Multiple Linear Regression

Sagar Vora

2024-03-11

To have a look at the data set-

```
carprice = read.csv("/Users/sagarvora/Downloads/CarPrice/CarPrice_Assignment.csv")
head(carprice)
```

```
##   car_ID symboling                  CarName fueltype aspiration doornumber
## 1      1         3       alfa-romero giulia      gas        std        two
## 2      2         3      alfa-romero stelvio      gas        std        two
## 3      3         1 alfa-romero Quadrifoglio      gas        std        two
## 4      4         2              audi 100 ls      gas        std       four
## 5      5         2               audi 100ls      gas        std       four
## 6      6         2                 audi fox      gas        std        two
##       carbody drivewheel enginelocation wheelbase carlength carwidth carheight
## 1 convertible        rwd          front      88.6     168.8     64.1      48.8
## 2 convertible        rwd          front      88.6     168.8     64.1      48.8
## 3   hatchback        rwd          front      94.5     171.2     65.5      52.4
## 4       sedan        fwd          front      99.8     176.6     66.2      54.3
## 5       sedan        4wd          front      99.4     176.6     66.4      54.3
## 6       sedan        fwd          front      99.8     177.3     66.3      53.1
##   curbweight enginetype cylindernumber enginesize fuelsystem boreratio stroke
## 1       2548       dohc           four        130       mpfi      3.47   2.68
## 2       2548       dohc           four        130       mpfi      3.47   2.68
## 3       2823       ohcv            six        152       mpfi      2.68   3.47
## 4       2337        ohc           four        109       mpfi      3.19   3.40
## 5       2824        ohc           five        136       mpfi      3.19   3.40
## 6       2507        ohc           five        136       mpfi      3.19   3.40
##   compressionratio horsepower peakrpm citympg highwaympg price
## 1              9.0        111    5000      21         27 13495
## 2              9.0        111    5000      21         27 16500
## 3              9.0        154    5000      19         26 16500
## 4             10.0        102    5500      24         30 13950
## 5              8.0        115    5500      18         22 17450
## 6              8.5        110    5500      19         25 15250
```

Full model and it's summary-

```
lm_full <- lm(price ~ doornumber + carbody + fueltype + carlength + horsepower + citympg + highwaympg,
summary(lm_full)
```

```
##
## Call:
## lm(formula = price ~ doornumber + carbody + fueltype + carlength +
##     horsepower + citympg + highwaympg, data = carprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9734.8 -2368.4  -127.2  1998.8 14482.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -24077.83    9546.38  -2.522  0.01247 *
## doornumbertwo        88.63     839.02   0.106  0.91598
## carbodyhardtop    -2747.83    2049.46  -1.341  0.18157
## carbodyhatchback  -6945.66    1623.88  -4.277 2.97e-05 ***
## carbodysedan      -5809.81    1756.82  -3.307  0.00112 **
## carbodywagon      -7837.67    1937.69  -4.045 7.55e-05 ***
## fueltypegas       -3233.37    1102.78  -2.932  0.00377 **
## carlength           191.72      42.31   4.531 1.03e-05 ***
## horsepower          131.87      11.84  11.141  < 2e-16 ***
## citympg             383.66     191.55   2.003  0.04658 *
## highwaympg         -337.73     172.05  -1.963  0.05108 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3745 on 194 degrees of freedom
## Multiple R-squared:  0.791,  Adjusted R-squared:  0.7802
## F-statistic: 73.42 on 10 and 194 DF,  p-value: < 2.2e-16
```

Null model and it's summary-

```
lm_null <- lm(price ~ 1, data = carprice)
summary(lm_null)
```

```
##
## Call:
## lm(formula = price ~ 1, data = carprice)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8159  -5489  -2982   3226  32123
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13277        558    23.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7989 on 204 degrees of freedom
```

Analysis of variance(ANOVA) of the null and the full model-

```
anova(lm_null,lm_full)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ 1
## Model 2: price ~ doornumber + carbody + fueltype + carlength + horsepower +
##     citympg + highwaympg
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    204 1.3020e+10
## 2    194 2.7212e+09 10 1.0298e+10 73.421 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The first reduced modlel and it’s summary-

Removed the predictors doornumber and highwaympg from the full model.

