HW5 - Logistic Regression and Multiple Linear Regression

Shiloh Bradley

6/10/2020

Toyota Corolla data

```
C <- read.csv("ToyotaCorolla2.csv", header=TRUE)
dim(C)  # 1436   11
```

```
## [1] 1436   11
```

```
head(C)
```

```
##   Price Age    KM Fuel_Type HP Met_Color Automatic   cc Doors
## 1 13500  23 46986    Diesel 90         1         0 2000     3
## 2 13750  23 72937    Diesel 90         1         0 2000     3
## 3 13950  24 41711    Diesel 90         1         0 2000     3
## 4 14950  26 48000    Diesel 90         0         0 2000     3
## 5 13750  30 38500    Diesel 90         0         0 2000     3
## 6 12950  32 61000    Diesel 90         0         0 2000     3
##   Quarterly_Tax Weight
## 1           210   1165
## 2           210   1165
## 3           210   1165
## 4           210   1165
## 5           210   1170
## 6           210   1170
```

```
C$Doors <- factor(C$Doors)
```

```
summary(C)
```

```
##      Price            Age              KM          Fuel_Type
##  Min.   : 4350   Min.   : 1.00   Min.   :     1   CNG   :  17
##  1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000   Diesel: 155
##  Median : 9900   Median :61.00   Median : 63390   Petrol:1264
##  Mean   :10731   Mean   :55.95   Mean   : 68533
##  3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021
##  Max.   :32500   Max.   :80.00   Max.   :243000
##        HP          Met_Color        Automatic             cc
##  Min.   : 69.0   Min.   :0.0000   Min.   :0.00000   Min.   : 1300
##  1st Qu.: 90.0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.: 1400
##  Median :110.0   Median :1.0000   Median :0.00000   Median : 1600
##  Mean   :101.5   Mean   :0.6748   Mean   :0.05571   Mean   : 1577
##  3rd Qu.:110.0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.: 1600
##  Max.   :192.0   Max.   :1.0000   Max.   :1.00000   Max.   :16000
##  Doors   Quarterly_Tax        Weight
##  2:  2   Min.   : 19.00   Min.   :1000
##  3:622   1st Qu.: 69.00   1st Qu.:1040
##  4:138   Median : 85.00   Median :1070
```

```
##  5:674   Mean   : 87.12   Mean   :1072
##          3rd Qu.: 85.00   3rd Qu.:1085
##          Max.   :283.00   Max.   :1615
```

Fit MLR Model

```
lm1 <- lm(Price~Age+KM+Fuel_Type+HP+Met_Color+Automatic+cc+Doors+Quarterly_Tax+Weight,data=C)
smre1 <- summary(lm1)
smre1
```

```
##
## Call:
## lm(formula = Price ~ Age + KM + Fuel_Type + HP + Met_Color +
##     Automatic + cc + Doors + Quarterly_Tax + Weight, data = C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11811.4   -756.6    -29.1    744.3   6771.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -7.564e+03  1.590e+03  -4.757 2.16e-06 ***
## Age             -1.229e+02  2.595e+00 -47.372  < 2e-16 ***
## KM              -1.686e-02  1.307e-03 -12.895  < 2e-16 ***
## Fuel_TypeDiesel  5.619e+02  3.763e+02   1.493    0.136
## Fuel_TypePetrol  2.486e+03  3.688e+02   6.741 2.27e-11 ***
## HP               2.253e+01  3.511e+00   6.417 1.89e-10 ***
## Met_Color        3.026e+01  7.491e+01   0.404    0.686
## Automatic        2.140e+02  1.588e+02   1.348    0.178
## cc              -6.421e-02  9.055e-02  -0.709    0.478
## Doors3          -6.961e+02  9.338e+02  -0.745    0.456
## Doors4          -4.848e+02  9.392e+02  -0.516    0.606
## Doors5          -8.517e+02  9.348e+02  -0.911    0.362
## Quarterly_Tax    1.256e+01  1.651e+00   7.609 4.99e-14 ***
## Weight           2.006e+01  1.254e+00  15.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1315 on 1422 degrees of freedom
## Multiple R-squared:  0.8697, Adjusted R-squared:  0.8685
## F-statistic: 729.8 on 13 and 1422 DF,  p-value: < 2.2e-16
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.6.2
```

```
vif(lm1) # all VIFs < 2
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## Age           1.931252  1        1.389695
## KM            1.993584  1        1.411943
## Fuel_Type     7.559246  2        1.658134
## HP            2.294335  1        1.514706
```

```
## Met_Color      1.021920  1        1.010900
## Automatic      1.100483  1        1.049039
## cc             1.224508  1        1.106575
## Doors          1.288879  3        1.043203
## Quarterly_Tax 3.822533  1        1.955130
## Weight         3.612684  1        1.900706
```

