

**Car Price Prediction Project**

***Team Members:***

***Suryateja Chalapati***

***Prof. Ron Satterfield***

***Mar 8th 2020***

**Managerial Decision Analysis**

WWW.DESCARGARPORTADAS.COM

**REGRESSION PROJECT:**

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**Dataset:**

**Source:** <https://www.kaggle.com/hellbuoy/car-price-prediction>

**Descriptions of Data:**

**Variables Used:**

Total Observations/Sample Size = 159.

Price of Cars ($), this is the Dependent/Response Variable.

Horsepower (hp), this is the First Independent/Explanatory Variable.

Peak RPM (RPM), this is the Second Independent/Explanatory Variable.

No. of Doors, this is the Third Independent/Explanatory Variable. This is the Binary variable with two levels.

Two Door = 1 (67 Observations, 42% of the total).

Four Door = 0 (92 Observations, 58% of the total).

(Binary Variable satisfies the 35/15 condition).

**Case:** We are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

**Dataset Attached Below: (Refer: *Sheet1*)**



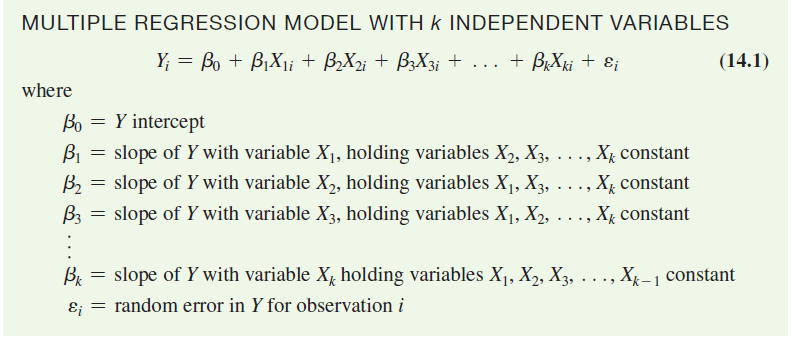
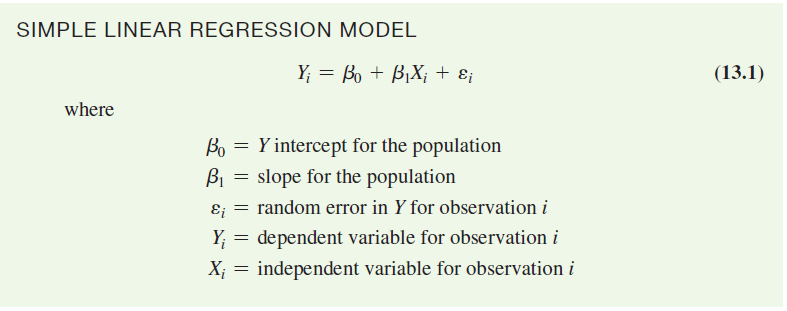
\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_



**Regression Analysis done on the following model combinations:**

1. (y, X1), (y, X­2), (y, X3)
2. (y, X1, X­2), (y, X1, X3), (y, X­2, X3)
3. (y, X1, X­2, X3)
4. (y, X1, X­2, X1X­2)
5. (y, X1, X1²) and (y, X­2, X­2²)

**Regression equations used:**

R, R-Studio and Minitab were used in conducting analysis for this project. We have used Three Independent Variables for this project with one being a Binary/Categorical Variable with two Factor-Levels. Following are the analysis results:

**Case\_1. (y, X1)**

**Analysis Results:**

a) Setting up the Structure of R Objects:

> price$doornumber\_x3 = as.factor(price$doornumber\_x3)

> str(price)

'data.frame': 159 obs. of 7 variables:

$ carname : chr "alfa-romero giulia" "alfa-romero stelvio" "audi 100 ls" "bmw 320i" ...

$ doors : chr "two" "two" "four" "two" ...

$ doornumber\_x3 : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...

$ cylindernumber: chr "four" "four" "four" "four" ...

$ horsepower\_x1 : num 111 111 102 101 101 70 70 68 68 102 ...

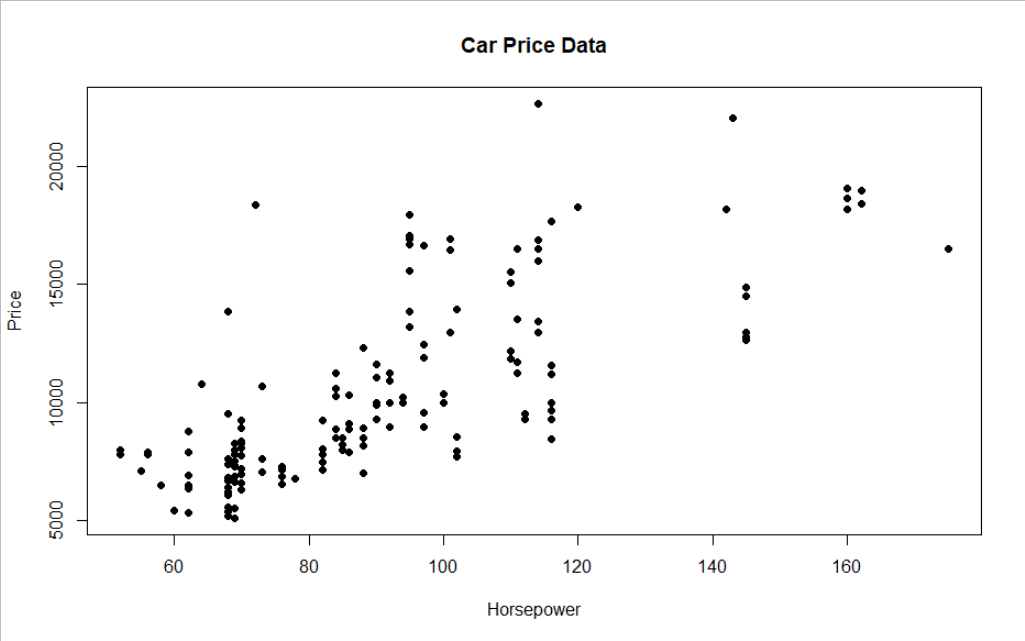
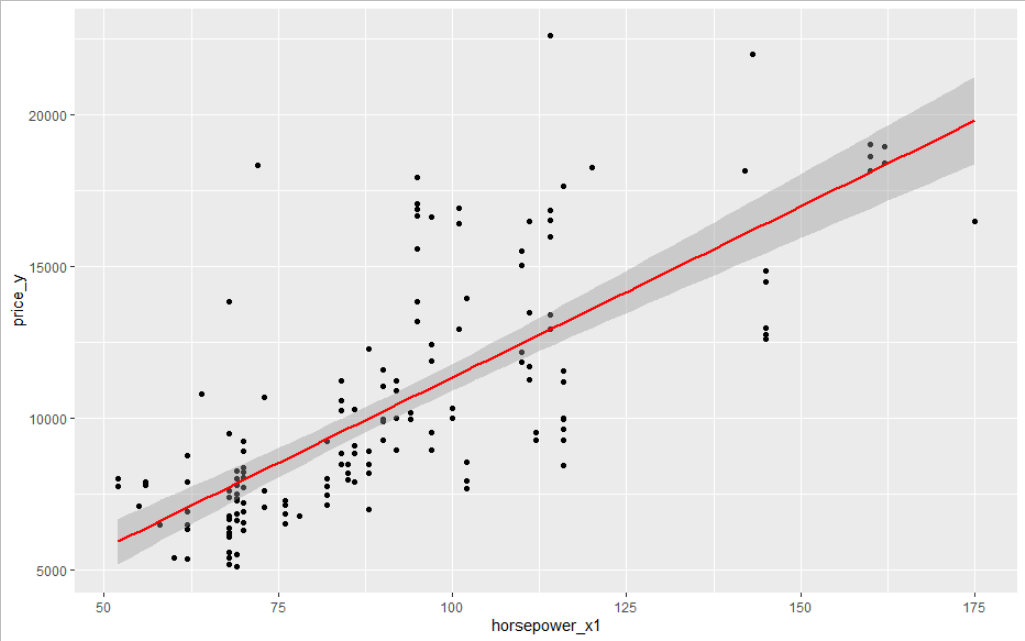
$ peakrpm\_x2 : num 5000 5000 5500 5800 5800 5400 5400 5500 5500 5500 ...

$ price\_y : num 13495 16500 13950 16430 16925 ...

b) Scatter Plot:

> # 1.1 (y,X1)

> # Basic scatterplot of the data.

c) Simple Regression:

> # 1.1 (y,X1)

> # Conducting a Simple regression (lm="Linear Model") on the data.

> price1.1=lm(price\_y~horsepower\_x1)

> summary(price1.1)

Call:

lm(formula = price\_y ~ horsepower\_x1)

Residuals:

Min 1Q Median 3Q Max

-4708.9 -1549.4 -630.8 949.3 10151.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 68.308 773.310 0.088 0.93

horsepower\_x1 112.841 8.219 13.730 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2652 on 157 degrees of freedom

Multiple R-squared: 0.5456, Adjusted R-squared: 0.5427

F-statistic: 188.5 on 1 and 157 DF, p-value: < 2.2e-16

> confint(price1.1)

2.5 % 97.5 %

(Intercept) -1459.12471 1595.7407

horsepower\_x1 96.60805 129.0742

Dependent variable (price\_y)

Y-Intercept for the sample

Slope for the sample

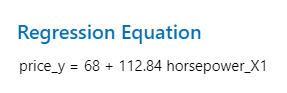
Independent variable (horsepower\_x1)

R-squared: 0.5456, Adjusted R-squared: 0.5427

d) Observation:

1. This looks like a linear distribution from the scatter plot above.

2. R-squared (Coefficient of determination) is 54.56%, which means about 54.56% of the variance in price is explained by the horsepower. Regression equation as follows,



3. Interpretation of the Regression Equation:

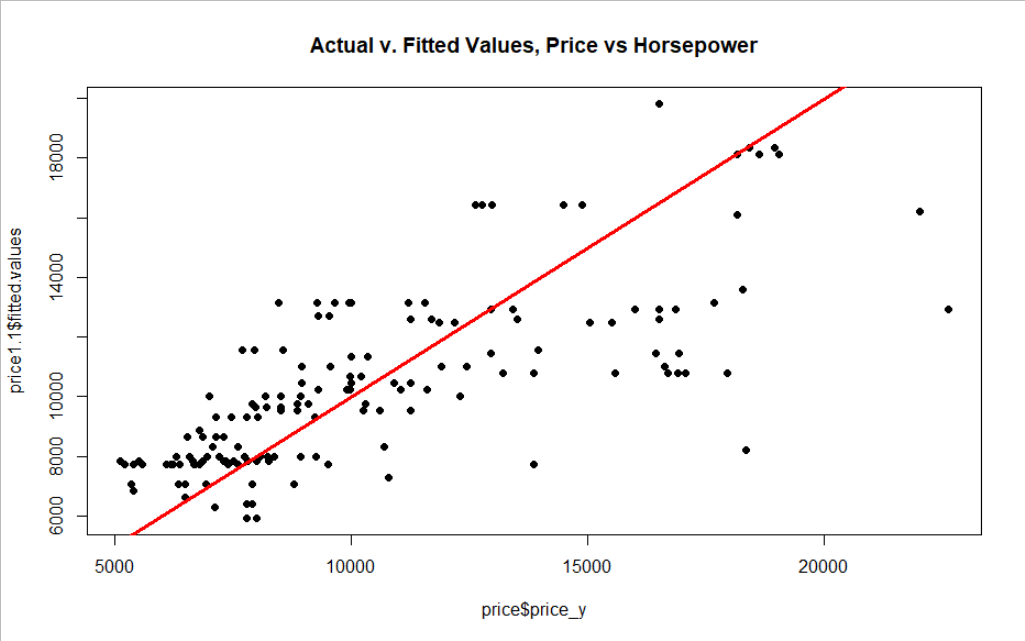
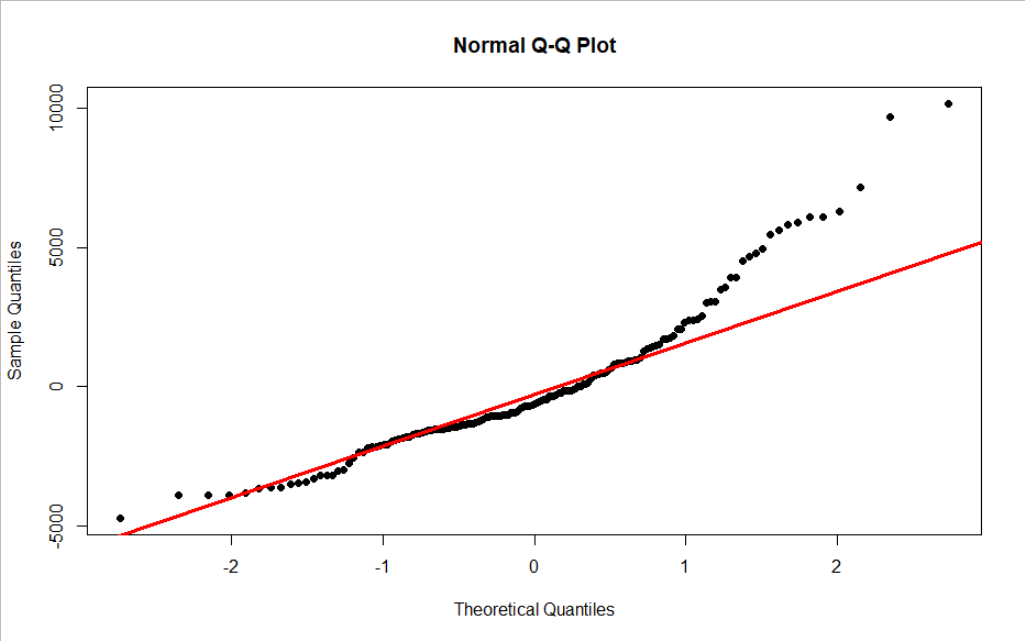
|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 68.3 | No change in horsepower. | The predicted mean price increases by $68.3. |
| *ß1* = 112.84 | Horsepower increases by 1hp. | The predicted mean price increases by $112.84. |