```
lm_1 <- lm(price ~ + carbody + fueltype + carlength + horsepower + citympg , data = carprice)
summary(lm_1)
```

```
##
## Call:
## lm(formula = price ~ +carbody + fueltype + carlength + horsepower +
##     citympg, data = carprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9851.9 -2327.7  -401.1  1977.2 14318.2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -27785.90    9032.92  -3.076 0.002397 **
## carbodyhardtop    -2902.38    2057.48  -1.411 0.159933
## carbodyhatchback  -7076.67    1629.10  -4.344 2.24e-05 ***
## carbodysedan      -6066.39    1672.66  -3.627 0.000366 ***
## carbodywagon      -7872.57    1831.86  -4.298 2.72e-05 ***
## fueltypegas       -3496.09    1099.80  -3.179 0.001719 **
## carlength           205.22      40.52   5.064 9.44e-07 ***
## horsepower          130.80      11.66  11.213  < 2e-16 ***
## citympg              48.06      87.51   0.549 0.583494
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3764 on 196 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.778
## F-statistic: 90.39 on 8 and 196 DF,  p-value: < 2.2e-16
```

Removed the predictor citympg from model 1.

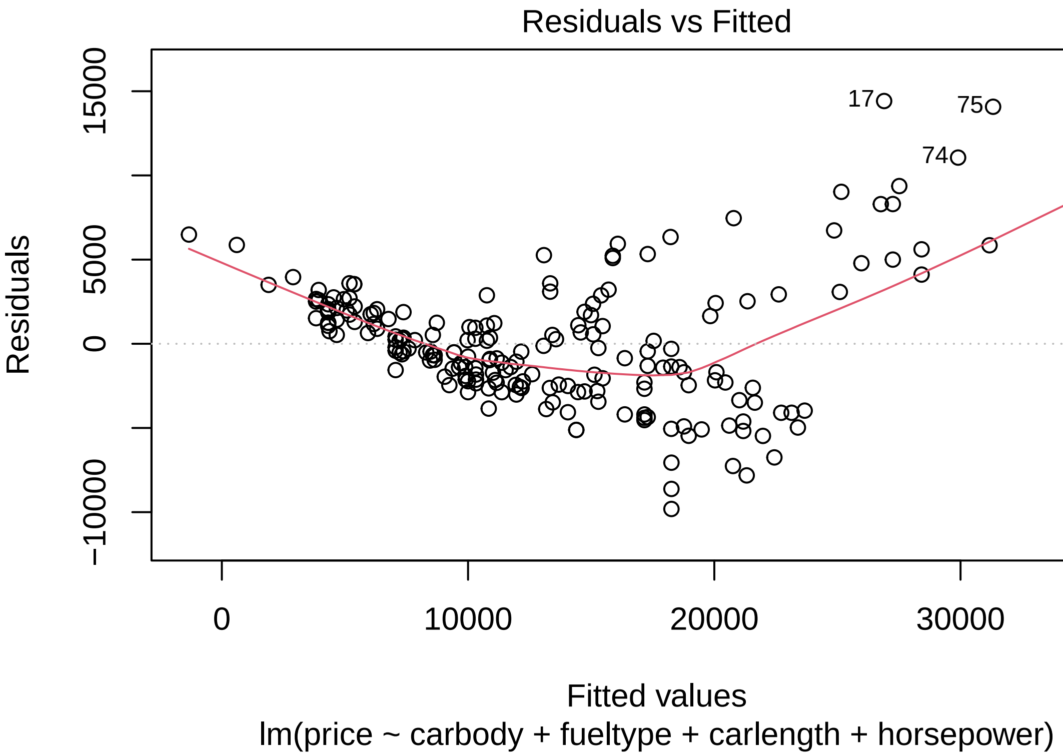
Second and the final reduced model and it's summary-