Drop insignificant predictors, P-value > 0.05

```
lm2 <- lm(Price~Age+KM+Fuel_Type+HP+Quarterly_Tax+Weight,data=C) ## Our final model -- explanatory
smre2 <- summary(lm2)
smre2
```

```
##
## Call:
## lm(formula = Price ~ Age + KM + Fuel_Type + HP + Quarterly_Tax +
##     Weight, data = C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11308.3   -781.1    -29.6    757.8   6827.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -7.202e+03  1.186e+03  -6.072 1.61e-09 ***
## Age             -1.226e+02  2.572e+00 -47.677  < 2e-16 ***
## KM              -1.726e-02  1.302e-03 -13.250  < 2e-16 ***
## Fuel_TypeDiesel  5.905e+02  3.682e+02   1.604    0.109
## Fuel_TypePetrol  2.374e+03  3.679e+02   6.452 1.51e-10 ***
## HP               2.329e+01  3.341e+00   6.973 4.74e-12 ***
## Quarterly_Tax    1.212e+01  1.649e+00   7.347 3.40e-13 ***
## Weight           1.903e+01  1.128e+00  16.873  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1319 on 1428 degrees of freedom
## Multiple R-squared:  0.8685, Adjusted R-squared:  0.8678
## F-statistic:  1347 on 7 and 1428 DF,  p-value: < 2.2e-16
```

```
vif(lm2) # all VIFs < 2
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## Age           1.889123  1        1.374454
## KM            1.969464  1        1.403376
## Fuel_Type     6.494914  2        1.596406
## HP            2.067076  1        1.437733
## Quarterly_Tax 3.796844  1        1.948549
## Weight        2.907560  1        1.705157
```

