Department of Mathematics and Actuarial Science

MACT 4231: Applied Regression Methods

Professor Ali S. Hadi

Final Regression Project

Cars Price Prediction

Youssef Nakhla

900201430

Contents

Problem Statement

Predicting the price of a car is no easy ordeal specifically since each car brand has numerous variations of car models and within each car model, there exists numerous

variations of cars varying in size, performance, etc. To the untrained eye or in simpler terms to anyone outside of the car industry, it may deem difficult to predict the price of a car with high accuracy, or even to the trained eye it would be a very difficult problem given the amount of possible combinations of features in cars. Hence, to tackle such problem we must turn to a method that can handle such complications, hence the goal of this project is to try and determine the most influential attributes that affect the price of a car and build a linear regression model to predict the price of any given car. This solution would help manufacturers and buyers alike in aiding to determine the best price of a car based on its capabilities and features.

Dataset

Source & Background

This dataset was obtained on Kaggle, a renowned platform for sharing datasets and work on said datasets. This dataset was created as part of a competition designed for students to practice the concepts covered academically. The context of this dataset is that there exists a Chinese car company that aspires to enter the US market and hence they would like to know what car features affect their price in the US market.

Description

As W. Edwards Deming says, “In God we trust, all others must bring data,” the data being used for such project must be carefully picked and examined to ensure the validity of the conclusions we will later reach. The dataset contains 26 variables; 1 response variable (the price) and 25 predictor variables with 205 observations with each observation representing a different car. A brief description of each variable can be found in the table below:

Column Name	Description	Data Type	Unit of Measurement
Car_ID	Unique id of each observation	Integer	N/A
Symboling	Assigned insurance risk rating	Categorical	N/A
CarName	Name of car company	Categorical	N/A
fueltype	Car fuel type (gas or diesel)	Categorical	N/A
aspiration	Aspiration used in a car	Categorical	N/A
doornumber	Number of doors in a car	Categorical	N/A
carbody	Body type of car	Categorical	N/A
drivewheel	Type of drive wheel	Categorical	N/A
enginelocation	Location of car engine	Categorical	N/A
wheelbase	Wheelbase of car	Numeric	Inches
carlength	Length of car	Numeric	Inches
carwidth	Width of car	Numeric	Inches
carheight	Height of car	Numeric	Inches
curbweight	Weight of car without occupants or baggage	Numeric	Pounds (lbs)
enginetype	Type of engine	Categorical	N/A
cylindernumber	Cylinder count in the car	Categorical	N/A
enginesize	Engine size	Numeric	Cubic inches
fuelsystem	Fuel system of car	Categorical	N/A
boreratio	Bore ratio of car	Numeric	Ratio
stroke	Stroke or volume inside the engine	Numeric	Inches
compressionratio	Compression ratio of car	Numeric	Ratio
horsepower	Horsepower	Numeric	Horsepower (hp)
peakrpm	Car peak rpm	Numeric	Revolutions per minute (RPM)
citympg	Mileage in city	Numeric	Miles per gallon (mpg)
highwaympg	Mileage on highway	Numeric	Miles per gallon (mpg)
price	Price of car	Numeric	US Dollars (USD)

Preprocessing

The first step to any linear regression process is to have an overlook on the dataset and ensure that it is suitable for regression or not. Firstly, the variable Car_ID is only used as a unique identifier for each observation, and hence does not contain any information that contributes to predicting the response variable, so it will be dropped. The variables cylinder number and door number have been encoded such that instead of the word “two” it takes on a

numeric form. In addition, it is clear that this data set contains a number of categorical variables that must be dealt with. These variables are fuel type, aspiration, car body, drive wheel, engine location and fuel system. For each categorical variable new columns have been defined and the categories are converted into binary variables (indicator variables). Another categorical variable is Car Name which contains 147 unique categories and hence to help reduce the number of categories, a new feature was extracted called car brand that only includes the car brand instead of the full name of the car and the original column Car Name was subsequently dropped. There were some discrepancies in some of the categories in some of the variables where there were typos and hence they were corrected before creating the indicator variables.

Multiple Linear Regression

Graphs before fitting

Response Variable

The first step to any linear regression model is to plot the data, as Dr Ali Hadi has always preached in all his classes. Hence it was vital to visualize the response variable and understand its distribution and whether it contains outliers or not. Hence a boxplot (Figure 1) was drawn, and it is obvious that the response variable contains a number of outliers which would require further investigation. Furthermore, a histogram (Figure 2) was plotted to examine the distribution of the response variable (price) which appears to be positively skewed (right skewed) with most prices concentrated toward the lower range and a tail extending toward higher prices which would make sense in this case as the Chinese car company is targeting a more affordable segment of the market and hence would require more data for that segment rather than the more expensive car companies. In addition, a statistical

summary was printed for the price, and it indicated that it had a median of 10295 which would in this case represent the ‘typical’ car price given this dataset.

Predictor Variables

Due to the big amount of predictor variables as well as the encoded indicator variables, it would be very time consuming to plot each of the predictor variables against easy other separately, hence, to combat these issues and still get a sense of relationship between the predictor variables, a subset of the variables only containing the numerical variables has been created and the scatter plot matrix has been plotted. From the pairs plot (Figure 3), it is evident that some of the predictors exhibit strong linear relationships with each other. For example, there is a near-perfect correlation between car length and car width, suggesting a strong dependency. Additionally, curb weight and engine size also show a noticeable linear relationship, indicating that larger engines are associated with heavier vehicles.

