Data 621 Blog 1

Bryan Persaud

10/3/2020

Simple Linear Regression

For my first blog I will be demonstrating how to create a simple linear regression model. A linear regression model is a model that shows the relationship between a dependent variable, y, and an independent variable, x.

Load Dataet

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

```r
library(ggplot2)
```

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

```r
summary(diamonds)
```

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

The summary function is used to take a look at the dataset and to see what we are working with.

Simple Linear Regression

Use the lm function to create your regression model.

```
model <- lm(data = diamonds)
model
```

```
##
## Call:
## lm(data = diamonds)
##
## Coefficients:
## (Intercept)        cut.L        cut.Q        cut.C        cut^4      color.L
##  -1.660e+00   -2.057e-02    9.257e-03   -6.879e-03    1.741e-03    1.085e-01
##     color.Q      color.C      color^4      color^5      color^6    clarity.L
##   4.106e-02    5.772e-03   -5.063e-03    5.713e-03    1.959e-03   -1.784e-01
##   clarity.Q    clarity.C    clarity^4    clarity^5    clarity^6    clarity^7
##   1.081e-01   -5.677e-02    1.727e-02   -1.232e-02   -1.793e-03    1.719e-04
##       depth        table        price            x            y            z
##   1.212e-02    2.203e-03    4.428e-05    2.425e-01    5.961e-03    4.586e-03
```

What is displayed is the coefficients and intercepts for each variable in the dataset.

```
summary(model)
```

```
##
## Call:
## lm(data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52251 -0.03060 -0.00102  0.02905  2.17188
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.660e+00  2.388e-02 -69.512  < 2e-16 ***
## cut.L       -2.057e-02  1.416e-03 -14.531  < 2e-16 ***
## cut.Q        9.257e-03  1.131e-03   8.186 2.76e-16 ***
## cut.C       -6.879e-03  9.715e-04  -7.081 1.45e-12 ***
## cut^4        1.741e-03  7.762e-04   2.244  0.02487 *
## color.L      1.085e-01  1.115e-03  97.305  < 2e-16 ***
## color.Q      4.106e-02  9.904e-04  41.461  < 2e-16 ***
## color.C      5.772e-03  9.243e-04   6.245 4.28e-10 ***
## color^4     -5.063e-03  8.482e-04  -5.969 2.40e-09 ***
## color^5      5.713e-03  8.014e-04   7.130 1.02e-12 ***
## color^6      1.959e-03  7.285e-04   2.689  0.00716 **
## clarity.L   -1.784e-01  2.058e-03 -86.670  < 2e-16 ***
## clarity.Q    1.081e-01  1.785e-03  60.536  < 2e-16 ***
## clarity.C   -5.677e-02  1.518e-03 -37.391  < 2e-16 ***
## clarity^4    1.727e-02  1.211e-03  14.259  < 2e-16 ***
## clarity^5   -1.232e-02  9.885e-04 -12.464  < 2e-16 ***
## clarity^6   -1.793e-03  8.602e-04  -2.084  0.03717 *
## clarity^7    1.719e-04  7.595e-04   0.226  0.82093
## depth        1.212e-02  2.801e-04  43.278  < 2e-16 ***
## table        2.203e-03  1.825e-04  12.073  < 2e-16 ***
## price        4.428e-05  1.913e-07 231.494  < 2e-16 ***
## x            2.425e-01  1.800e-03 134.732  < 2e-16 ***
## y            5.961e-03  1.212e-03   4.917 8.82e-07 ***
## z            4.586e-03  2.100e-03   2.184  0.02899 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07088 on 53916 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9776
## F-statistic: 1.025e+05 on 23 and 53916 DF,  p-value: < 2.2e-16
```

The summary function is used to show the coefficients and intercepts, as well as other information such as the r-squared and adjusted r-squared.

This model includes all the variables in the dataset. To get a much better model it is better to create a model based on a relationship between two variables. Let's create a new model using the carat and price variables and the lm function again.