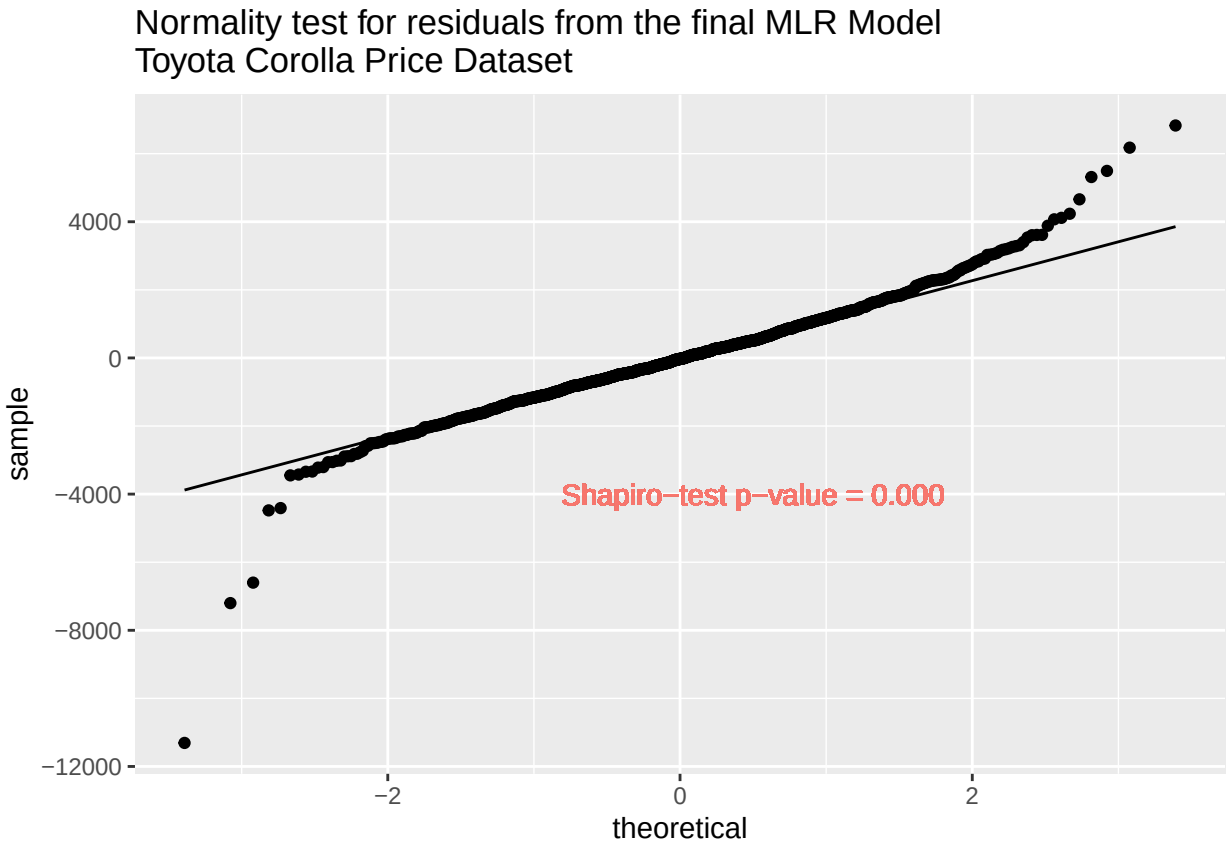
```
write.csv(smre2$coefficients,"MLR Final Model Toyota Corolla Price Dataset 052920.csv")
```

Verify normality of residuals from final MLR model lm2

```
shapiro.test(lm2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm2$residuals
## W = 0.95945, p-value < 2.2e-16
```

```
df.resid <- as.data.frame(lm2$residuals)
colnames(df.resid) <- "Residuals"
#qq plot with normal line (normality test for residuals from lm2)
ggplot(df.resid)+stat_qq(aes(sample=Residuals)) +
  geom_qq_line(aes(sample=Residuals))+
  geom_text(aes(x=0.5, y=-4000, color="red", label="Shapiro-test p-value = 0.000"))+
  theme(legend.position="none")+ggtitle("Normality test for residuals from the final MLR Model\nToyota C
```



Root Mean Square Error (RMSE)

```
MSE <- sum(lm2$residuals ** 2) / nrow(C)
MSE
```

```
## [1] 1729103
```

```
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 1314.954
```

Rhop50

```
C <- read.csv("rhop50.csv", header=TRUE)
dim(C)
```

```
## [1] 150   6
```

```
head(C)
```

```
##   X        Y     X1     X2     X3     X4
## 1 1 11952.35  99.72  99.05  99.25  99.30
## 2 2 12022.39  99.44  99.89  99.84 100.44
## 3 3 11988.09 100.43  98.62  99.87  99.83
## 4 4 11991.31 100.10  99.61  98.75 101.28
## 5 5 12240.01 101.43 101.20 101.38 102.34
## 6 6 12019.97 100.26  99.18 100.04 100.22
```

```
summary(C)
```

```
##        X                Y               X1              X2
##  Min.   :  1.00   Min.   :11660   Min.   : 97.32   Min.   : 97.36
##  1st Qu.: 38.25   1st Qu.:11923   1st Qu.: 99.08   1st Qu.: 98.89
##  Median : 75.50   Median :12004   Median : 99.84   Median : 99.79
##  Mean   : 75.50   Mean   :12007   Mean   : 99.88   Mean   : 99.78
##  3rd Qu.:112.75   3rd Qu.:12067   3rd Qu.:100.58   3rd Qu.:100.50
##  Max.   :150.00   Max.   :12313   Max.   :102.75   Max.   :103.20
##        X3               X4
##  Min.   : 96.29   Min.   : 96.76
##  1st Qu.: 99.19   1st Qu.: 99.19
##  Median : 99.78   Median : 99.75
##  Mean   : 99.82   Mean   : 99.88
##  3rd Qu.:100.48   3rd Qu.:100.48
##  Max.   :102.67   Max.   :103.20
```