Model 1

The response variable was regressed (Figure 4) on all predictor variables and the results of the regression can be seen below. The preliminary results show that 88.6% of the variability in price can be explained by the predictor variables, as indicated by the adjusted R-squared value. The F-statistic is 67.08 which suggests that the model is statistically significant overall. Yet, no definitive statistical inference can be made by this model until the assumptions of linear regression are validated. As a next step, these assumptions (e.g., normality, homoscedasticity, multicollinearity, etc.) will be thoroughly checked and addressed.

Checking the assumptions

There are 10 assumptions to regression, 3 assumptions of these cannot be proved which are that the predictor variables are non-random, that they have no errors, and that all observations are equally reliable. The rest of the assumptions will be tested in this section. From the diagnostic (Figures 5 & 6), it is evident that the residuals are not normally distributed, as the Normal Q-Q Plot shows deviations from the expected straight-line pattern for some residuals. This indicates a violation of the normality assumption. Additionally, the index plot of the standardized residuals and the plot of standardized residuals versus fitted values reveal noticeable patterns and clustering, suggesting potential issues with the independence of residuals. Furthermore, the residuals do not exhibit constant variance, indicating a violation of the homoscedasticity assumption. These observations highlight the need for possible transformations to address these violations. Furthermore, it is evident that the implicit assumptions of linear regression are violated. The dataset contains high-leverage and influential points that may significantly impact the regression results and require further investigation. Hence, some steps need to be taken to satisfy the regression assumptions before continuing; these steps are highlighted in the table below.

Correction Steps

Step	Objective
Coding Indicator Variables into Binary Variables	Properly encode categorical variables to make them suitable for regression analysis.
Transformation of Variables	Address violations of normality, linearity, and homoscedasticity to improve model fit.
Multicollinearity Problem	Identify and mitigate issues caused by highly correlated predictors, ensuring stable estimates.
Detection of Outliers	Identify and handle high-leverage and influential points that may distort regression results.

Since the plot of the studentized residuals versus fitted values and the index plot of standardized residuals indicate the presence of a pattern, categorical variables were converted into binary (dummy) variables to prepare them for regression analysis. To address collinearity, one variable from each category was dropped, serving as the base variable. The usual approach to test the assumptions is to plot the residuals vs each predictor variable but since the number of predictors is large, it would be computationally complex to do so and hence I will be plotting the fitted values vs the residuals and if it deemed necessary plot each predictor variable with the residuals.

Model 2

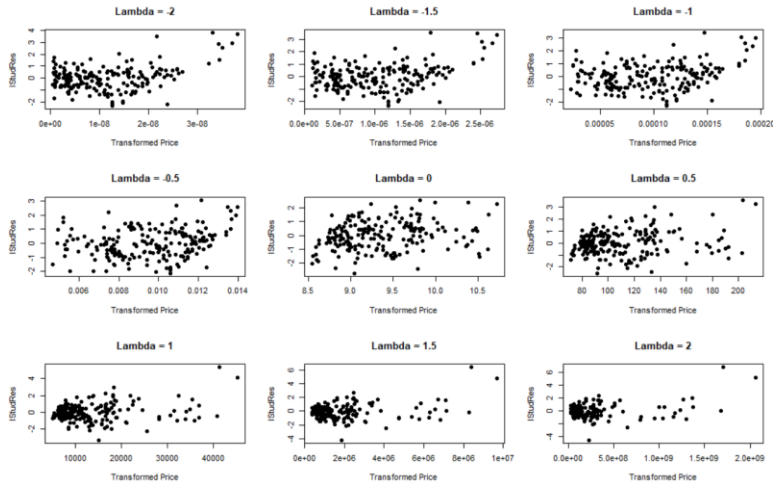
The multiple R-squared has massively improved to read 96.18% (Figure 7) of the variability in price is accounted for by the predictor variables. Some of the predictor variables became more significant. Yet, 2 variables “idi” and “saab” are not defined due to singularities, likely caused by perfect multicollinearity or redundancy in the dataset hence they will be dropped.

Checking assumptions

The index plot of residuals and the plot of studentized residuals vs fitted values (Figure 8) were plotted to validate the assumptions. From these plots, it can be seen that the residual plots have significantly improved showing no clustering. The regression model largely meets the assumptions of normality and independence, though there are minor deviations from normality and potential outliers. Yet the normality assumption has still not been satisfied and requires deeper investigation.

Power Transformation

To address the issue of normality, power transformations will be implemented on the response variable (price) yet to find the most appropriate transformation, we must try different lambda values and the plot of the studentized residuals versus the response variable to find the best lambda. The ladder of transformation is defined to take values between -2.0 to 2.0 in increments of 0.5. As can seen below the best transformation is the log transformation since the best graph among all is that of lambda equals 0. Hence the response variable will be transformed accordingly, and the model will be fitted again.