4. p-value for Y-Intercept *ß0* is 0.93 (Must be below 0.05), other p-value interpretations as follows:

|  |  |  |
| --- | --- | --- |
| p-value | *ß0* = 0.93 | *ß1* < 2e-16 |
| Independent Variables | Fail to Reject Null, Not Significant | Reject the Null, Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Highly Significant | |

e) L.I.N.E Assumptions:

Linearity & Normality:

Independence of errors:

> # Independence of the observations.

> dwtest(lm(price\_y~horsepower\_x1))

Durbin-Watson test

data: lm(price\_y ~ horsepower\_x1)

DW = 0.97404, p-value = 2.787e-11

alternative hypothesis: true autocorrelation is greater than 0

> chisq.test(horsepower\_x1,price\_y, correct = FALSE)

Pearson's Chi-squared test

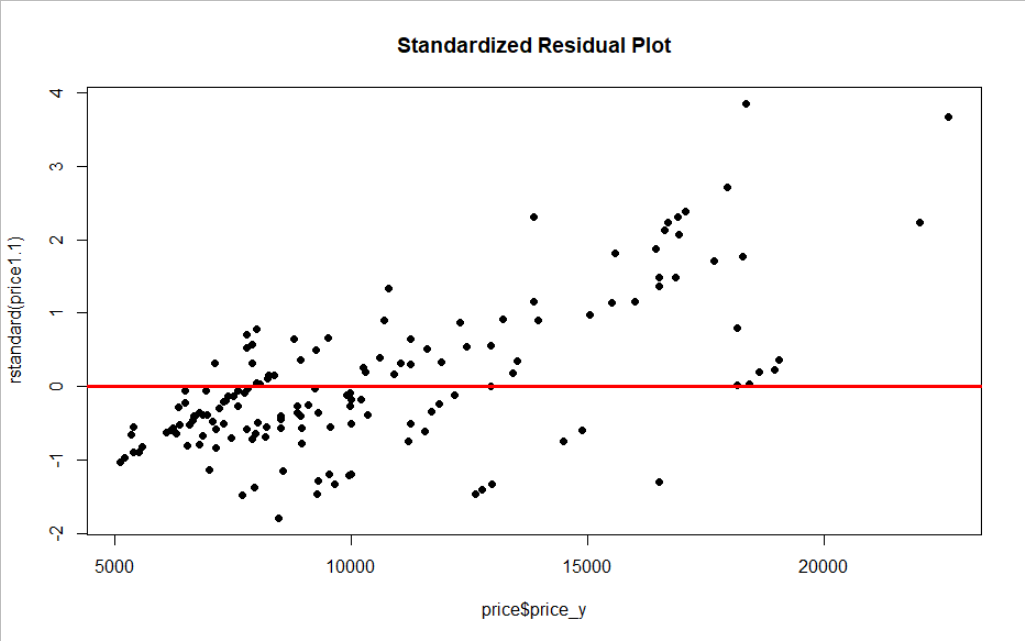
data: horsepower\_x1 and price\_y

X-squared = 5749.2, df = 5472, p-value = 0.004495

From Durbin-Watson test, DW=0.97, p-value = 2.787e-11, residuals are autocorrelated.

From Chi-squared test, p-value = 0.004495, means the variables are related.

Equality of Variances:



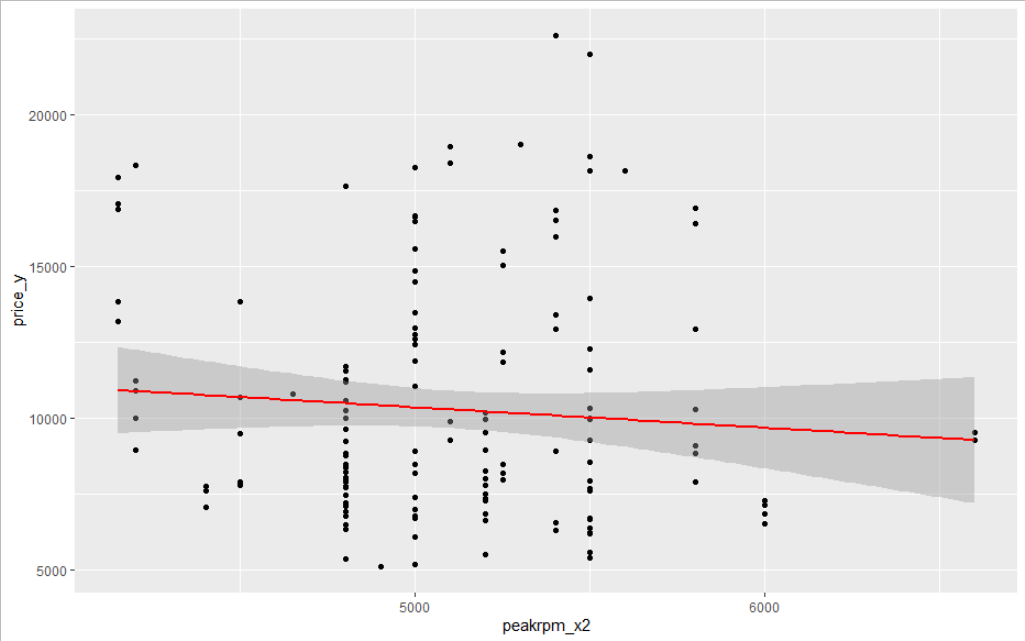
**Case\_2. (y, X2)**

**Analysis Results:**

a) Scatter Plot:

> # 1.2 (y,X2)

> # Basic scatterplot of the data.

b) Simple Regression:

> # 1.2 (y,X2)

> # Conducting a Simple regression (lm="Linear Model") on the data.

> price1.2=lm(price\_y~peakrpm\_x2)

> summary(price1.2)

Call:

lm(formula = price\_y ~ peakrpm\_x2)

Residuals:

Min 1Q Median 3Q Max

-5311 -2895 -1239 2171 12533

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13736.3099 3475.8807 3.952 0.000117 \*\*\*

peakrpm\_x2 -0.6749 0.6771 -0.997 0.320438

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3922 on 157 degrees of freedom

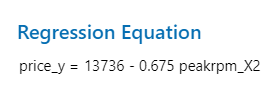
Multiple R-squared: 0.006288, Adjusted R-squared: -4.156e-05

F-statistic: 0.9934 on 1 and 157 DF, p-value: 0.3204

> confint(price1.2)

2.5 % 97.5 %

(Intercept) 6870.787931 2.060183e+04

peakrpm\_x2 -2.012273 6.625268e-01

Dependent variable (price\_y)

Y-Intercept for the sample

Slope for the sample

Independent variable (peakrpm\_x2)

R-squared: 0.006288, Adjusted R-squared: -4.156e-05

c) Observation:

1. This looks like a curved distribution from the scatter plot above. R-squared (Coefficient of determination) is 0.63%, which means about 0.63% of the variance in price is explained by the peak RPM. Regression equation,

2. Interpretation of the Regression Equation:

|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 13736.3 | No change in peak RPM. | The predicted mean price increases by $13736.3. |
| *ß1* = -0.675 | Peak RPM increases by 1RPM. | The predicted mean price decreases by $0.675. |

**Case\_3. (y, X3)**

**Analysis Results:**

a) Simple Regression:

> # 1.3 (y,X3)

> # Conducting a Simple regression (lm="Linear Model") on the data.

> price1.3=lm(price\_y~doornumber\_x3)

> summary(price1.3)

Call:

lm(formula = price\_y ~ doornumber\_x3)

Residuals:

Min 1Q Median 3Q Max

-4672 -2982 -1382 2076 12578

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10901.5 403.1 27.043 <2e-16 \*\*\*

doornumber\_x31 -1461.2 621.0 -2.353 0.0199 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3866 on 157 degrees of freedom

Multiple R-squared: 0.03406, Adjusted R-squared: 0.02791

F-statistic: 5.537 on 1 and 157 DF, p-value: 0.01986

> confint(price1.3)

2.5 % 97.5 %

(Intercept) 10105.27 11697.7017

doornumber\_x31 -2687.78 -234.6353

Dependent variable (price\_y)

Y-Intercept for the sample

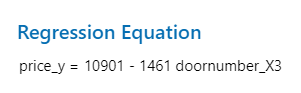
Slope for the sample

Independent variable (doornumber\_x3, Two Door = 1)

R-squared: 0.03406, Adjusted R-squared: 0.02791

b) Observation:

1. R-squared (Coefficient of determination) is 3.4%, which means about 3.4% of the variance in price is explained by the No. of doors. Regression equation as follows,



2. Interpretation of the Regression Equation:

|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 10901.5 | No change in No. of doors. | The predicted mean price increases by $10901.5. |
| *ß1* = -1461.2 | No. of doors increases by one. | The predicted mean price decreases by $1461.2. |

3. p-value interpretations as follows:

|  |  |  |
| --- | --- | --- |
| p-value | *ß0* < 2e-16 | *ß­1* = 0.0199 |
| Independent Variables | Reject the Null, Highly Significant | Reject the Null, Significant |
| Overall p-value = 0.01986 | Reject the Null, Significant | |

**Case\_4. (y, X1, X2)**

**Analysis Results:**

a) Multiple Regression:

> # 2.1 (y,X1,X2)

> # Multiple regression for Horsepower and Peak RPM.