Fit MLR Model

```
lm1 <- lm(Y~X1+X2+X3+X4,data=C)
smre1 <- summary(lm1)
smre1
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.4810  -5.7592   0.1916   6.4896  26.4827
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  100.763     85.290   1.181    0.239
## X1            19.119      1.112  17.187   <2e-16 ***
## X2            20.974      1.063  19.722   <2e-16 ***
## X3            39.080      1.119  34.923   <2e-16 ***
## X4            40.075      1.034  38.765   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.17 on 145 degrees of freedom
## Multiple R-squared:  0.9928, Adjusted R-squared:  0.9926
## F-statistic:  4975 on 4 and 145 DF,  p-value: < 2.2e-16
```

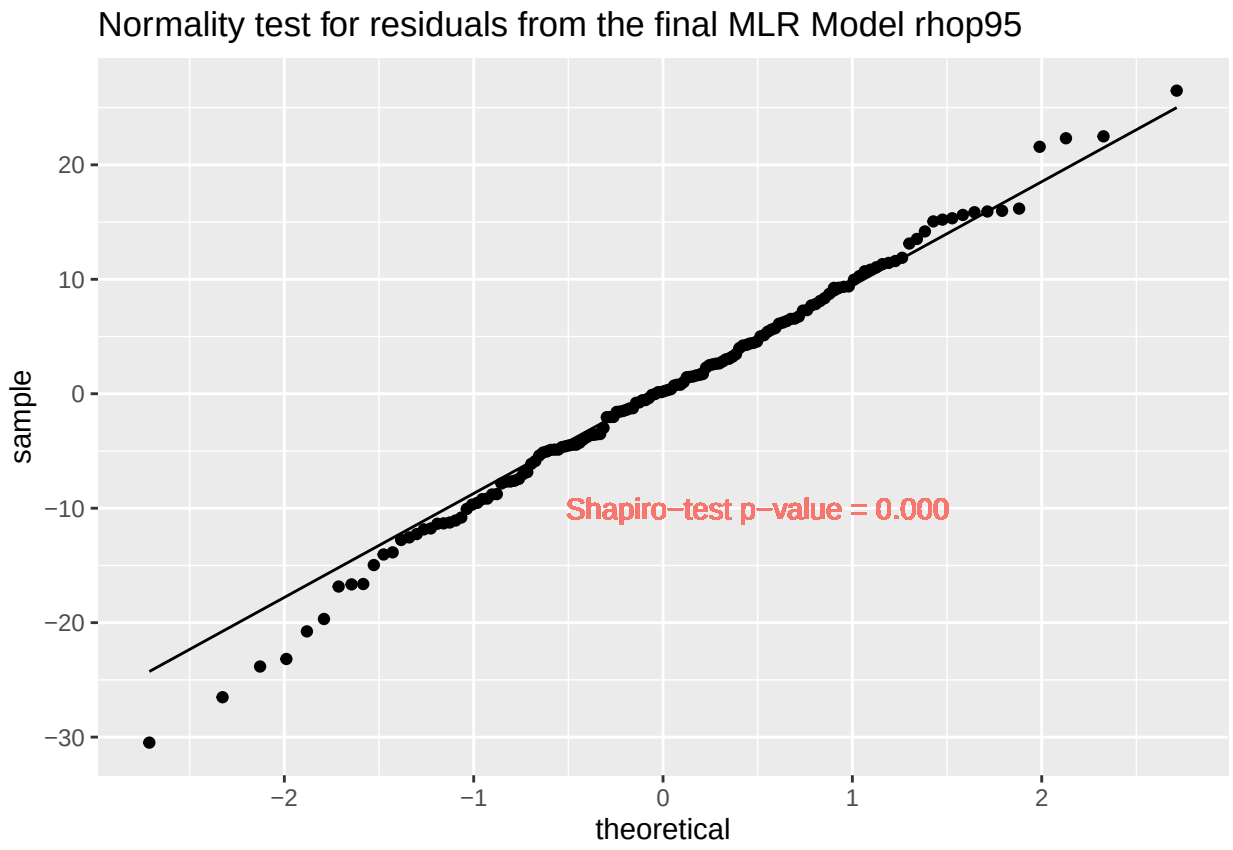
```
library(car)
vif(lm1) # all VIFs < 2
```

```
##       X1       X2       X3       X4
## 2.502759 2.177050 2.103689 2.182237
```

```
shapiro.test(lm1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm1$residuals
## W = 0.99414, p-value = 0.8074
```

```
df.resid <- as.data.frame(lm1$residuals)
colnames(df.resid) <- "Residuals"
#qq plot with normal line (normality test for residuals from lm2)
ggplot(df.resid)+stat_qq(aes(sample=Residuals)) +
  geom_qq_line(aes(sample=Residuals))+
  geom_text(aes(x=0.5, y=-10, color="red", label="Shapiro-test p-value = 0.000"))+
  theme(legend.position="none")+
  ggtitle("Normality test for residuals from the final MLR Model rhop95")
```



