Replace ??? by numerical values.

$$\text{price} = ??? + ??? \cdot \text{carat} + ??? \cdot \text{depth} + ??? \cdot \text{table}$$

Diamond dataset

We will include categorical variables in the following analysis. Read the data.

```
diamonds <- read.csv("diamonds.csv")
diamonds <- diamonds[sample(1:nrow(diamonds), 1000), ]
head(diamonds)
```

```
##       carat     cut color clarity depth table price length.in.mm width.of.mm
## 18047  0.35   Ideal     H     VS1  62.0    56   614         4.54        4.56
## 51655  0.26   Ideal     H    VVS1  62.2    55   545         4.09        4.11
## 16522  1.31 Premium     H     SI2  61.4    59  6602         6.99        6.96
## 12338  1.22 Premium     G     SI2  61.6    62  5226         6.79        6.75
## 12107  1.01   Ideal     H     VS2  62.7    56  5166         6.35        6.40
## 50723  0.55   Ideal     F    VVS2  61.5    57  2294         5.24        5.27
##       depth.in.mm
## 18047        2.82
## 51655        2.55
## 16522        4.28
## 12338        4.17
## 12107        4.00
## 50723        3.23
```

Fit a linear regression.

```
diamond.fit <- lm(price ~ carat + cut + color + clarity + depth + table, data = diamonds)
summary(diamond.fit)
```

```
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
##     table, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5527.5  -682.5  -160.2   496.0  7425.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2535.72    2515.06  -1.008 0.313600
## carat         9047.55      85.30 106.072  < 2e-16 ***
## cutGood        707.08     244.68   2.890 0.003939 **
## cutIdeal       921.85     241.65   3.815 0.000145 ***
## cutPremium     917.03     237.93   3.854 0.000124 ***
## cutVery Good   947.84     237.02   3.999 6.84e-05 ***
## colorE        -370.69     130.75  -2.835 0.004675 **
## colorF        -229.47     131.38  -1.747 0.081009 .
## colorG        -556.51     124.16  -4.482 8.27e-06 ***
## colorH       -1138.73     134.13  -8.490  < 2e-16 ***
## colorI       -1591.92     157.03 -10.138  < 2e-16 ***
## colorJ       -2512.03     187.33 -13.410  < 2e-16 ***
## clarityIF     3813.31     408.35   9.338  < 2e-16 ***
## claritySI1    2294.46     360.82   6.359 3.11e-10 ***
## claritySI2    1479.70     363.15   4.075 4.98e-05 ***
## clarityVS1    3306.24     363.44   9.097  < 2e-16 ***
## clarityVS2    2997.79     360.72   8.311 3.15e-16 ***
```

```
## clarityVVS1   4067.77     379.80  10.710  < 2e-16 ***
## clarityVVS2   3947.43     375.53  10.512  < 2e-16 ***
## depth          -43.20      27.83  -1.552 0.120918
## table          -17.53      20.72  -0.846 0.397816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1102 on 979 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9255
## F-statistic: 621.7 on 20 and 979 DF,  p-value: < 2.2e-16
```

Exercise 4

(a) Do regression diagnostics using the `plot` function.

```
# insert your code here to do regression diagnostics.
```

(b) Answer the following TRUE/FALSE questions based on the diagnostics plot. Uncomment your answer.

```
### I. The plot indicates heteroscedasticity.
# TRUE
# FALSE
### II. There are non-linearity between the explantory variable and response variable.
# TRUE
# FALSE
### III. The normal assumtion holds for this model.
# TRUE
# FALSE
```