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

```
##
## Call:
## lm(formula = carat ~ price, data = diamonds)
##
## Coefficients:
## (Intercept)        price
##   0.3672972    0.0001095
```

```
summary(model2)
```

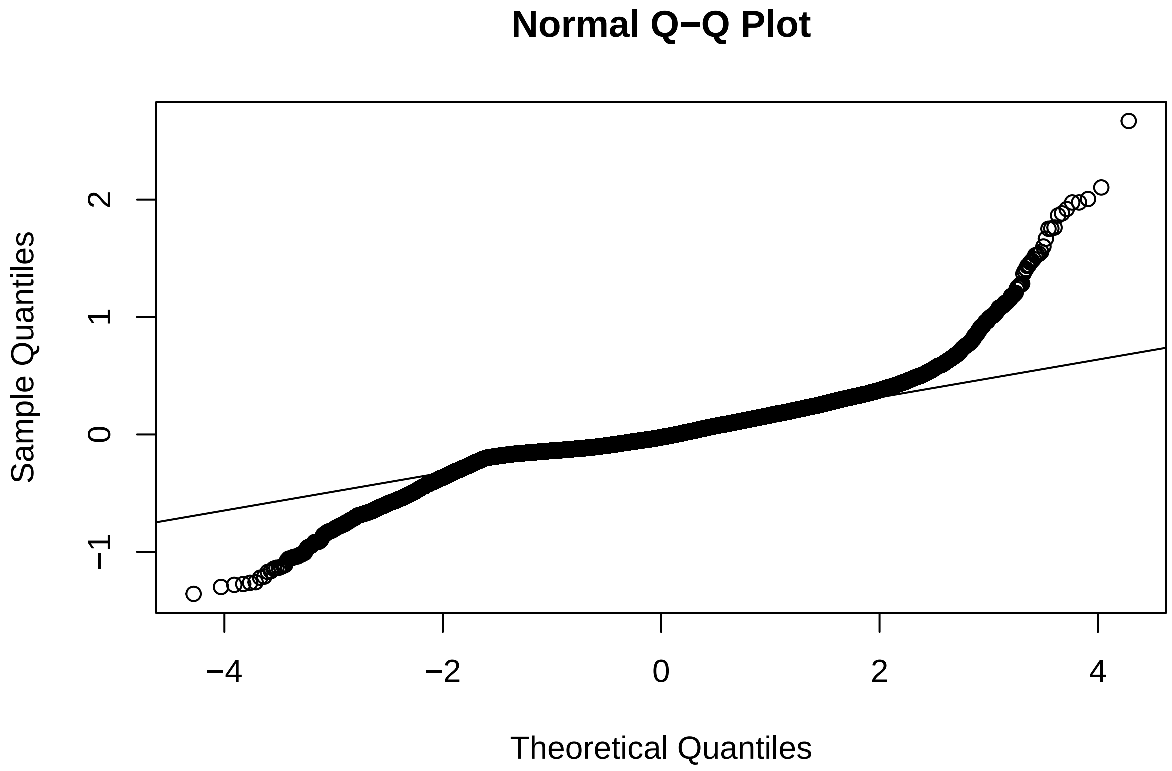
```
##
## Call:
## lm(formula = carat ~ price, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35765 -0.11329 -0.02442  0.10344  2.66973
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.673e-01  1.112e-03   330.2   <2e-16 ***
## price       1.095e-04  1.986e-07   551.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.184 on 53938 degrees of freedom
## Multiple R-squared:  0.8493, Adjusted R-squared:  0.8493
## F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

Here we can see the coefficients, intercepts, and other information from the summary function. This all comes from the model created from the carat and price variables. The equation of the model is y = 0.0001095x + 0.3672972.

Plot

Another good thing to do is plot the residuals to check for normality. Residuals are the distance between data points and the regression line. The qqnorm function is used to plot the residuals and the qqline function adds a line to the plot that passes through the first and third quartiles.

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



Here we see most of the residuals follow the straight line. There are a good amount of points that deviate from the line. We can say that this distribution follows a nearly normal distribution. We can say that the model fits the data. This means that there is a strong relationship between carat and price, meaning that a higher carat diamond would have a higher price to it.