Root Mean Square Error (RMSE)

```
MSE <- sum(lm1$residuals ** 2) / nrow(C)
MSE
```

```
## [1] 100.0518
```

```
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 10.00259
```

Rhop 95

```
C <- read.csv("rhop95.csv", header=TRUE)
dim(C)
```

```
## [1] 150   6
```

```
head(C)
```

```
##   X        Y     X1     X2     X3     X4
## 1 1 12257.19 101.60 101.93 102.20 101.86
## 2 2 12280.40 101.80 101.69 102.35 101.68
## 3 3 12034.48 100.10  99.71 100.07 100.22
## 4 4 12062.48 100.18 100.42 100.51 100.27
## 5 5 12031.48  99.79  99.87 100.36 100.02
## 6 6 12066.00 100.45 100.22 100.36 100.13
```

```
summary(C)
```

```
##        X                Y               X1              X2
##  Min.   :  1.00   Min.   :11771   Min.   : 97.83   Min.   : 98.04
##  1st Qu.: 38.25   1st Qu.:11949   1st Qu.: 99.36   1st Qu.: 99.31
##  Median : 75.50   Median :12033   Median :100.00   Median :100.05
##  Mean   : 75.50   Mean   :12033   Mean   :100.04   Mean   :100.04
##  3rd Qu.:112.75   3rd Qu.:12099   3rd Qu.:100.65   3rd Qu.:100.62
##  Max.   :150.00   Max.   :12358   Max.   :102.87   Max.   :102.86
##        X3               X4
##  Min.   : 97.81   Min.   : 97.65
##  1st Qu.: 99.38   1st Qu.: 99.33
##  Median :100.04   Median :100.11
##  Mean   :100.08   Mean   :100.08
##  3rd Qu.:100.69   3rd Qu.:100.68
##  Max.   :102.66   Max.   :102.90
```

Fit MLR model

```
lm1 <- lm(Y~X1+X2+X3+X4,data=C)
smre1 <- summary(lm1)
smre1
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = C)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.9183  -6.3479  -0.2488   6.4844  29.7771
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -82.512     88.185  -0.936    0.351
## X1            17.099      3.609   4.738 5.09e-06 ***
## X2            20.093      3.335   6.026 1.33e-08 ***
## X3            42.794      3.557  12.032  < 2e-16 ***
## X4            41.090      3.973  10.342  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.32 on 145 degrees of freedom
## Multiple R-squared:  0.9927, Adjusted R-squared:  0.9925
## F-statistic:  4948 on 4 and 145 DF,  p-value: < 2.2e-16
```

```
library(car)
vif(lm1) # all VIFs < 2
```

