Data 621 Blog 2

Bryan Persaud

10/10/2020

Multiple Linear Regression

For my second blog I will continue to demonstrate a linear regression model by showing how to do a multiple linear regression model. A multiple linear regression model is a model that shows the relationship between an dependent variable, y, and one or more independent variables.

Load Dataset

I will be using the diamonds dataset again to show an example on how to create a multiple linear regression model. The diamond dataset is under the ggplot2 library.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

Multiple Linear Regression Model

Use the lm model again to create the model.

```
model <- lm(price ~ carat + cut + color + clarity + depth + table + x + y + z, data = diamonds)
model
```

```
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
##     table + x + y + z, data = diamonds)
##
## Coefficients:
## (Intercept)        carat        cut.L        cut.Q        cut.C        cut^4
##    5753.762    11256.978      584.457     -301.908      148.035      -20.794
##     color.L      color.Q      color.C      color^4      color^5      color^6
##   -1952.160     -672.054     -165.283       38.195      -95.793      -48.466
##   clarity.L    clarity.Q    clarity.C    clarity^4    clarity^5    clarity^6
##    4097.431    -1925.004      982.205     -364.918      233.563        6.883
##   clarity^7        depth        table            x            y            z
##      90.640      -63.806      -26.474    -1008.261        9.609      -50.119
```

Here we see a model created with every variable in the diamond dataset and their corresponding coefficients and intercepts displayed.

```
summary(model)
```

```
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
##     table + x + y + z, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21376.0   -592.4   -183.5    376.4  10694.2
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  5753.762    396.630   14.507  < 2e-16 ***
## carat       11256.978     48.628  231.494  < 2e-16 ***
## cut.L         584.457     22.478   26.001  < 2e-16 ***
## cut.Q        -301.908     17.994  -16.778  < 2e-16 ***
## cut.C         148.035     15.483    9.561  < 2e-16 ***
## cut^4         -20.794     12.377   -1.680  0.09294 .
## color.L     -1952.160     17.342 -112.570  < 2e-16 ***
## color.Q      -672.054     15.777  -42.597  < 2e-16 ***
## color.C      -165.283     14.725  -11.225  < 2e-16 ***
## color^4        38.195     13.527    2.824  0.00475 **
## color^5       -95.793     12.776   -7.498 6.59e-14 ***
## color^6       -48.466     11.614   -4.173 3.01e-05 ***
## clarity.L    4097.431     30.259  135.414  < 2e-16 ***
## clarity.Q   -1925.004     28.227  -68.197  < 2e-16 ***
## clarity.C     982.205     24.152   40.668  < 2e-16 ***
## clarity^4    -364.918     19.285  -18.922  < 2e-16 ***
## clarity^5     233.563     15.752   14.828  < 2e-16 ***
## clarity^6       6.883     13.715    0.502  0.61575
## clarity^7      90.640     12.103    7.489 7.06e-14 ***
## depth         -63.806      4.535  -14.071  < 2e-16 ***
## table         -26.474      2.912   -9.092  < 2e-16 ***
## x           -1008.261     32.898  -30.648  < 2e-16 ***
## y               9.609     19.333    0.497  0.61918
## z             -50.119     33.486   -1.497  0.13448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 53916 degrees of freedom
## Multiple R-squared:  0.9198, Adjusted R-squared:  0.9198
## F-statistic: 2.688e+04 on 23 and 53916 DF,  p-value: < 2.2e-16
```

The summary function is once again used to show additional information of the model.