Model 3

The regression model with the log-transformed price shows a significant improvement, with a Multiple R-squared of 96.33%**,** indicating that 96.33% of the variability in the log-transformed price is accounted for by the predictor variables (Figure 9). Several predictor variables have become more significant after the transformation, suggesting a better fit of the model to the data.

Checking assumptions

The histogram of residuals (Figure 10) demonstrates a roughly symmetric, bell-shaped distribution, indicating that the normality assumption is approximately satisfied. This is further supported by the Q-Q plot, where most residuals align with the line. The index plot of residuals shows a random scatter of residuals across the index, confirming the assumption of independence. Lastly, the residuals vs. fitted values plot illustrates an even spread of residuals around zero, with no discernible patterns or fan-shaped structures, validating the assumption of homoscedasticity (constant variance). Overall, the model's assumptions are largely satisfied, ensuring that the regression results have improved, yet the issue of multicollinearity and existence of outliers still exists. Even though the transformation of the response variable validated all the assumptions, the residuals were plotted vs each predictor variables just to ensure that there were no more needed transformations. All plots did not show any signs of non-linearity or violation of any assumptions, samples of these plots can be seen in the appendix (Figures 15 and 16).

Detection of outliers

The outlier detection plots of Cook’s distance, Hadi’s influence, and the Potential residuals plot (Figure 11, 12, 13) were plotted to detect the outliers in the data, as seen below there exists outliers and hence the implicit assumption is violated. These outliers were identified as observations 3, 126, 127, 17, 50, 135 and 75 yet there does not seem to be an obvious reason to why they are outliers hence I have decided to drop these outliers and check the effect of their dropping on the model. The model fitted after dropping the outliers remained almost the same in terms of its R-squared value, yet the assumptions have all been validated.

Multicollinearity

As seen in all previous models, the issue of multicollinearity stands, which can be observed by examining the high standard errors of the regression coefficients and hence the issue of multicollinearity is required to be dealt with. In order to validate this assumption, we use the the VIF (Variance Inflation Factor) scores and conditional indices. The VIF values indicate a significant presence of multicollinearity among several variables in the model. Notably, the variable diesel exhibits extremely high VIF, suggesting severe collinearity with other predictors. Variables related to vehicle dimensions and engine specifications, such as curbweight, enginesize, compressionratio, carwidth, carlength, wheelbase, and cylindernumber, also show high VIF values, indicating strong linear relationships among them. Additionally, power and efficiency-related variables, including horsepower, citympg, and highwaympg, exhibit moderate to high collinearity, likely due to their functional interdependence. Variables such as ohcf, ohcv, and dohc further contribute to collinearity, likely due to their connection to engine configurations.

```
> vif(reg3)
      symboling          diesel           turbo     convertible         hardtop       hatchback
       4.663355      182.504652        6.481736        2.454976        1.991945        2.753351
          wagon          fourwd             rwd            rear       wheelbase       carlength
       2.037028        2.091494        7.749614        4.839046       18.435476       21.553755
       carwidth       carheight      curbweight            dohc           dohcv               l
      14.008754        6.896263       39.853406        2.705871        3.311208       18.950592
           ohcf            ohcv           rotor   cylindernumber      enginesize          `1bbl`
       4.813564        5.478327        7.871846       28.564728       62.694495        8.100949
         twobbl         fourbbl            spdi       boreratio          stroke compressionratio
       4.935098        4.280925        2.958778        9.582988        4.796297      172.328745
     horsepower         peakrpm         citympg      highwaympg      alfa_romero            audi
      42.583025        5.624704       46.771763       40.019714        2.146180        3.109166
            bmw       chevrolet           dodge           honda           isuzu          jaguar
       2.814931        5.098207        2.028334        2.486570        8.857704        1.530990
     volkswagen           mazda           buick           Other       mitsubishi          nissan
       3.251238        2.398665        3.898501        2.560581        1.580543       23.785402
        peugeot        plymouth         porsche         renault          toyota           volvo
       2.091597        3.502188        1.672824        3.469307        2.339900        3.629905
```

The condition numbers (kappa values) indicate significant multicollinearity in the dataset, particularly among variables such as diesel, wheelbase, carlength, curbweight, enginesize, compressionratio, and horsepower.

```
> kappa
 [1] 1.000000e+00 1.444827e+00 1.773340e+00 1.836570e+00 1.921545e+00 2.023856e+00 2.097864e+00 2.171867e+00
 [9] 2.336536e+00 2.443285e+00 2.499827e+00 2.619789e+00 2.662396e+00 2.716430e+00 2.730463e+00 2.833122e+00
[17] 2.893745e+00 2.922293e+00 2.970888e+00 2.982498e+00 2.998676e+00 3.077996e+00 3.135967e+00 3.270558e+00
[25] 3.307258e+00 3.676385e+00 4.025440e+00 4.198127e+00 4.334911e+00 4.410982e+00 4.918854e+00 5.280216e+00
[33] 5.435301e+00 5.734439e+00 6.555752e+00 6.936714e+00 7.405827e+00 7.597670e+00 8.630479e+00 9.139710e+00
[41] 9.592947e+00 9.670330e+00 1.210512e+01 1.316107e+01 1.383862e+01 1.411367e+01 1.519440e+01 1.926558e+01
[49] 2.296329e+01 2.483324e+01 2.778201e+01 4.816934e+01 6.302408e+01 2.357223e+08
```

This proves the existence of multicollinearity amongst the variables in our data and these variables cause issues with our regression model.