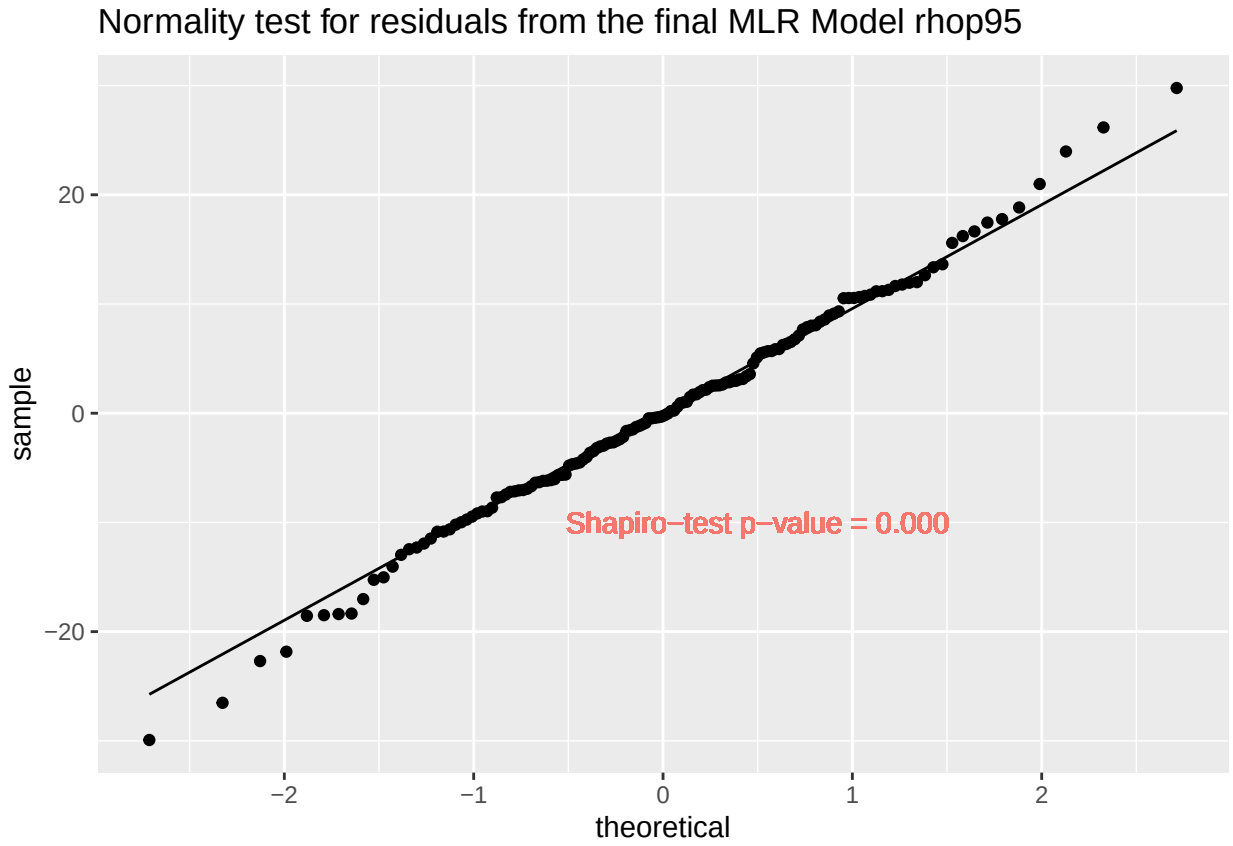
```
##       X1       X2       X3       X4
## 18.80025 16.30793 18.07852 20.43866
```

Verify normaly of residuals from final MLR model lm1

```
shapiro.test(lm1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm1$residuals
## W = 0.9952, p-value = 0.9064
```

```
df.resid <- as.data.frame(lm1$residuals)
colnames(df.resid) <- "Residuals"
#qq plot with normal line (normality test for residuals from lm2)
ggplot(df.resid)+stat_qq(aes(sample=Residuals)) +
  geom_qq_line(aes(sample=Residuals))+
  geom_text(aes(x=0.5, y=-10, color="red", label="Shapiro-test p-value = 0.000"))+
  theme(legend.position="none")+
  ggtitle("Normality test for residuals from the final MLR Model rhop95")
```



Root Mean Square Error (RMSE)

```
MSE <- sum(lm1$residuals ** 2) / nrow(C)
MSE
```

```
## [1] 102.9734
```

```
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 10.14758
```