```
lm_2<- lm(price ~ carbody + fueltype + carlength + horsepower, data = carprice)
summary(lm_2)
```

```
##
## Call:
## lm(formula = price ~ carbody + fueltype + carlength + horsepower,
##     data = carprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9810.5 -2316.0  -439.9  1908.2 14417.7
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -23831.961   5445.735  -4.376 1.96e-05 ***
## carbodyhardtop     -2759.000   2037.226  -1.354 0.177194
## carbodyhatchback   -6964.090   1613.287  -4.317 2.51e-05 ***
## carbodysedan       -5891.939   1639.306  -3.594 0.000411 ***
## carbodywagon       -7722.881   1808.259  -4.271 3.03e-05 ***
## fueltypegas        -3774.519    974.252  -3.874 0.000146 ***
## carlength            192.316     32.962   5.834 2.19e-08 ***
## horsepower           127.059      9.457  13.436  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3757 on 197 degrees of freedom
## Multiple R-squared:  0.7864, Adjusted R-squared:  0.7788
## F-statistic: 103.6 on 7 and 197 DF,  p-value: < 2.2e-16
```

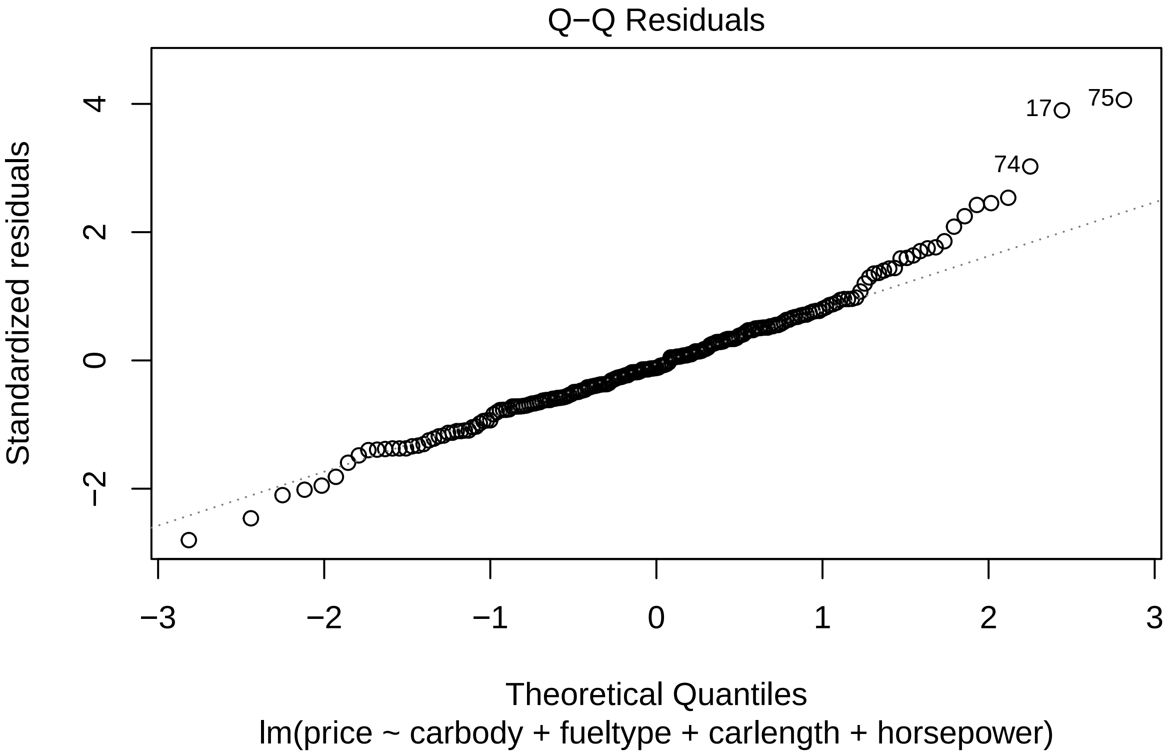
Residuals vs Fitted Plot-