Model 4

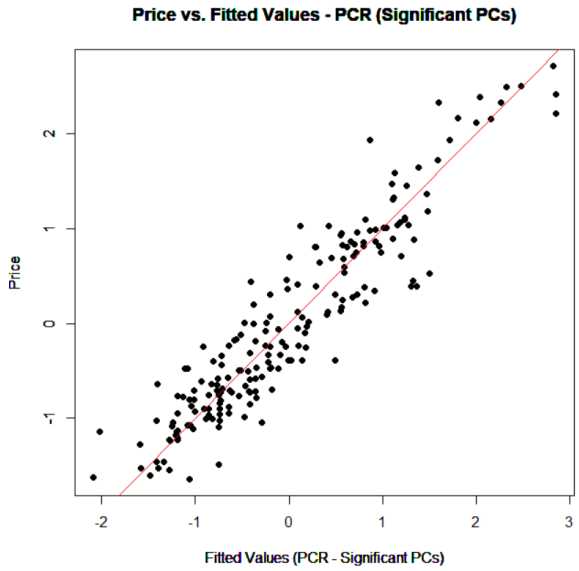
PCA

Hence to deal with this problem, principal components regression was implemented to get rid of collinearity and provide us with a better performing model. In this approach, the predictors were transformed into uncorrelated principal components derived from the eigenvectors of the correlation matrix. The regression model was initially fitted using all principal components, and p-values were used to identify the statistically significant components. Only these significant components were retained in the final model to ensure that irrelevant or redundant predictors were excluded.

```
> print(significant_pcs)
 W1  W2 W24
  1   2  24
```

The coefficients of the significant components were transformed back into the original predictor space to interpret their contributions. Then, the fitted values were calculated based on these coefficients, providing a robust model that effectively mitigates multicollinearity while retaining strong predictive power. Finally, the fitted values were plotted against the price to view the robustness and accuracy of the model and based on the graph the majority of the fitted values align closely with the actual prices, as indicated by their proximity to the red diagonal line in the graph. Overall, the PCR model performs well for this dataset,

providing a robust framework that mitigates the impact of multicollinearity and delivers accurate predictions.



Moreover, the model was refitted only using the significant PCs (Figure 12); to compare the performance of the model before the implementation of PCR and after, the table below was constructed. Even though the R-squared value decreased drastically, but the PCR model explains 88.35% of the variance in the target variable using just three principal components, demonstrating strong performance with fewer predictors. Furthermore, the RSE is very low (0.344), indicating the model has minimal error in predicting the target variable when scaled appropriately.

Metric	PCR Model	Full Model
RSE	0.344	1820.000
Multiple R^2	0.8835	0.9618
Adjusted R^2	0.882	0.9481
Predictors Used	3 PCs	54 predictors

Conclusion

After analyzing the data and addressing various issues affecting the regression model, it is evident that Principal Component Regression (PCR) provides a robust solution to mitigate multicollinearity and enhance the model's performance. Initially, multicollinearity posed significant challenges, as observed through high VIF values and condition numbers, indicating strong dependencies among several predictor variables. To address this, PCR was implemented by transforming the predictors into uncorrelated principal components and fitting the regression model. Insignificant components were identified and excluded based on their p-values, ensuring that only relevant components were retained in the final model. The coefficients of significant components were transformed back into the original predictor space for interpretability, and the fitted values were calculated and visualized. The plot of fitted values against actual prices demonstrated strong alignment, showcasing the model's accuracy and robustness. Despite minor deviations, the PCR model effectively handled multicollinearity and delivered reliable predictions. Overall, this approach underscores the importance of addressing multicollinearity in regression analysis to ensure stability and interpretability, making PCR a valuable technique in this context.