> price2.1 = lm(price\_y~horsepower\_x1+peakrpm\_x2,data = price)

> summary(price2.1)

Call:

lm(formula = price\_y ~ horsepower\_x1 + peakrpm\_x2, data = price)

Residuals:

Min 1Q Median 3Q Max

-5303.3 -1471.9 -433.3 858.7 10053.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7785.9168 2305.0893 3.378 0.000923 \*\*\*

horsepower\_x1 116.7813 8.0103 14.579 < 2e-16 \*\*\*

peakrpm\_x2 -1.5792 0.4463 -3.539 0.000530 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2560 on 156 degrees of freedom

Multiple R-squared: 0.5794, Adjusted R-squared: 0.574

F-statistic: 107.4 on 2 and 156 DF, p-value: < 2.2e-16

> confint(price2.1)

2.5 % 97.5 %

(Intercept) 3232.702662 12339.130975

horsepower\_x1 100.958615 132.604000

peakrpm\_x2 -2.460725 -0.697713

Dependent variable (price\_y)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

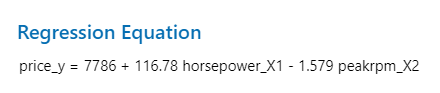
First Independent variable (horsepower\_x1)

Second Independent variable (peakrpm\_x2)

R-squared: 0.5794, Adjusted R-squared: 0.574

b) Observation:

1. R-squared (Coefficient of determination) is 57.94%, which means about 57.94% of the variance in price is explained by both horsepower and peak RPM. Regression equation as follows,



2. Interpretation of the Regression Equation:

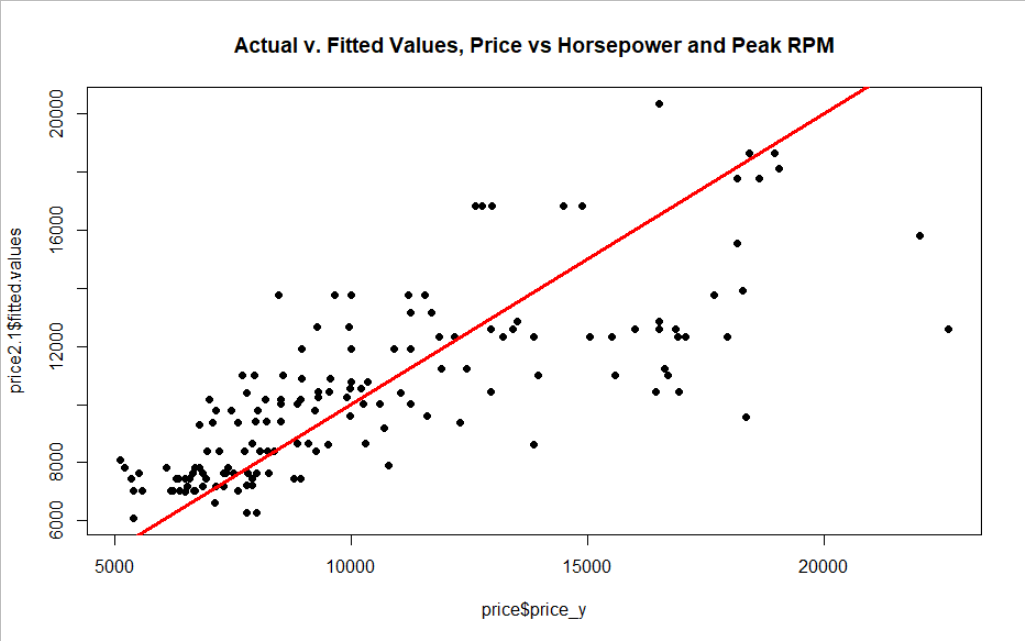
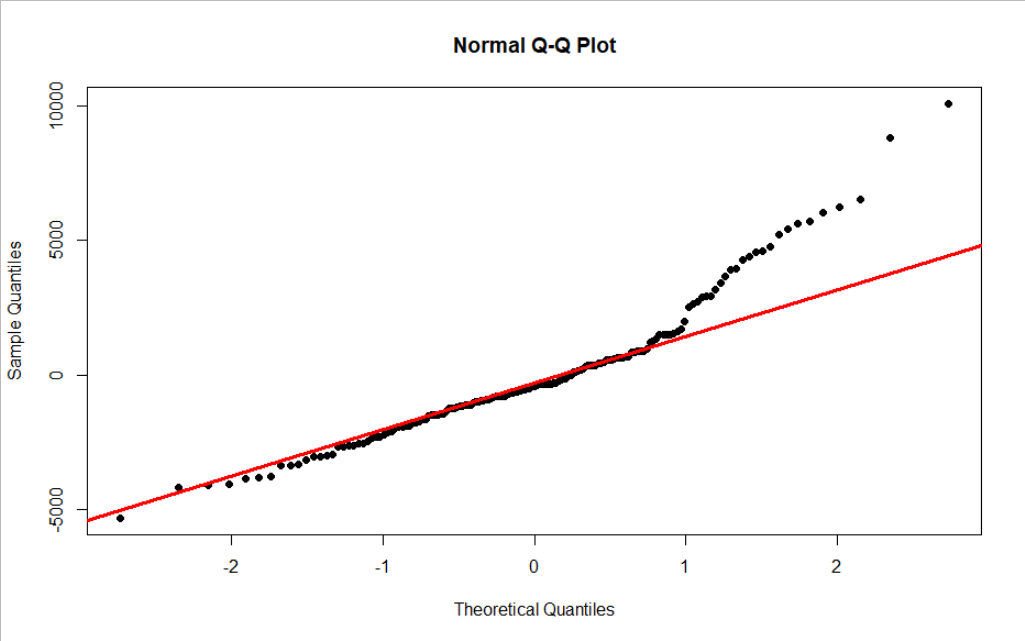
|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 7785.9 | No change in the other two variables horsepower and peak RPM. | The predicted mean price increases by $7785.9. |
| *ß1* = 116.78 | Horsepower increases by 1hp. | The predicted mean price increases by $116.78, holding peak RPM constant. |
| *ß2* = -1.579 | Peak RPM increases by 1RPM. | The predicted mean price decreases by $1.579, holding horsepower constant. |

3. p-value interpretations as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| p-value | *ß0* = 0.000923 | *ß1* < 2e-16 | *ß2* = 0.000530 |
| Independent Variables | Reject the Null, Significant | Reject the Null, Highly Significant | Reject the Null, Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Highly Significant | | |

c) L.I.N.E Assumptions:

Linearity & Normality:

Independence of errors:

> # Independence of the observations.

> dwtest(price2.1)

Durbin-Watson test

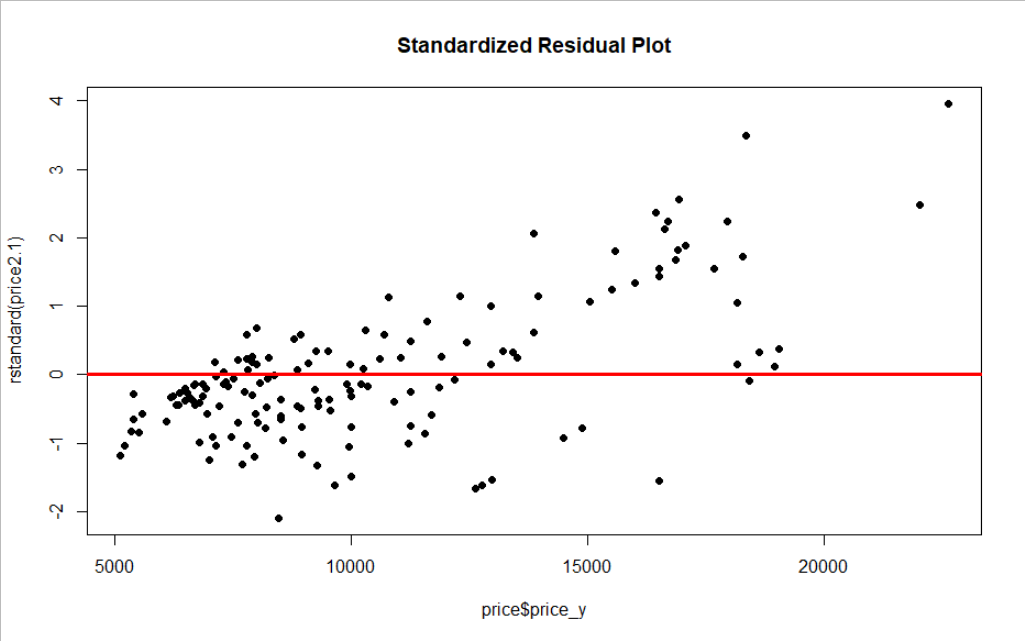
data: price2.1

DW = 1.0392, p-value = 2.928e-10

alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test, DW=1.0392, p-value = 2.928e-10, residuals are autocorrelated.

Equality of Variances:



**Case\_5. (y, X1, X3)**

**Analysis Results:**

a) Multiple Regression:

> # 2.2 (y,X1,X3)

> # Multiple regression for Horsepower and No. of Doors.

> price2.2 = lm(price\_y~horsepower\_x1+doornumber\_x3,data = price)

> summary(price2.2)

Call:

lm(formula = price\_y ~ horsepower\_x1 + doornumber\_x3, data = price)

Residuals:

Min 1Q Median 3Q Max

-4708 -1771 -536 1017 9420

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 640.96 739.52 0.867 0.387

horsepower\_x1 115.05 7.76 14.825 < 2e-16 \*\*\*

doornumber\_x31 -1833.17 402.17 -4.558 1.04e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2499 on 156 degrees of freedom

Multiple R-squared: 0.599, Adjusted R-squared: 0.5939

F-statistic: 116.5 on 2 and 156 DF, p-value: < 2.2e-16

> confint(price2.2)

2.5 % 97.5 %

(Intercept) -819.80306 2101.7204

horsepower\_x1 99.71919 130.3767

doornumber\_x31 -2627.57285 -1038.7717

Dependent variable (price\_y)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

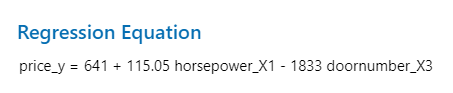
First Independent variable (horsepower\_x1)

Third Independent variable (doornumber\_x3, Two Door = 1)

R-squared: 0.599, Adjusted R-squared: 0.5939

b) Observation:

1. R-squared (Coefficient of determination) is 59.9%, which means about 59.9% of the variance in price is explained by both horsepower and No. of doors. Regression equation as follows,



2. Interpretation of the Regression Equation:

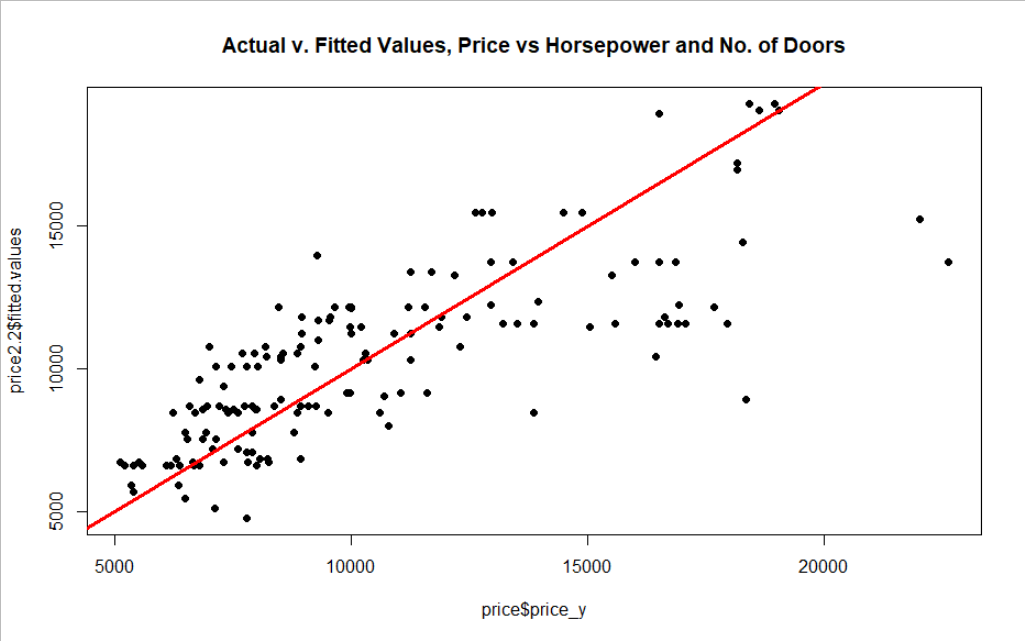
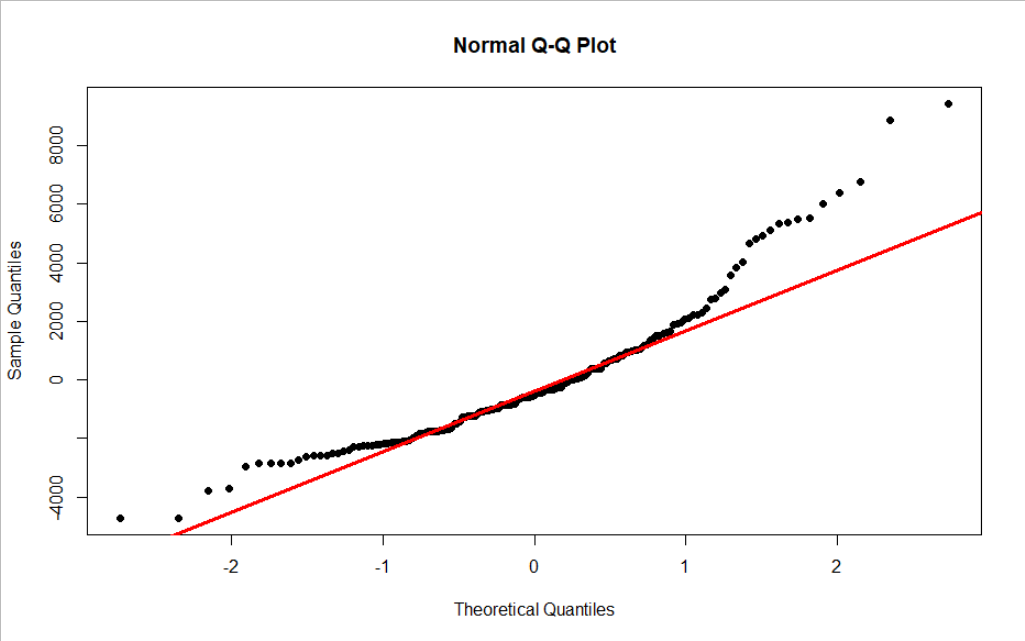
|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 640.96 | No change in the other two variables horsepower and No. of doors. | The predicted mean price increases by $640.96. |
| *ß1* = 115.05 | Horsepower increases by 1hp. | The predicted mean price increases by $115.05, holding No. of doors constant. |
| *ß2* = -1833.17 | No. of doors increases by one. | The predicted mean price decreases by $1833.17, holding horsepower constant. |