```
plot(lm_2,which = 1)
```



Normal Q-Q plot/Scatterplot-

```
plot(lm_2,which = 2)
```



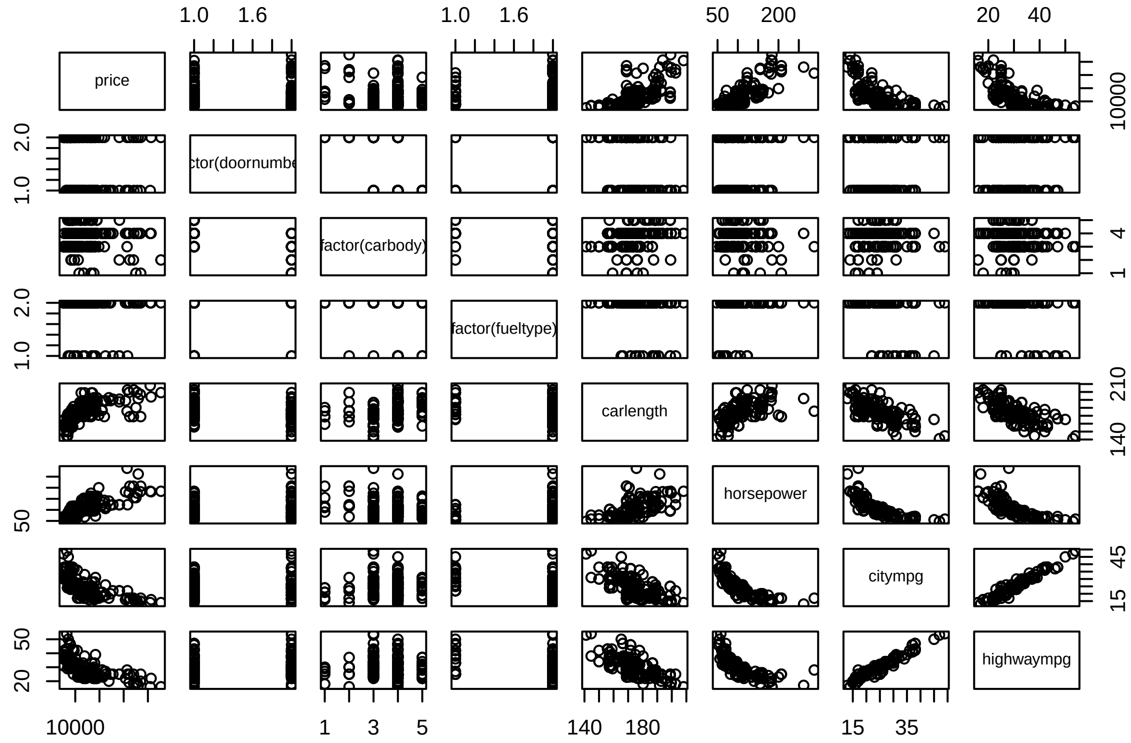
Analysis of Variance (ANOVA) of the final model and the full model-

```
anova(lm_2,lm_full)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ carbody + fueltype + carlength + horsepower
## Model 2: price ~ doornumber + carbody + fueltype + carlength + horsepower +
##     citympg + highwaympg
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    197 2780690444
## 2    194 2721173467  3  59516977 1.4144 0.2398
```

Scatterplot Matrix for the full model-