Appendix

Figure 1

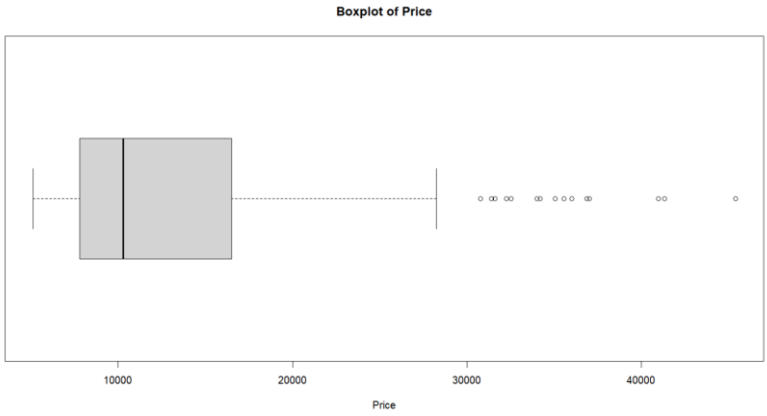


Figure 2

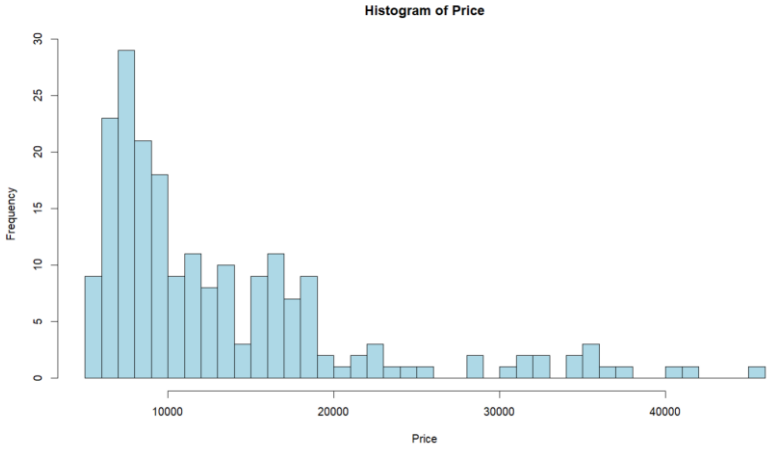


Figure 3

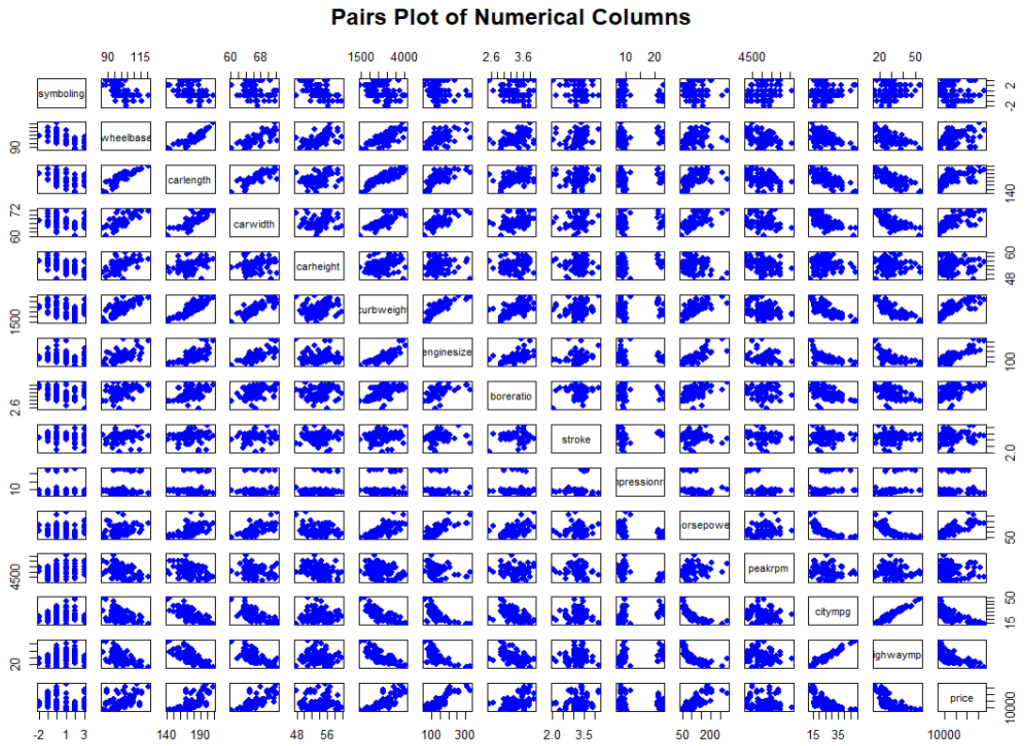


Figure 4

```
Call:
lm(formula = price ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7276.2 -1614.4    58.7  1369.2 14346.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -7.903e+04  2.171e+04  -3.641 0.000355 ***
symboling        -6.202e+01  2.620e+02  -0.237 0.813152
fueltype          4.744e+03  6.596e+03   0.719 0.472929
aspiration        6.433e+02  8.829e+02   0.729 0.467153
doornumber        2.928e+02  3.203e+02   0.914 0.361824
carbody          -6.864e+02  3.726e+02  -1.842 0.067133 .
drivewheel        8.610e+02  5.398e+02   1.595 0.112475
enginelocation    1.093e+04  2.083e+03   5.249 4.28e-07 ***
wheelbase         9.204e+01  1.026e+02   0.897 0.371100
carlength        -2.653e+01  5.373e+01  -0.494 0.622123
carwidth          6.827e+02  2.434e+02   2.805 0.005584 **
carheight         2.550e+02  1.300e+02   1.962 0.051355 .
curbweight        1.922e+00  1.580e+00   1.216 0.225426
enginetype        1.390e+02  2.116e+02   0.657 0.512036
cylindernumber   -9.953e+02  6.538e+02  -1.522 0.129669
enginesize        1.109e+02  2.546e+01   4.356 2.22e-05 ***
fuelsystem        7.039e+01  1.965e+02   0.358 0.720590
boreratio        -2.400e+03  1.504e+03  -1.595 0.112473
stroke           -3.450e+03  8.802e+02  -3.920 0.000126 ***
compressionratio  4.628e+02  4.702e+02   0.984 0.326333
horsepower        3.583e+01  1.805e+01   1.985 0.048620 *
peakrpm           1.044e+00  6.599e-01   1.581 0.115533
citympg          -6.136e+01  1.658e+02  -0.370 0.711761
highwaympg        7.374e+01  1.459e+02   0.505 0.613967
car_brand        -2.186e+02  3.836e+01  -5.699 4.84e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2697 on 180 degrees of freedom
Multiple R-squared:  0.8994,    Adjusted R-squared:  0.886
F-statistic: 67.08 on 24 and 180 DF,  p-value: < 2.2e-16
```