3. p-value interpretations as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| p-value | *ß0* = 0.387 | *ß1* < 2e-16 | *ß2* = 1.04e-05 |
| Independent Variables | Fail to Reject the Null, Not Significant | Reject the Null, Highly Significant | Reject the Null, Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Highly Significant | | |

c) L.I.N.E Assumptions:

Linearity & Normality:

Independence of errors:

> # Independence of the observations.

> dwtest(price2.2)

Durbin-Watson test

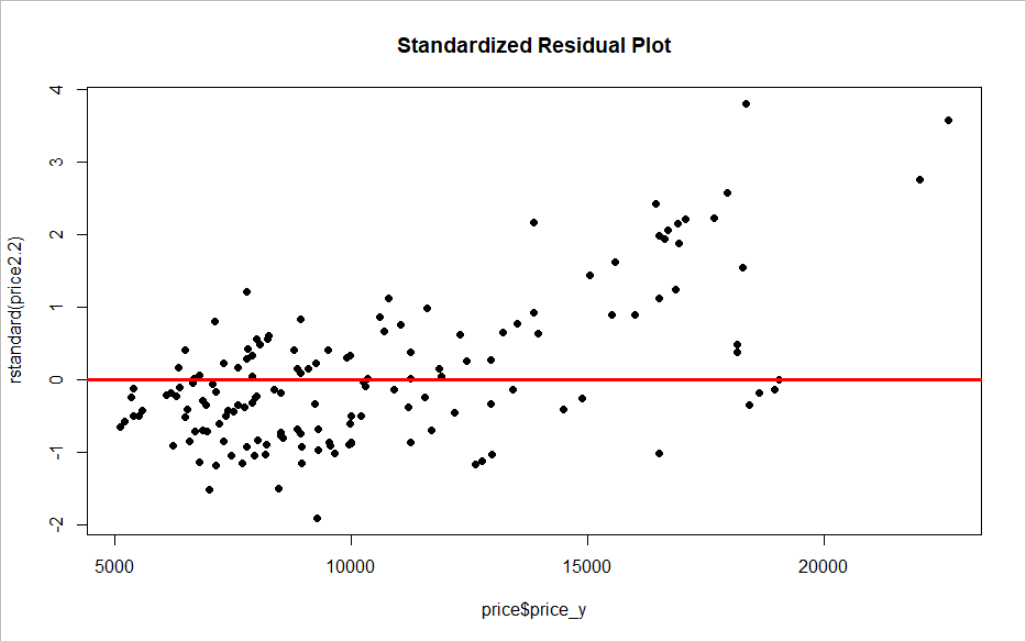
data: price2.2

DW = 1.2012, p-value = 1.341e-07

alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test, DW=1.2012, p-value = 1.341e-07, residuals are autocorrelated.

Equality of Variances:



**Case\_6. (y, X2, X3)**

**Analysis Results:**

a) Multiple Regression:

> # 2.3 (y,X2,X3)

> # Multiple regression for Peak RPM and No. of Doors.

> price2.3 = lm(price\_y~peakrpm\_x2+doornumber\_x3,data = price)

> summary(price2.3)

Call:

lm(formula = price\_y ~ peakrpm\_x2 + doornumber\_x3, data = price)

Residuals:

Min 1Q Median 3Q Max

-4508 -3034 -1352 2081 12672

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12650.9818 3470.9899 3.645 0.000364 \*\*\*

peakrpm\_x2 -0.3480 0.6858 -0.507 0.612531

doornumber\_x31 -1390.4147 637.9030 -2.180 0.030782 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3876 on 156 degrees of freedom

Multiple R-squared: 0.03566, Adjusted R-squared: 0.02329

F-statistic: 2.884 on 2 and 156 DF, p-value: 0.05889

> confint(price2.3)

2.5 % 97.5 %

(Intercept) 5794.778872 19507.184819

peakrpm\_x2 -1.702564 1.006547

doornumber\_x31 -2650.456457 -130.372847

Dependent variable (price)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

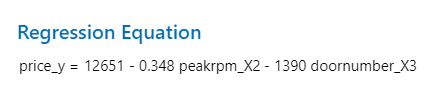
Second Independent variable (peakrpm\_x2)

Third Independent variable (doornumber\_x3, Two Door = 1)

R-squared: 0.03566, Adjusted R-squared: 0.02329

b) Observation:

1. R-squared (Coefficient of determination) is 3.56%, which means about 3.56% of the variance in price is explained by both peak RPM and No. of doors. Regression equation as follows,



2. Interpretation of the Regression Equation:

|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 12650.98 | No change in the other two variables peak RPM and No. of doors. | The predicted mean price increases by $12650.98. |
| *ß1* = -0.348 | Peak RPM increases by 1RPM. | The predicted mean price decreases by $0.348, holding No. of doors constant. |
| *ß2* = -1390.41 | No. of doors increases by one. | The predicted mean price decreases by $1390.41, holding peak RPM constant. |

**Case\_7. (y, X1, X2, X3)**

**Analysis Results:**

a) Full Multiple Regression:

> # Q3 (y,X1,X2,X3)

> # Multiple regression for Horsepower, Peak RPM and No .of Doors.

> price3 = lm(price\_y~horsepower\_x1+peakrpm\_x2+doornumber\_x3,data = price)

> summary(price3)

Call:

lm(formula = price\_y ~ horsepower\_x1 + peakrpm\_x2 + doornumber\_x3,

data = price)

Residuals:

Min 1Q Median 3Q Max

-4290.8 -1605.8 -432.5 990.2 9252.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6489.3409 2227.9719 2.913 0.004113 \*\*

horsepower\_x1 117.7856 7.6624 15.372 < 2e-16 \*\*\*

peakrpm\_x2 -1.2119 0.4366 -2.776 0.006187 \*\*

doornumber\_x31 -1595.4907 402.9991 -3.959 0.000114 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2447 on 155 degrees of freedom

Multiple R-squared: 0.618, Adjusted R-squared: 0.6106

F-statistic: 83.59 on 3 and 155 DF, p-value: < 2.2e-16

> confint(price3)

2.5 % 97.5 %

(Intercept) 2088.233982 10890.4478845

horsepower\_x1 102.649285 132.9218330

peakrpm\_x2 -2.074419 -0.3494229

doornumber\_x31 -2391.569911 -799.4113977

Dependent variable (price\_y)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

Third Slope for the sample

First Independent variable (horsepower\_x1)

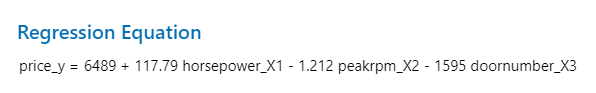
Second Independent variable (peakrpm\_x2)

Third Independent variable (doornumber\_x3, Two Door = 1)

R-squared: 0.618, Adjusted R-squared: 0.6106

b) Observation:

1. R-squared (Coefficient of determination) is 61.8%, which means about 61.8% of the variance in price is explained by the independent variable’s horsepower, peak RPM and No. of doors. Regression equation as follows,



2. Interpretation of the Regression Equation:

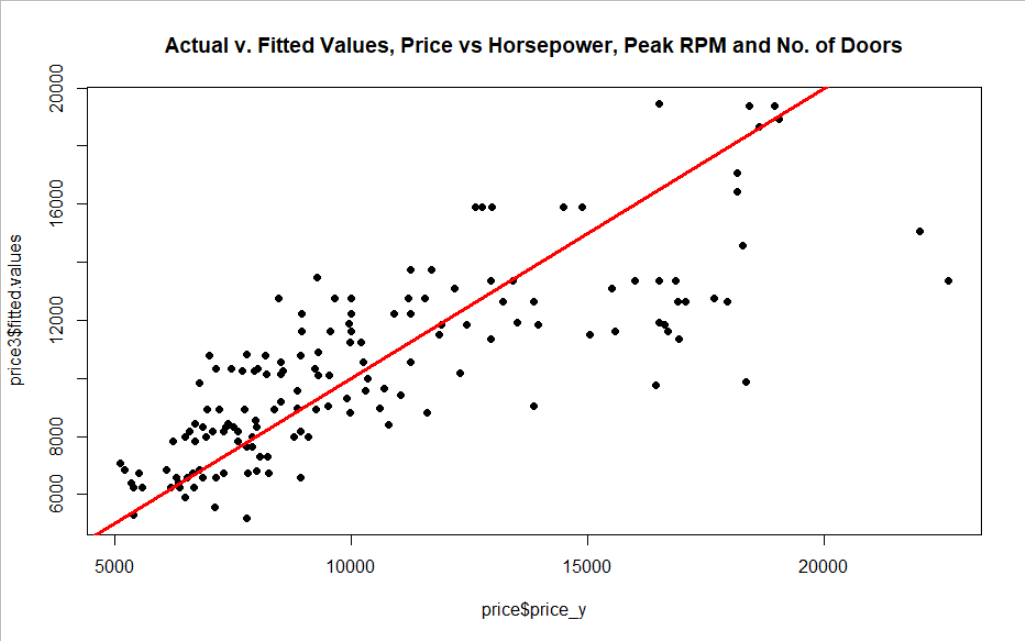
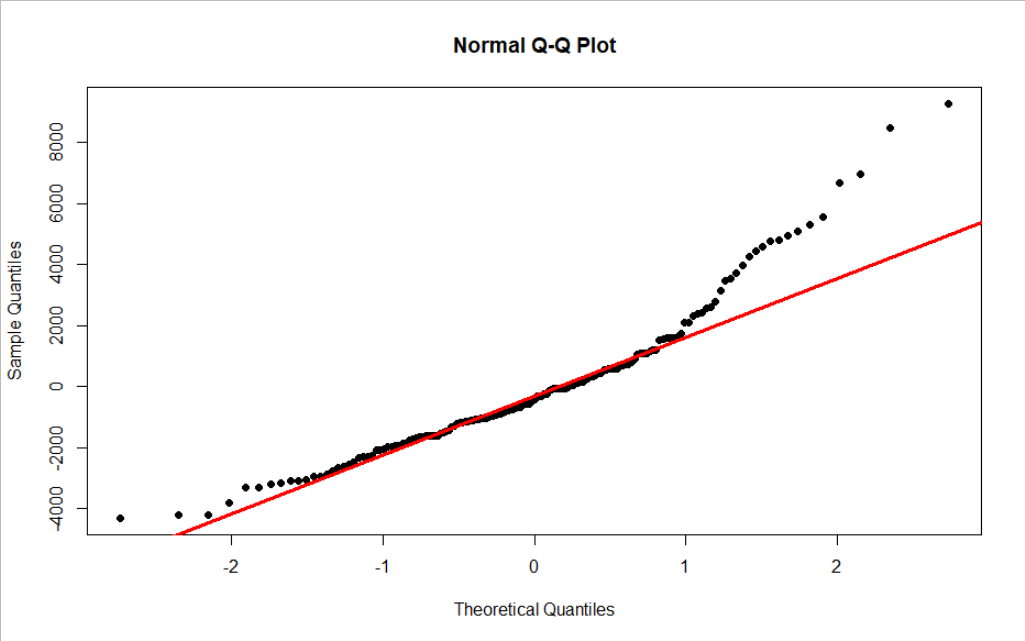
|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 6489.34 | No change in the other three variables horsepower, peak RPM and No. of doors. | The predicted mean price increases by $6489.34 |
| *ß1* = 117.78 | Horsepower increases by 1hp. | The predicted mean price increases by $117.78 holding peak RPM and No. of doors constant |
| *ß2* = -1.212 | Peak RPM increases by 1RPM. | The predicted mean price decreases by $1.212 holding horsepower and No. of doors constant |
| *ß3* = -1595.49 | No. of doors increases by one. | The predicted mean price decreases by $1595.49 holding horsepower and peak RPM constant |

3. p-value interpretations as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| p-value | *ß0* = 0.004113 | *ß1* < 2e-16 | *ß2* = 0.006187 | *ß3* = 0.000114 |
| Independent Variables | Reject the Null, Significant | Reject the Null, Highly Significant | Reject the Null, Significant | Reject the Null, Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Highly Significant | | | |

c) L.I.N.E Assumptions:

Linearity & Normality:

Independence of errors:

> # Independence of the observations.