Let's create a model where we narrow down some of the variables. Let's create a model using price, carat, and depth.

```r
model2 <- lm(price ~ carat + depth, data = diamonds)
model2
```

```
##
## Call:
## lm(formula = price ~ carat + depth, data = diamonds)
##
## Coefficients:
## (Intercept)        carat        depth
##      4045.3       7765.1       -102.2
```

```r
summary(model2)
```

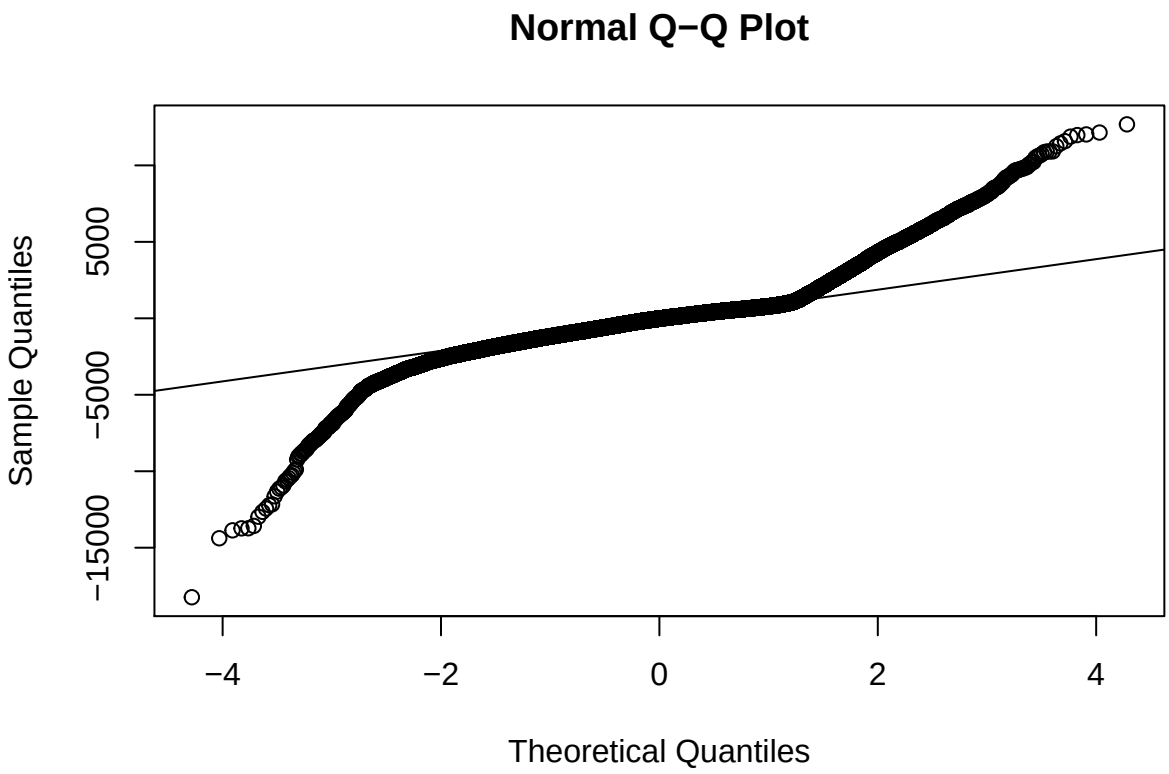
```
##
## Call:
## lm(formula = price ~ carat + depth, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18238.9   -801.6    -19.6    546.3  12683.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4045.333    286.205   14.13   <2e-16 ***
## carat       7765.141     14.009  554.28   <2e-16 ***
## depth       -102.165      4.635  -22.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1542 on 53937 degrees of freedom
## Multiple R-squared:  0.8507, Adjusted R-squared:  0.8507
## F-statistic: 1.536e+05 on 2 and 53937 DF,  p-value: < 2.2e-16
```

This is the information for the model created by the price, carat, and depth variables.

Plot

Let's plot the residuals to check for normality. This is done by using the qqnorm and qqline functions.

```r
qqnorm(model2$residuals)
qqline(model2$residuals)
```



Here we see that a good amount of residuals follow the straight line, but there are a lot that deviate away from the line. We can say that the distribution is nearly normal. The model is an okay fit to the data, but there could be a model that is a better fit. This means that there is not really a strong relationship and having both carat and depth doesn't affect the price of a diamond too much.