```
pairs(price ~ factor(doornumber) + factor(carbody) + factor(fueltype) + carlength + horsepower + citympg
```

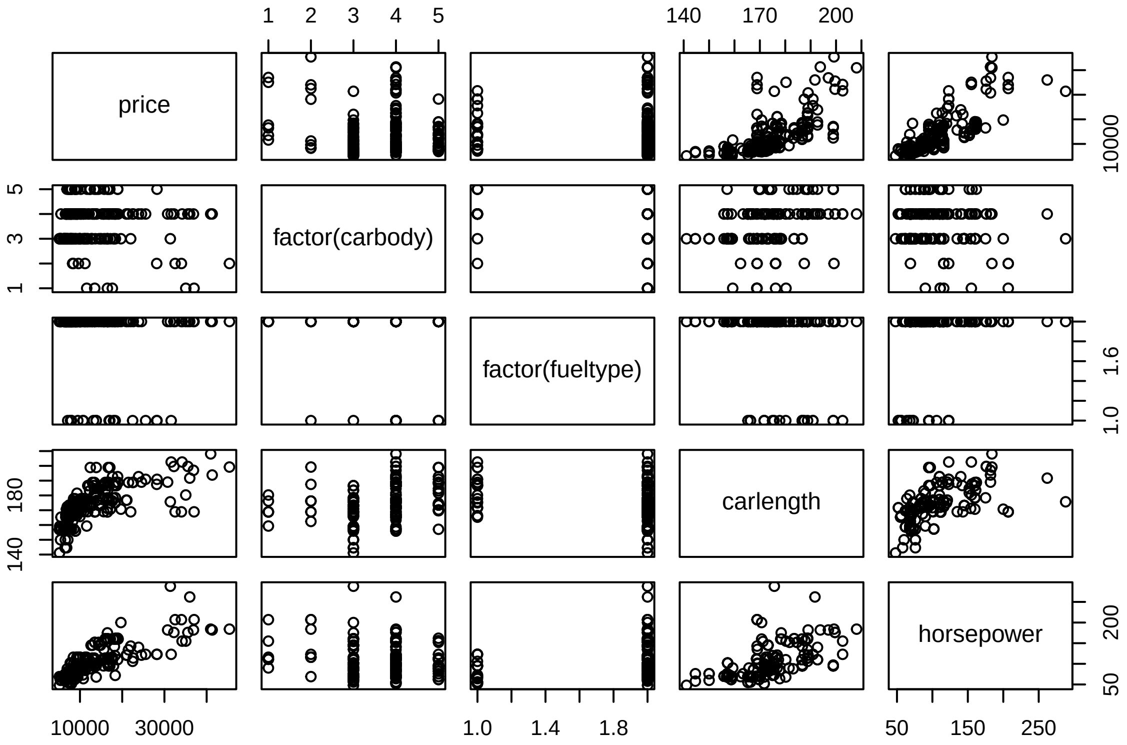