> dwtest(price3)

Durbin-Watson test

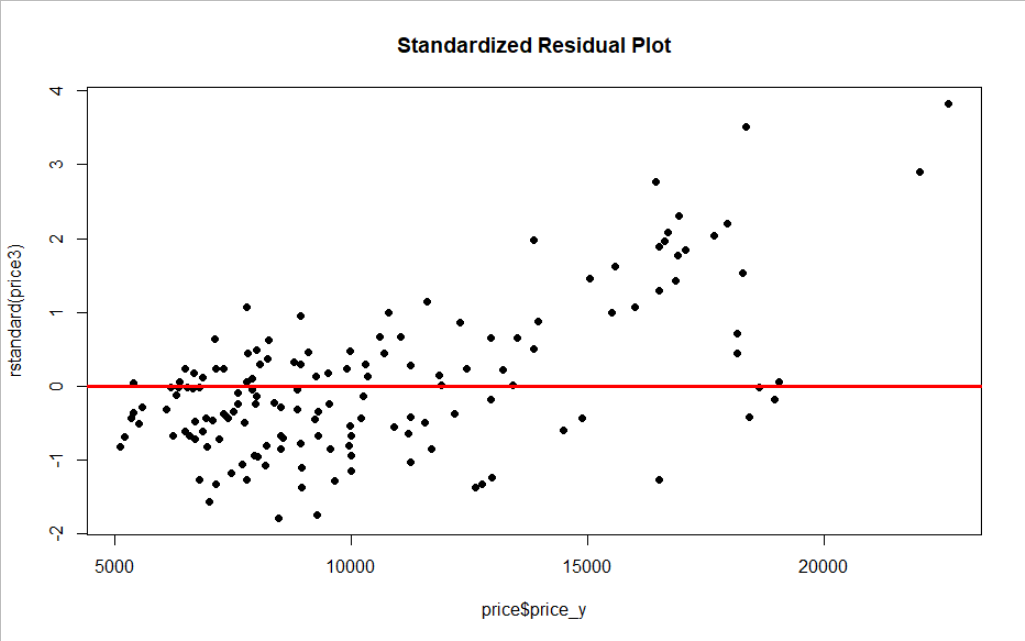
data: price3

DW = 1.2107, p-value = 1.426e-07

alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test, DW=1.2107, p-value = 1.426e-07, residuals are autocorrelated.

Equality of Variances:



**Case\_8. (y, X1, X2, X1X2)**

**Analysis Results:**

a) Multiple Regression Using Interaction Term:

> # Q4 (y,X1,X2,X1X2)

> # Multiple regression for Horsepower and Peak RPM with interactions

> price4=lm(price\_y~horsepower\_x1+peakrpm\_x2+I(horsepower\_x1\*peakrpm\_x2),

+ data=price)

> summary(price4)

Call:

lm(formula = price\_y ~ horsepower\_x1 + peakrpm\_x2 + I(horsepower\_x1 \*

peakrpm\_x2), data = price)

Residuals:

Min 1Q Median 3Q Max

-4624.4 -1514.2 -395.2 888.5 9756.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.682e+04 1.093e+04 3.369 0.000951 \*\*\*

horsepower\_x1 -2.090e+02 1.202e+02 -1.738 0.084118 .

peakrpm\_x2 -7.245e+00 2.132e+00 -3.399 0.000861 \*\*\*

I(horsepower\_x1 \* peakrpm\_x2) 6.334e-02 2.332e-02 2.716 0.007365 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2509 on 155 degrees of freedom

Multiple R-squared: 0.5985, Adjusted R-squared: 0.5907

F-statistic: 77.01 on 3 and 155 DF, p-value: < 2.2e-16

> confint(price4)

2.5 % 97.5 %

(Intercept) 1.523529e+04 58412.3568710

horsepower\_x1 -4.464565e+02 28.4837584

peakrpm\_x2 -1.145519e+01 -3.0339920

I(horsepower\_x1 \* peakrpm\_x2) 1.726574e-02 0.1094084

Dependent variable (price\_y)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

Third Slope for the sample with the interaction term

First Independent variable (horsepower\_x1)

Second Independent variable (peakrpm\_x2)

Third Independent variable (Interaction term: horsepower\_x1\*peakrpm\_x2)

R-squared: 0.5985, Adjusted R-squared: 0.5907

b) Observation:

1. R-squared (Coefficient of determination) is 59.85%, which means about 59.85% of the variance in price is explained by the independent variable’s horsepower, peak RPM and the Interaction term.

2. Interpretation of the Regression Equation:

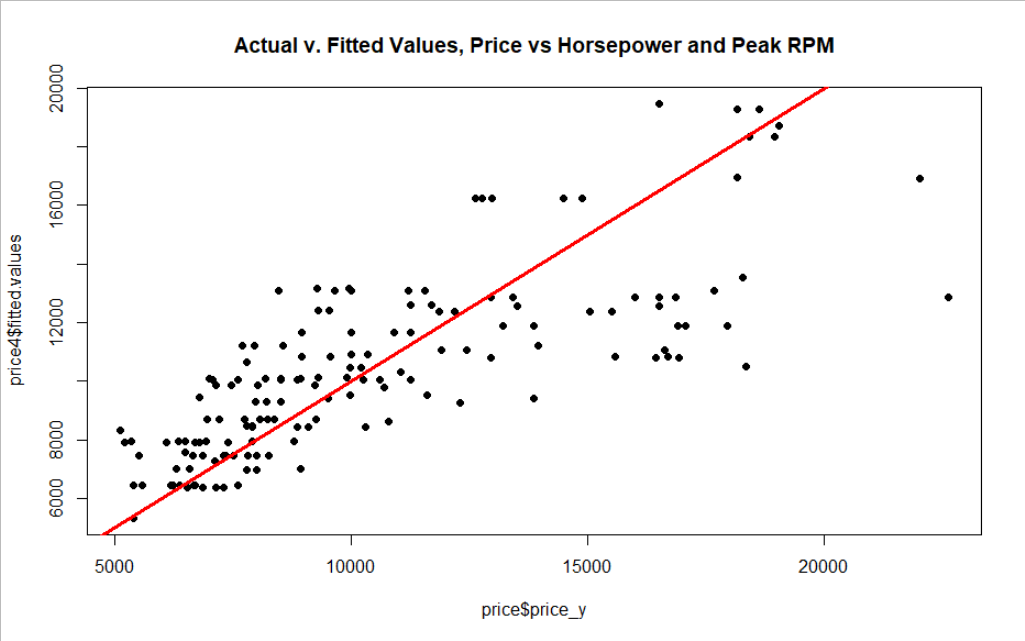
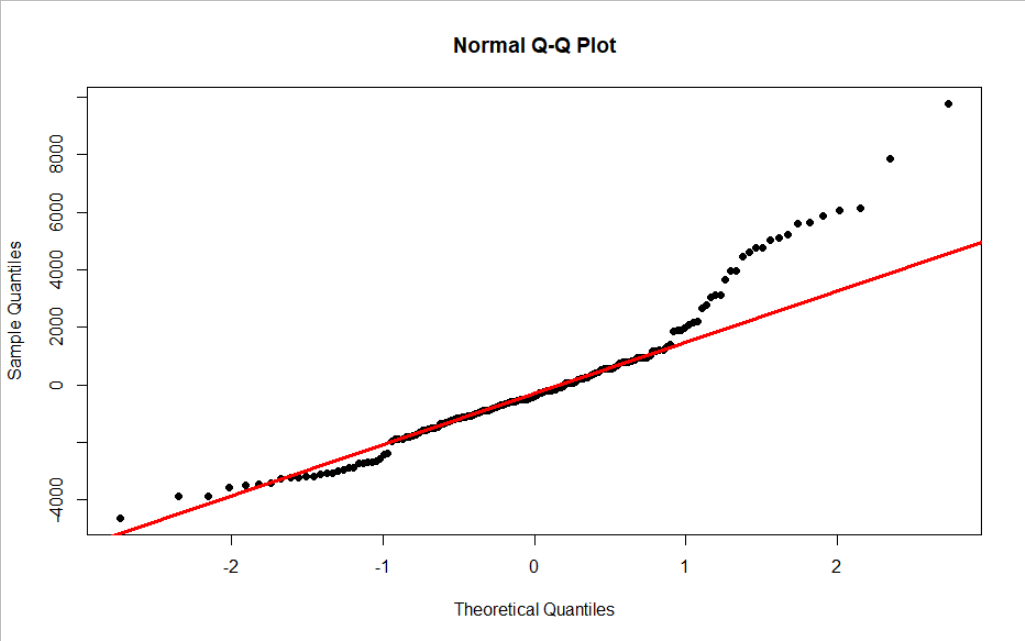
|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 3.682e+04 | No change in the other two variables horsepower and peak RPM. | The predicted mean price increases by $3.682e+04. |
| *ß1* = -2.090e+02 | Horsepower increases by 1hp. | The predicted mean price changes by $[(6.334e-02)-(2.090e+02)] with effect of horsepower on price is different for different values of peak RPM. |
| *ß2* = -7.245e+00 | Peak RPM increases by 1RPM. | The predicted mean price changes by $[(6.334e-02)-(7.245e+00)] with effect of peak RPM on price is different for different values of horsepower. |

3. p-value interpretations as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| p-value | *ß0* = 0.000951 | *ß1* = 0.084118 | *ß2* = 0.000861 | *ß3* = 0.007365 |
| Independent Variables | Reject the Null, Significant | Fail to Reject the Null, Not Significant | Reject the Null, Significant | Reject the Null, Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Highly Significant | | | |

c) L.I.N.E Assumptions:

Linearity & Normality:

Independence of errors:

> # Independence of the observations.

> dwtest(price4)

Durbin-Watson test

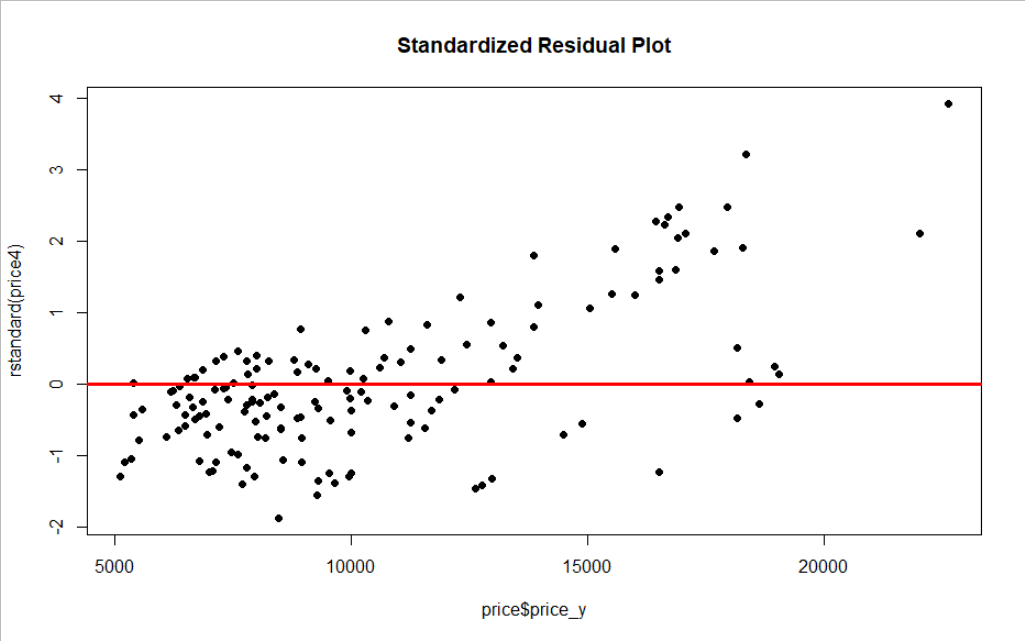
data: price4

DW = 0.95068, p-value = 5.47e-12

alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test, DW=0.95068, p-value = 5.47e-12, residuals are autocorrelated.