Figure 5

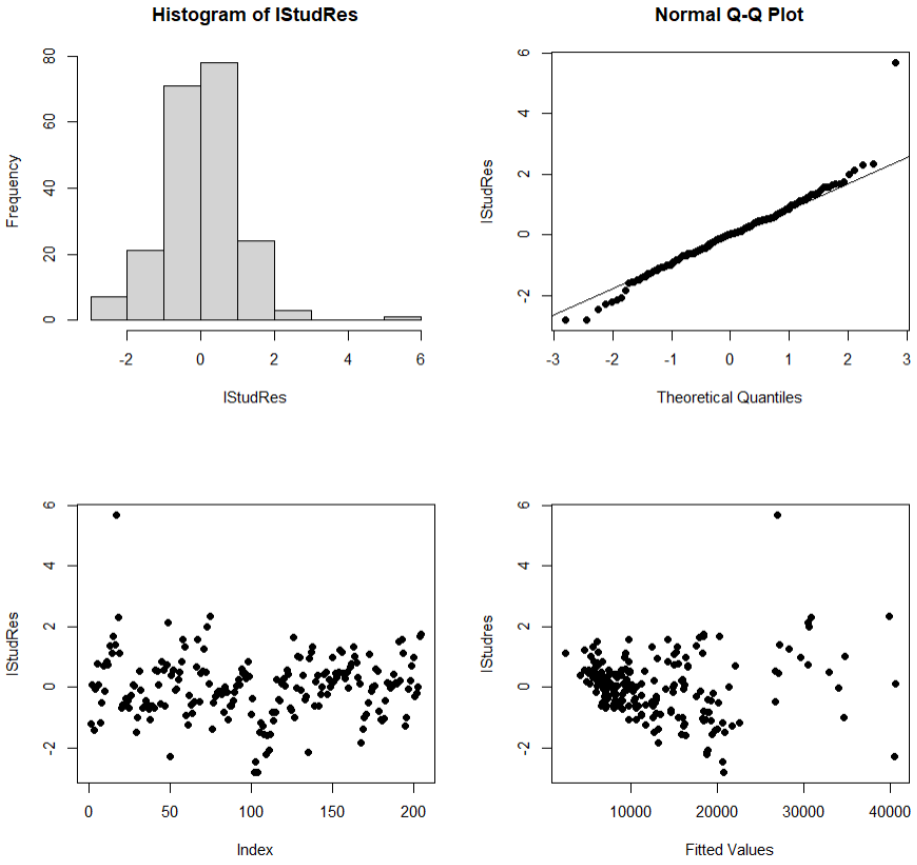


Figure 6

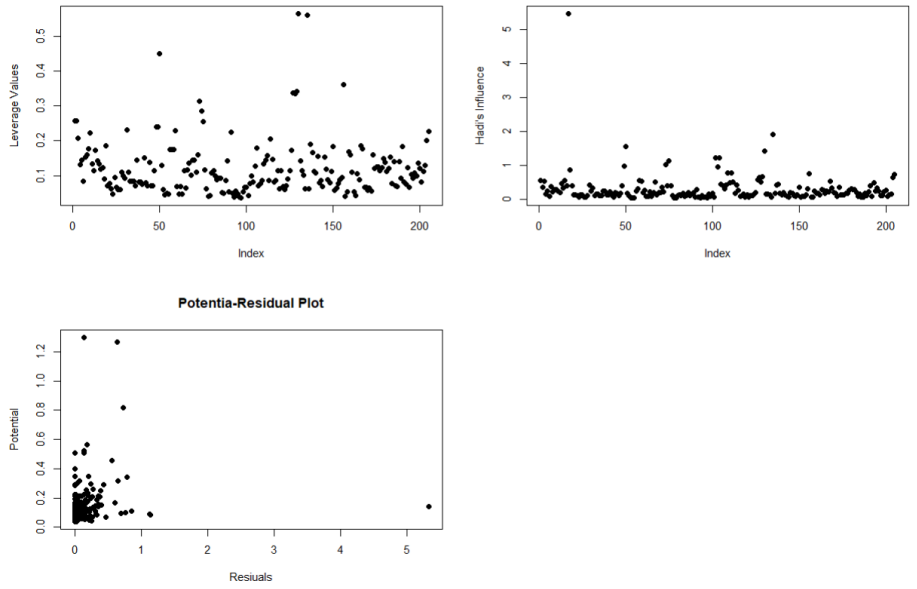


Figure 7

```
Call:
lm(formula = df_new$price ~ ., data = df_new)

Residuals:
     Min       1Q   Median       3Q      Max
-3615.7  -930.6    -2.3   875.1  8568.8

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.116e+04  1.606e+04  -2.563 0.011349 *
symboling        -1.013e+02  2.210e+02  -0.459 0.647137
diesel            8.908e+03  5.787e+03   1.539 0.125828
turbo             2.141e+03  8.414e+02   2.544 0.011961 *
convertible       2.564e+03  1.182e+03   2.170 0.031558 *
hardtop           7.121e+01  9.264e+02   0.077 0.938827
hatchback        -8.383e+02  4.448e+02  -1.885 0.061387 .
wagon            -2.069e+02  5.544e+02  -0.373 0.709475
fourwd            4.298e+02  8.972e+02   0.479 0.632637
rwd               4.466e+02  7.326e+02   0.610 0.543049
rear              1.235e+04  2.328e+03   5.303 4.01e-07 ***
wheelbase         2.453e+02  9.085e+01   2.699 0.007742 **
carlength        -1.178e+02  4.795e+01  -2.458 0.015116 *
carwidth          6.709e+02  2.223e+02   3.018 0.002994 **
carheight        -2.508e+02  1.369e+02  -1.831 0.069038 .
curbweight        4.876e+00  1.545e+00   3.156 0.001934 **
dohc             -3.736e+02  8.907e+02  -0.419 0.675485
dohcv             2.510e+03  3.320e+03   0.756 0.450724
l                 4.904e+03  2.357e+03   2.081 0.039164 *
ohcf              5.136e+02  1.071e+03   0.480 0.632217
ohcv             -1.722e+03  1.221e+03  -1.410 0.160527
rotor             7.637e+03  2.578e+03   2.962 0.003553 **
cylindernumber   -3.451e+02  6.301e+02  -0.548 0.584667
enginesize        1.195e+02  2.423e+01   4.932 2.14e-06 ***
`1bbl`           -1.958e+03  1.605e+03  -1.219 0.224638
twobbl            8.553e+02  6.044e+02   1.415 0.159060
fourbbl          -2.263e+03  2.190e+03  -1.033 0.303087