```
summary(lm_full)
```

```
##
## Call:
## lm(formula = price ~ doornumber + carbody + fueltype + carlength +
##     horsepower + citympg + highwaympg, data = carprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9734.8 -2368.4  -127.2  1998.8 14482.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -24077.83    9546.38  -2.522  0.01247 *
## doornumbertwo        88.63     839.02   0.106  0.91598
## carbodyhardtop    -2747.83    2049.46  -1.341  0.18157
## carbodyhatchback  -6945.66    1623.88  -4.277 2.97e-05 ***
## carbodysedan      -5809.81    1756.82  -3.307  0.00112 **
## carbodywagon      -7837.67    1937.69  -4.045 7.55e-05 ***
## fueltypegas       -3233.37    1102.78  -2.932  0.00377 **
## carlength           191.72      42.31   4.531 1.03e-05 ***
## horsepower          131.87      11.84  11.141  < 2e-16 ***
## citympg             383.66     191.55   2.003  0.04658 *
## highwaympg         -337.73     172.05  -1.963  0.05108 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3745 on 194 degrees of freedom
## Multiple R-squared:  0.791,  Adjusted R-squared:  0.7802
## F-statistic: 73.42 on 10 and 194 DF,  p-value: < 2.2e-16
```

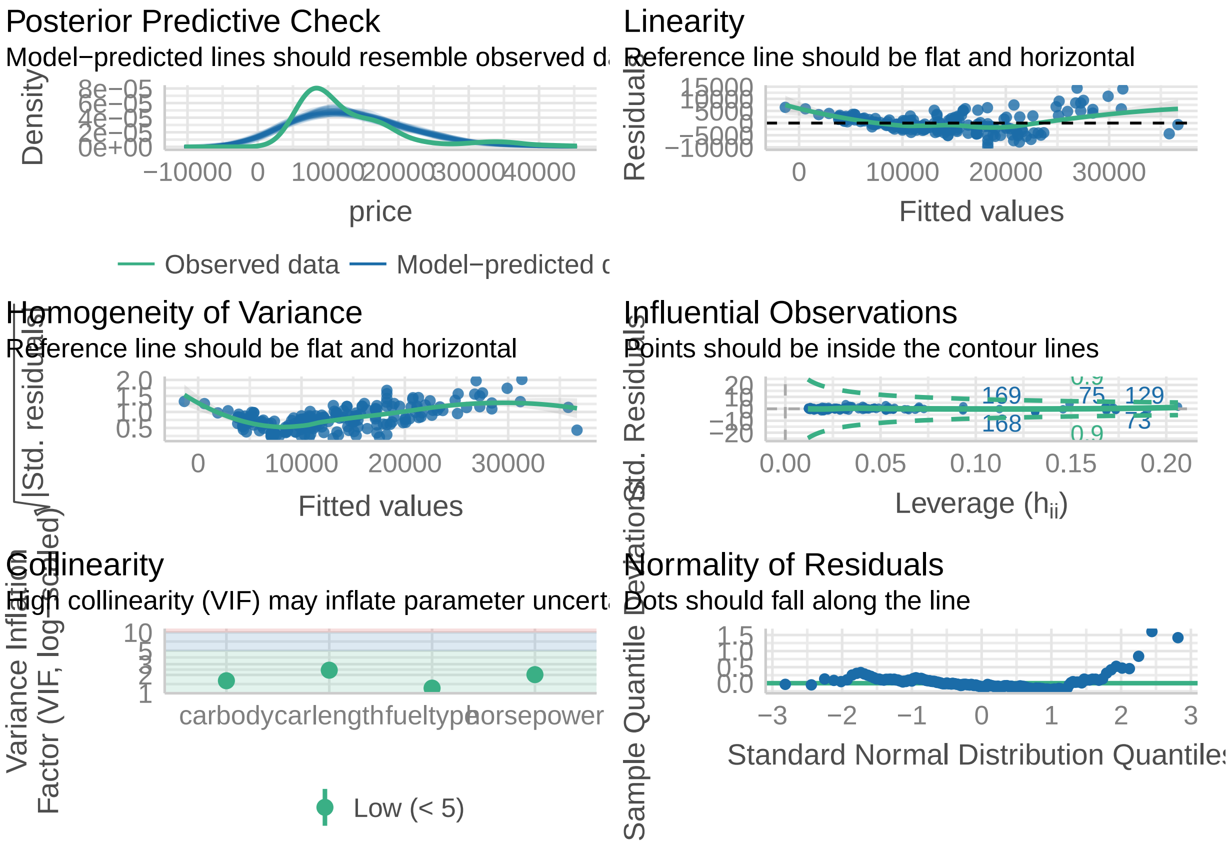
Scatterplot Matrix for the final model-

```
pairs(price ~ factor(carbody) + factor(fueltype) + carlength + horsepower, data = carprice)
```



## Checking all the assumptions for the final model-

```
performance::check_model(lm_2)
```


Posterior Predictive Check
Model−predicted lines should resemble observed d
Density
8e−05
6e−05
4e−05
2e−05
0e+00
−10000
0
10000
20000
30000
40000
price
Observed data
Model−predicted d
Linearity
Reference line should be flat and horizontal
Residuals
15000
10000
5000
0
−5000
−10000
0
10000
20000
30000
Fitted values
Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
2.0
1.5
1.0
0.5
0
10000
20000
30000
Fitted values
Influential Observations
Points should be inside the contour lines
Std. Residuals
20
10
0
−10
−20
0.9
169
75
129
168
73
0.9
0.00
0.05
0.10
0.15
0.20
Leverage (h_{ii})
Collinearity
High collinearity (VIF) may inflate parameter uncert
Variance Inflation
Factor (VIF, log-scaled)
10
5
3
1
carbody
carlength
fueltype
horsepower
Low (< 5)
Normality of Residuals
Dots should fall along the line
Sample Quantile Deviations
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3
Standard Normal Distribution Quantiles