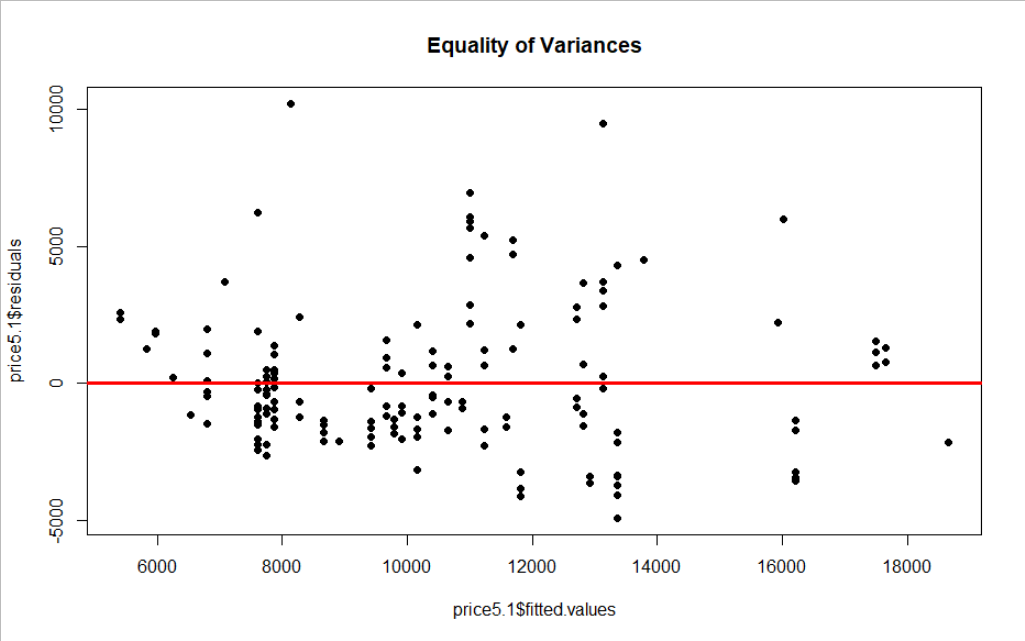
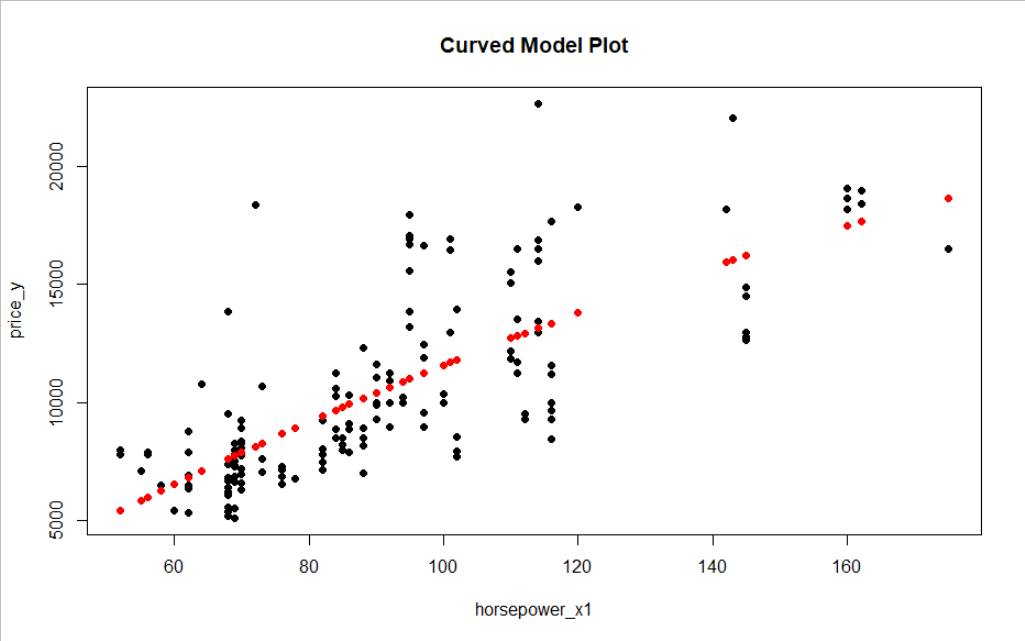
Equality of Variances:



**Case\_9. (y, X1, X1²)**

**Analysis Results:**

a) Curved and Equality of Variances Plots:



b) Simple Regression for Non-Linearity Test:

> # 5.1 (y,X1,X1^2)

> # Simple regression with correcting for the non-linearity for Horsepower.

> price5.1=lm(price\_y~horsepower\_x1+I(horsepower\_x1^2),data=price)

> summary(price5.1)

Call:

lm(formula = price\_y ~ horsepower\_x1 + I(horsepower\_x1^2), data = price)

Residuals:

Min 1Q Median 3Q Max

-4904.6 -1591.3 -665.5 1199.9 10206.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2717.9476 2561.4304 -1.061 0.29028

horsepower\_x1 170.7897 51.4511 3.319 0.00112 \*\*

I(horsepower\_x1^2) -0.2779 0.2436 -1.141 0.25566

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2649 on 156 degrees of freedom

Multiple R-squared: 0.5494, Adjusted R-squared: 0.5436

F-statistic: 95.09 on 2 and 156 DF, p-value: < 2.2e-16

Dependent variable (price\_y)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

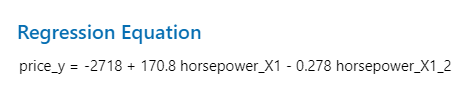
First Independent variable (horsepower\_x1)

Second Independent variable corrected for Non-Linearity (horsepower\_x1^2)

R-squared: 0.5494, Adjusted R-squared: 0.5436

c) Observation:

1. R-squared (Coefficient of determination) is 54.94%, which means about 54.94% of the variance in price is explained by the independent variable’s horsepower and horsepower^2. Regression equation as follows,



2. Interpretation of the Regression Equation:

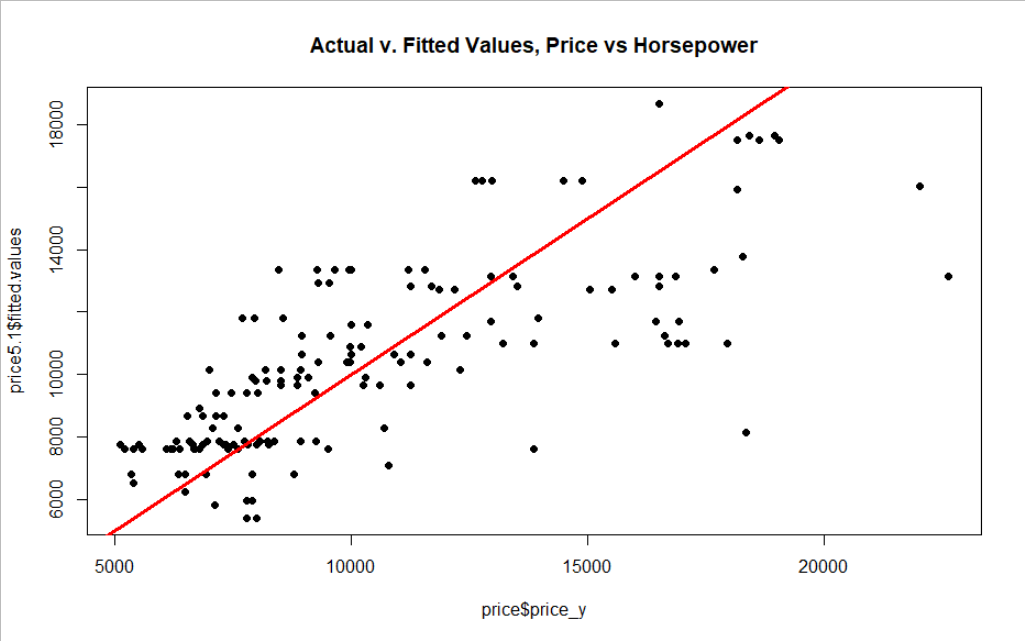
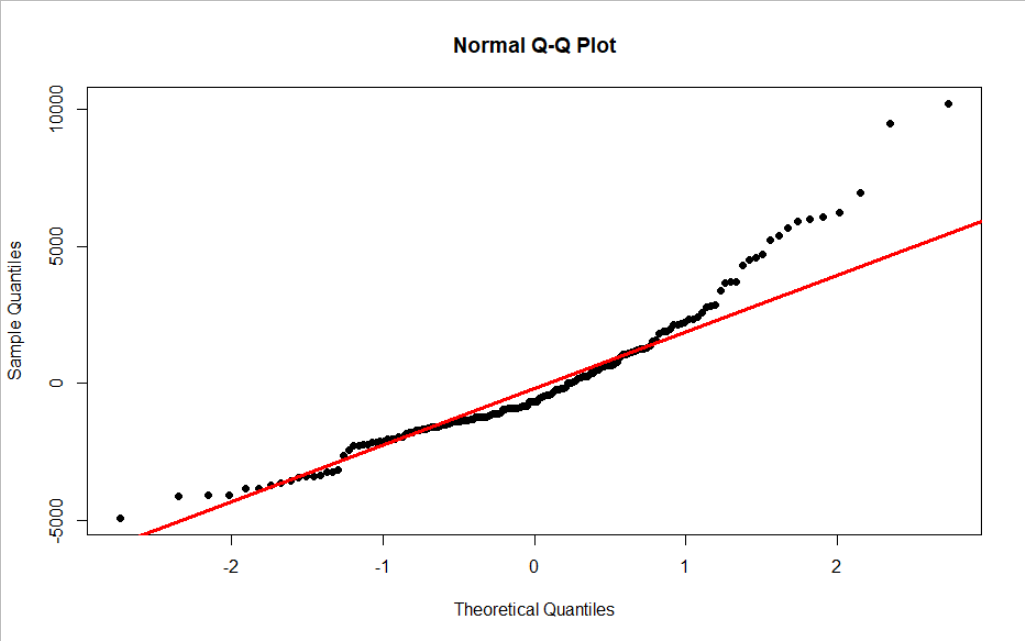
|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = -2717.94 | No change in the horsepower. | The predicted mean price decreases by $2717.94. |
| *ß1* = 170.78 | Horsepower increases by 1hp. | The predicted mean price increases by $170.50. |

3. p-value interpretations as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| p-value | *ß0* = 0.29028 | *ß1* = 0.00112 | *ß2* = 0.25566 |
| Independent Variables | Fail to Reject the Null, Not Significant | Reject the Null, Significant | Fail to Reject the Null, Not Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Significant | | |

d) L.I.N.E Assumptions:

Linearity & Normality:

Independence of errors:

> # Independence of the observations.