spdi             -9.670e+01  1.015e+03  -0.095 0.924226
boreratio        -3.372e+03  1.456e+03  -2.316 0.021928 *
stroke           -1.403e+03  8.898e+02  -1.577 0.116929
compressionratio -7.537e+02  4.211e+02  -1.790 0.075520 .
horsepower       -1.528e+01  2.103e+01  -0.727 0.468609
peakrpm           2.427e+00  6.335e-01   3.831 0.000187 ***
citympg           1.550e+01  1.332e+02   0.116 0.907493
highwaympg        1.226e+02  1.171e+02   1.047 0.296603
alfa_romero       3.096e+03  1.551e+03   1.997 0.047683 *
audi              2.152e+03  1.234e+03   1.744 0.083224 .
bmw               9.175e+03  1.101e+03   8.332 4.66e-14 ***
chevrolet         5.282e+03  1.482e+03   3.564 0.000490 ***
dodge            -1.723e+03  1.507e+03  -1.143 0.254915
honda            -2.167e+03  9.783e+02  -2.215 0.028245 *
isuzu             1.164e+03  1.552e+03   0.750 0.454462
jaguar            3.484e+02  1.137e+03   0.306 0.759750
volkswagen        1.252e+03  1.909e+03   0.656 0.512858
mazda             1.029e+03  7.339e+02   1.403 0.162757
buick            -2.641e+03  1.030e+03  -2.564 0.011322 *
Other            -6.001e+01  7.187e+02  -0.084 0.933564
mitsubishi        8.149e+02  1.155e+03   0.705 0.481702
nissan           -6.786e+03  2.751e+03  -2.467 0.014751 *
peugeot          -2.215e+03  1.012e+03  -2.188 0.030194 *
plymouth          5.310e+03  1.720e+03   3.088 0.002402 **
porsche          -6.049e+02  1.673e+03  -0.362 0.718102
renault           4.858e+03  1.405e+03   3.459 0.000706 ***
toyota            8.932e+02  8.282e+02   1.078 0.282585
volvo             1.515e+03  1.075e+03   1.410 0.160565
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1820 on 150 degrees of freedom
Multiple R-squared:  0.9618,    Adjusted R-squared:  0.9481
F-statistic: 70.02 on 54 and 150 DF,  p-value: < 2.2e-16
```