> dwtest(price5.1)

Durbin-Watson test

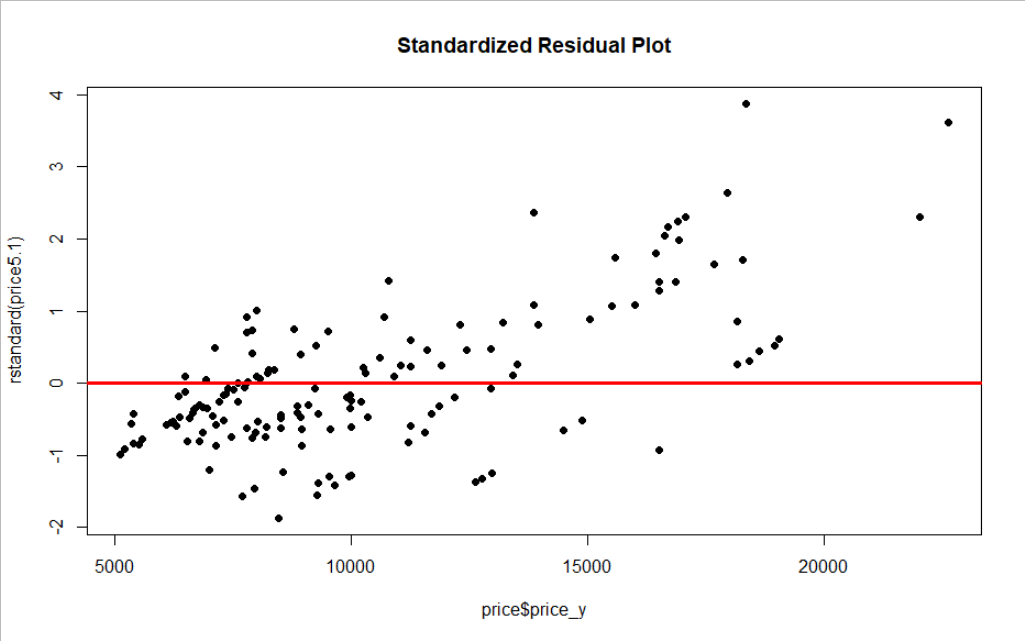
data: price5.1

DW = 0.97814, p-value = 2.741e-11

alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test, DW=0.97814, p-value = 2.741e-11, residuals are autocorrelated.

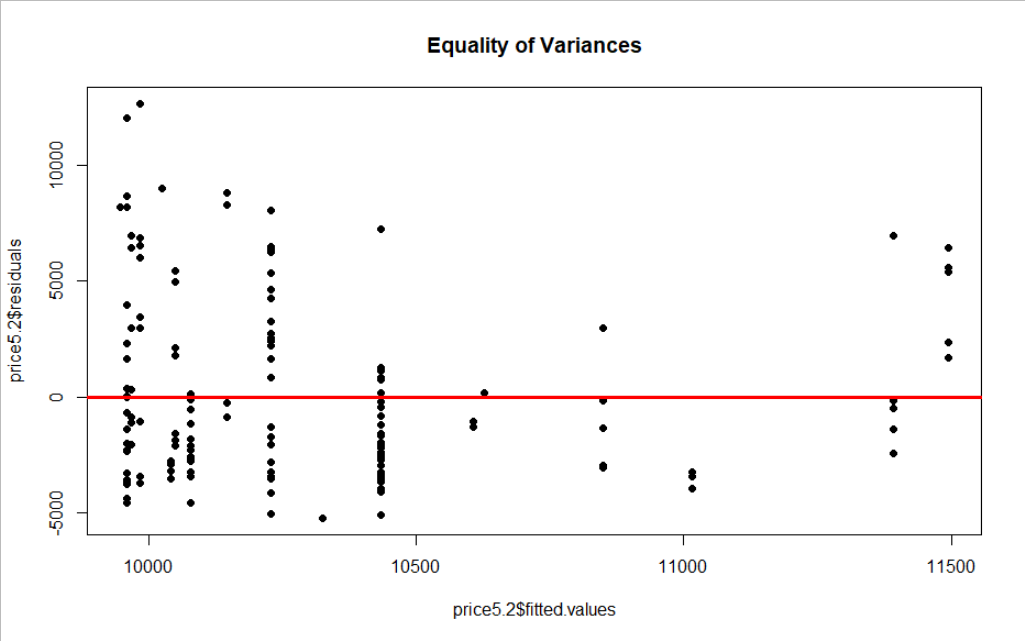
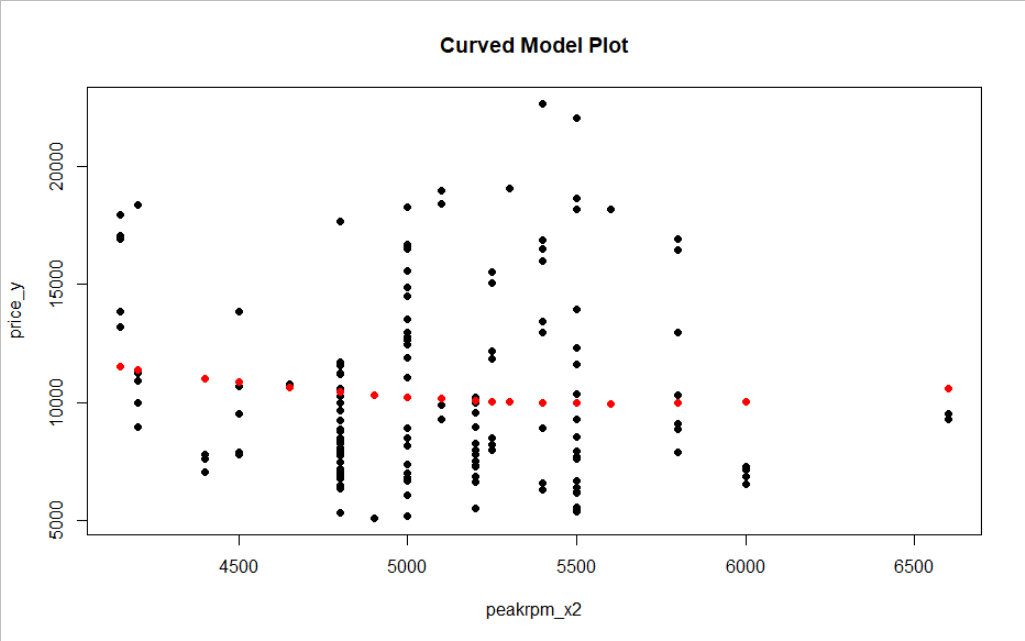
Equality of Variances:



**Case\_10. (y, X2, X2²)**

**Analysis Results:**

a) Curved and Equality of Variances Plots:



b) Simple Regression for Non-Linearity Test:

> # 5.2 (y,X2,X2^2)

> # Simple regression with correcting for the non-linearity for Peak RPM.

> price5.2=lm(price\_y~peakrpm\_x2+I(peakrpm\_x2^2),data=price)

> summary(price5.2)

Call:

lm(formula = price\_y ~ peakrpm\_x2 + I(peakrpm\_x2^2), data = price)

Residuals:

Min 1Q Median 3Q Max

-5207 -2933 -1308 2166 12640

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.229e+04 2.493e+04 1.295 0.197

peakrpm\_x2 -7.934e+00 9.681e+00 -0.820 0.414

I(peakrpm\_x2^2) 7.044e-04 9.370e-04 0.752 0.453

Residual standard error: 3927 on 156 degrees of freedom

Multiple R-squared: 0.009874, Adjusted R-squared: -0.00282

F-statistic: 0.7779 on 2 and 156 DF, p-value: 0.4612

Dependent variable (price\_y)

Y-Intercept for the sample

First Slope for the sample

Second Slope for the sample

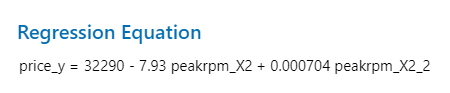
First Independent variable (peakrpm\_x2)

Second Independent variable corrected for Non-Linearity (peakrpm\_x2^2)

R-squared: 0.009874, Adjusted R-squared: -0.00282

c) Observation:

1. R-squared (Coefficient of determination) is 0.9%, which means about 0.9% of the variance in price is explained by the independent variable’s peak RPM and peak RPM^2. Regression equation as follows,



2. Rejecting this case due to high p-values and low R-squared.

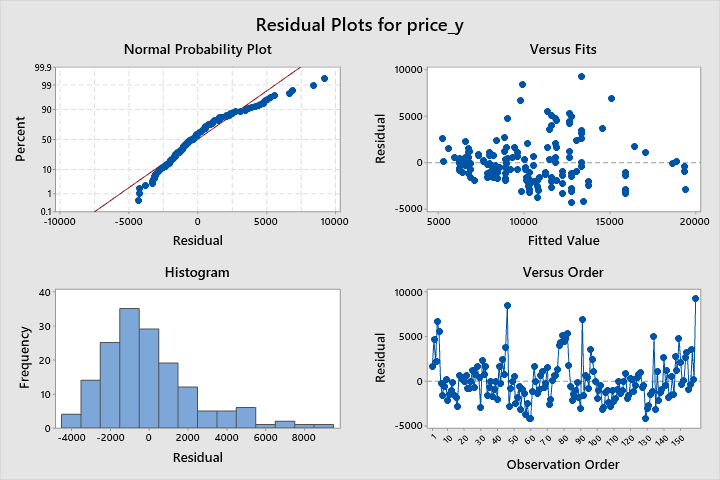
**BEST FIT:**

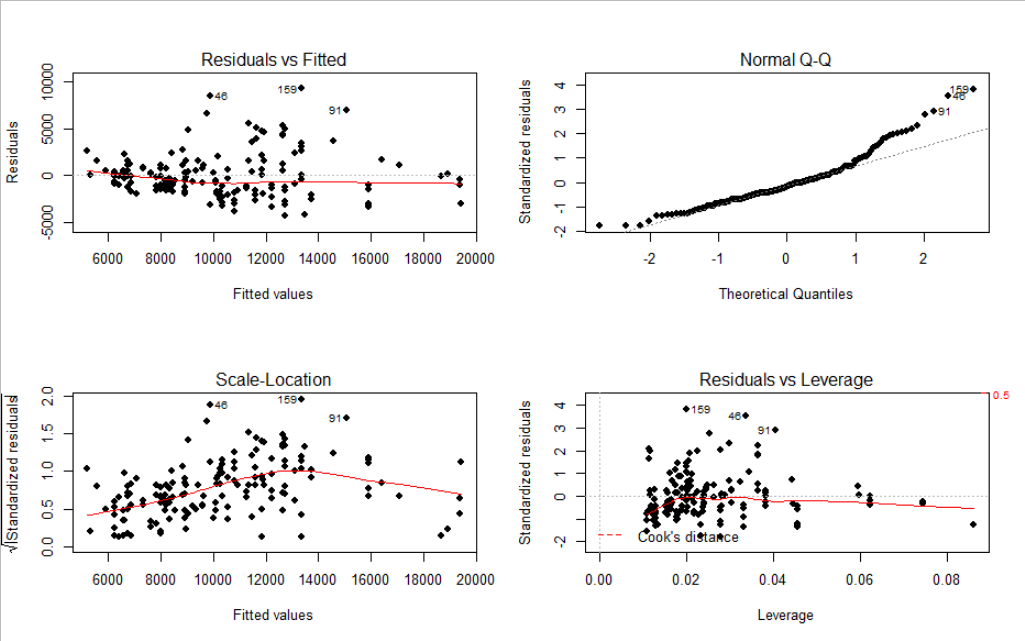
**Case\_7. (y, X1, X2, X3)**

**a) Justification for selecting this model as “Best Fit”:**

1. Out of the 10 models we tested for regression, we found this model to be the best fit. We choose this model based on the highest R-squared (61.8%) and Adjusted R-squared (61.06%) which means about 61% of the variance in price is explained by the independent variable’s horsepower, peak RPM and No. of doors.

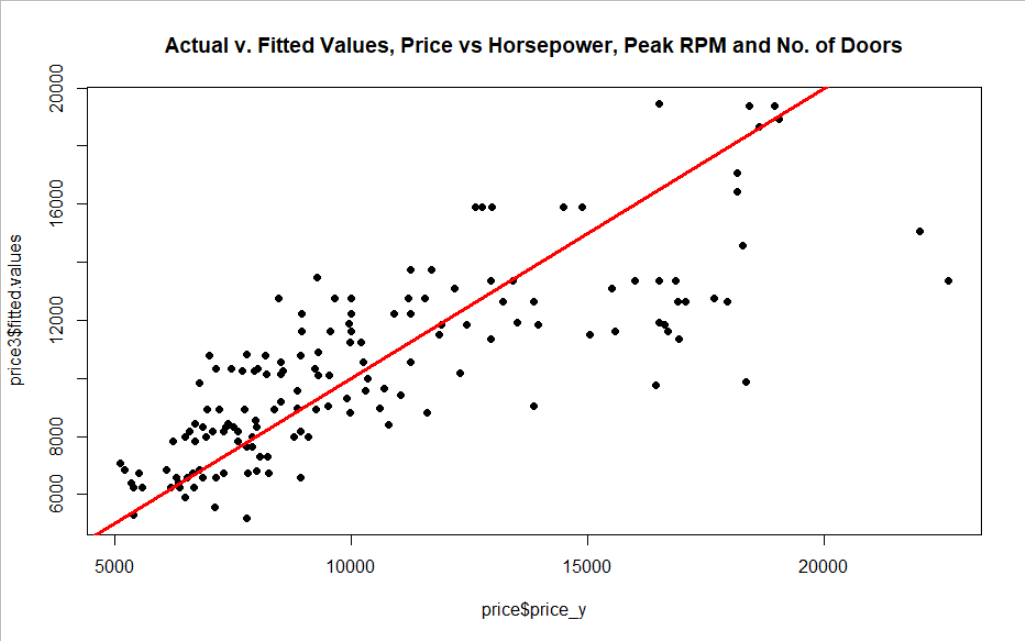
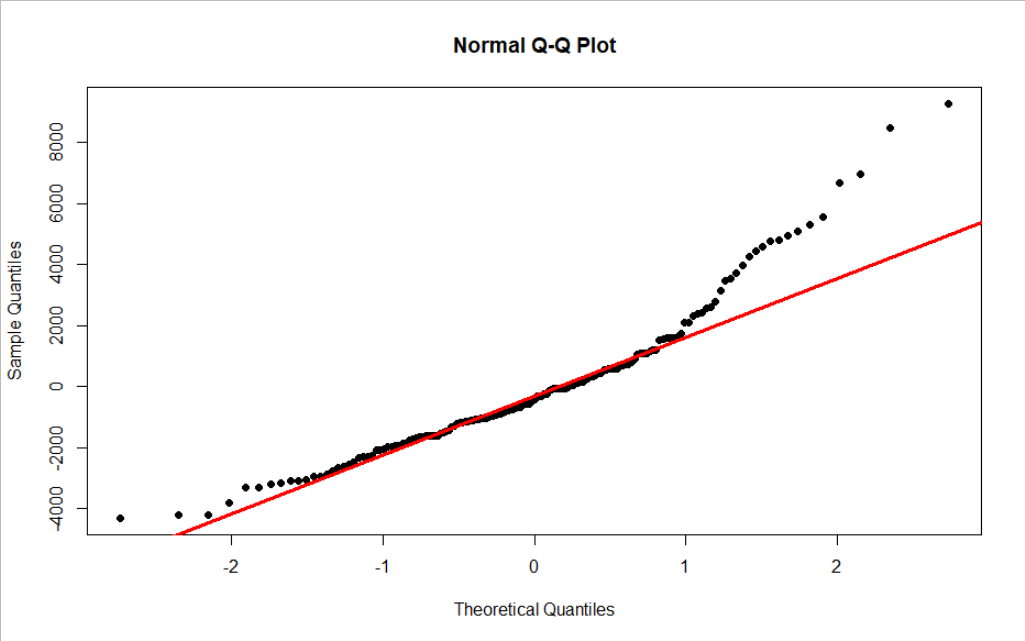
2. This model also has the best distributed residuals which is also one of the factors for choosing this model. From the Versus Fits graph below, we can see that most of our data points are close to zero and there is no upward or downward bias in the plot.





**b) L.I.N.E Assumptions:**

**Linearity & Normality:**

1. This model does voilate the Linearity assumption, as its curved at the end. But since the R-squared is the highest in this case, we chose this model above others.

2. From the histogram and Q-Q plots above, the model doesn’t deviate much from the Normality plot so we say it satisfies the Normality assumption.

**Independence of errors:**

> # Independence of the observations.

> dwtest(price3)

Durbin-Watson test

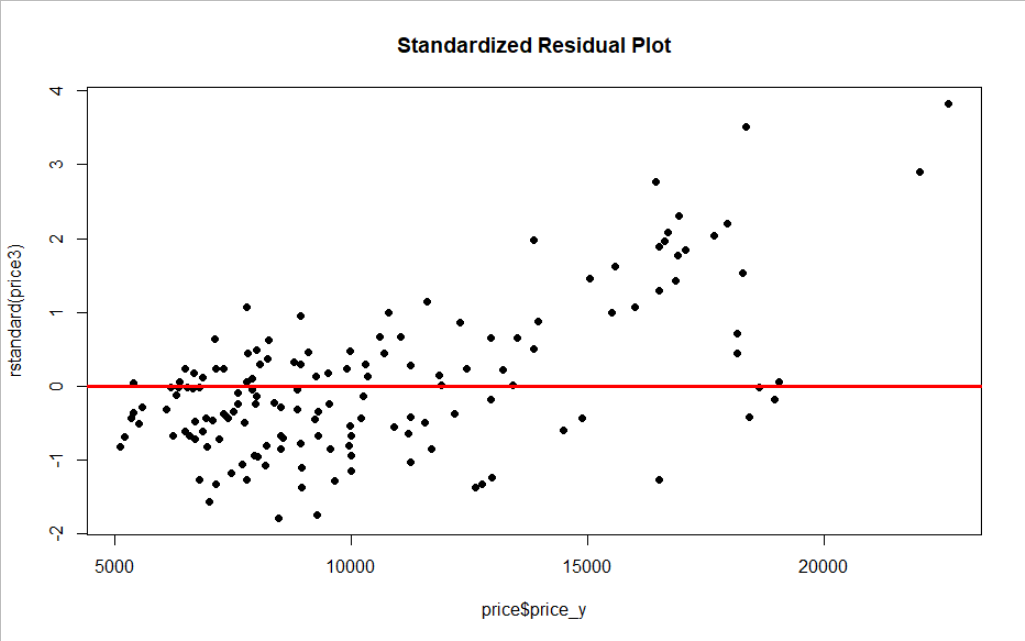
data: price3

DW = 1.2107, p-value = 1.426e-07

alternative hypothesis: true autocorrelation is greater than 0

From Durbin-Watson test, DW=1.2107, p-value = 1.426e-07, residuals are positively autocorrelated which means we have autocorrelation of residuals. This means that our model violates Independence of errors assumption.

**Equality of Variances:**

Here, we do not see the residuals “dots” fanning out in any triangular fashion, so we can say Equality of Variance assumption is met.

**c) Interpretations of models slope and intercept coefficients:**

Residuals:

Min 1Q Median 3Q Max

-4290.8 -1605.8 -432.5 990.2 9252.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6489.3409 2227.9719 2.913 0.004113 \*\*

horsepower\_x1 117.7856 7.6624 15.372 < 2e-16 \*\*\*

peakrpm\_x2 -1.2119 0.4366 -2.776 0.006187 \*\*

doornumber\_x31 -1595.4907 402.9991 -3.959 0.000114 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2447 on 155 degrees of freedom

Multiple R-squared: 0.618, Adjusted R-squared: 0.6106

F-statistic: 83.59 on 3 and 155 DF, p-value: < 2.2e-16

**Price for a car in dollars (Dependent/Response variable (price\_y))**

**Y-Intercept for the sample**

**Slope of price with horsepower\_x1, holding peakrpm\_x2 and doornumber\_x3 constant (First Regression Coefficient for the sample)**

**Slope of price with peakrpm\_x2, holding horsepower\_x1 and doornumber\_x3 constant (Second Regression Coefficient for the sample)**

**Slope of price with doornumber\_x3, holding horsepower\_x1 and peakrpm\_x2 constant (Third Regression Coefficient for the sample)**

**Horsepower in hp (First Independent/Explanatory/Predictor variable (horsepower\_x1))**

**Peak RPM in RPM (Second Independent/Explanatory/Predictor variable (peakrpm\_x2))**

**No. of Doors on the car (Third Independent/Explanatory/Predictor variable (doornumber\_x3, Two Door = 1))**

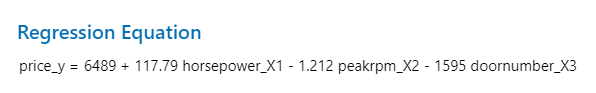
1. Interpretation of the Regression Equation:

|  |  |  |
| --- | --- | --- |
| *ß - Coefficient* | Independent Variable Change | Net Effect |
| *ß0* = 6489.34 | No change in the other three variables horsepower, peak RPM and No. of doors. | The predicted mean price increases by $6489.34. |
| *ß1* = 117.78 | Horsepower increases by 1hp. | The predicted mean price increases by $117.78 holding peak RPM and No. of doors constant. |
| *ß2* = -1.212 | Peak RPM increases by 1RPM. | The predicted mean price decreases by $1.212 holding horsepower and No. of doors constant. |
| *ß3* = -1595.49 | No. of doors increases by one. | The predicted mean price decreases by $1595.49 holding horsepower and peak RPM constant. |

2. p-value interpretations as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| p-value | *ß0* = 0.004113 | *ß1* < 2e-16 | *ß2* = 0.006187 | *ß3* = 0.000114 |
| Independent Variables | Reject the Null, Significant | Reject the Null, Highly Significant | Reject the Null, Significant | Reject the Null, Significant |
| Overall p-value < 2.2e-16 | Reject the Null, Highly Significant | | | |

**d) Estimation and prediction intervals:**



\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**Business Study:** Let’s say the company is going to introduce a new model and they want to predict the pricing, the new model specifications are as below:

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**Estimate\_1:** (High End Sports Model)

Horsepower: 200HP

Peak RPM: 7000RPM

No. of Doors: Two

Plugging-in the values in the Regression equation,

So, the predicted price of the new model is **$19965.85.**

**Estimate\_2:** (Cheapest Model)

Horsepower: 50HP

Peak RPM: 3500RPM

No. of Doors: Two

Plugging-in the values in the Regression equation,

So, the predicted price of the new model is **$6540.85.**

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Prediction Interval as follows:

> # Prediction Interval

> sun1=predict(price3,price,interval = "predict")

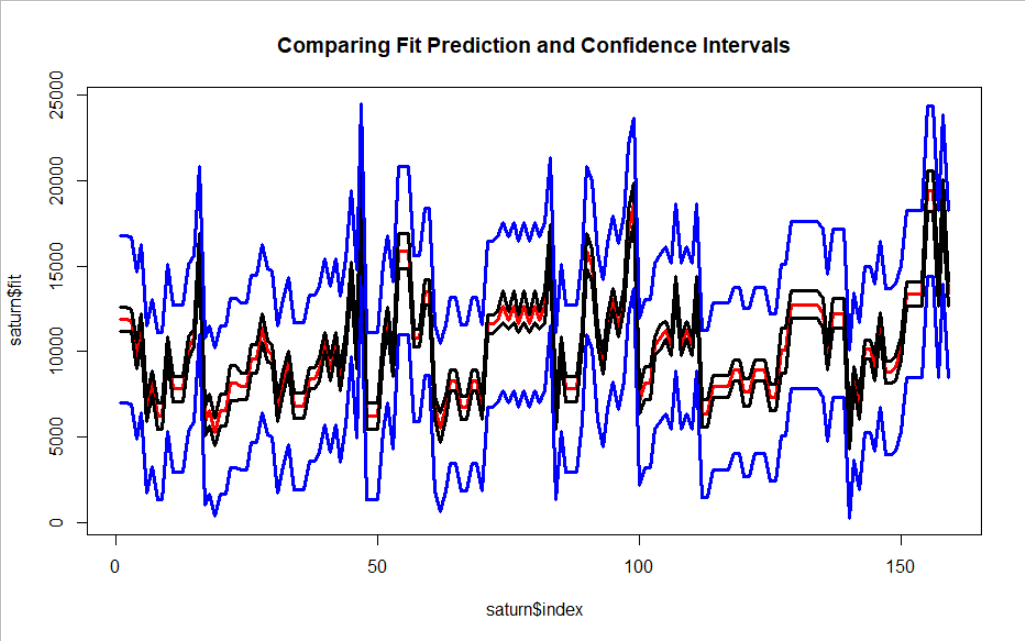
> max(sun1)

[1] 24484.85

> min(sun1)

[1] 285.138

Comparing Fit Prediction and Confidence Intervals:



-----------------------------------------------------------------------------------------------------------

-- | End | --