Figure 8

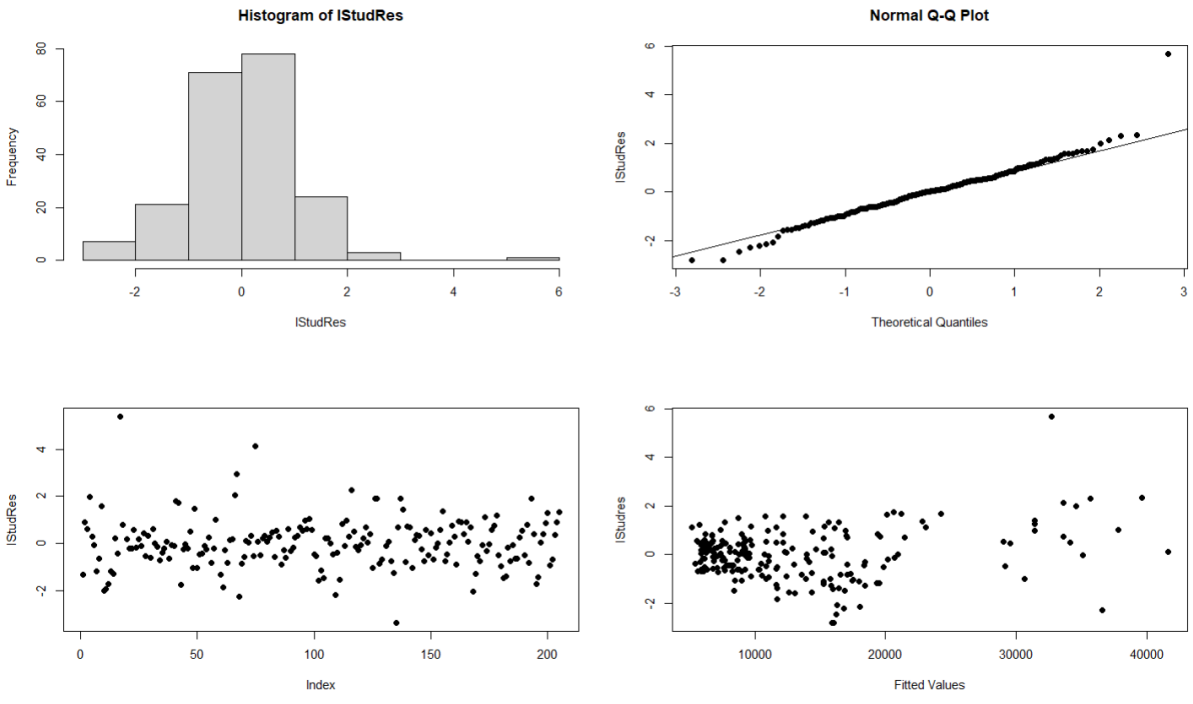


Figure 9

```
Call:
lm(formula = df_transformed$price ~ ., data = df_transformed)

Residuals:
      Min        1Q    Median        3Q       Max
-0.278485 -0.067990  0.002186  0.059653  0.239644

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.819e+00  9.931e-01   6.866 1.64e-10 ***
symboling        -2.403e-03  1.367e-02  -0.176 0.860664
diesel           -3.576e-03  3.579e-01  -0.010 0.992042
turbo             1.235e-01  5.204e-02   2.373 0.018893 *
convertible       1.100e-01  7.308e-02   1.505 0.134346
hardtop          -3.891e-02  5.730e-02  -0.679 0.498080
hatchback        -7.065e-02  2.751e-02  -2.568 0.011199 *
wagon            -2.314e-03  3.429e-02  -0.067 0.946288
fourwd            4.813e-02  5.549e-02   0.867 0.387168
rwd               1.147e-02  4.531e-02   0.253 0.800511
rear              7.085e-01  1.440e-01   4.920 2.25e-06 ***
wheelbase         1.558e-02  5.619e-03   2.772 0.006278 **
carlength        -5.923e-03  2.966e-03  -1.997 0.047594 *
carwidth          3.388e-02  1.375e-02   2.464 0.014867 *
carheight        -2.688e-02  8.470e-03  -3.174 0.001823 **
curbweight        4.617e-04  9.555e-05   4.832 3.31e-06 ***
dohc              2.238e-02  5.509e-02   0.406 0.685177
dohcv            -2.289e-02  2.053e-01  -0.112 0.911365
l                 2.235e-01  1.458e-01   1.533 0.127378
ohcf             -2.550e-02  6.623e-02  -0.385 0.700802
ohcv             -5.453e-02  7.550e-02  -0.722 0.471316
rotor             1.875e-01  1.595e-01   1.176 0.241435
cylindernumber   -2.858e-02  3.897e-02  -0.733 0.464447
enginesize        3.226e-03  1.498e-03   2.153 0.032946 *
`ibbl`           -2.166e-01  9.930e-02  -2.182 0.030685 *
twobbl           -7.062e-02  3.738e-02  -1.889 0.060787 .
fourbbl          -1.571e-01  1.355e-01  -1.160 0.247853
spdi             -2.034e-02  6.278e-02  -0.324 0.746408
boreratio        -1.050e-01  9.008e-02  -1.165 0.245693
stroke           -4.713e-02  5.504e-02  -0.856 0.393218
compressionratio -1.705e-03  2.605e-02  -0.065 0.947881
horsepower       -5.935e-05  1.301e-03  -0.046 0.963658
peakrpm           5.335e-05  3.919e-05   1.362 0.175391
citympg          -1.542e-02  8.238e-03  -1.871 0.063256 .
highwaympg        1.241e-02  7.240e-03   1.714 0.088538 .
alfa_romero       2.089e-01  9.591e-02   2.178 0.030947 *
audi              1.908e-01  7.633e-02   2.500 0.013496 *
bmw               5.122e-01  6.811e-02   7.520 4.66e-12 ***
chevrolet         1.670e-01  9.166e-02   1.822 0.070511 .
dodge            -3.928e-02  9.324e-02  -0.421 0.674128
honda            -1.277e-01  6.051e-02  -2.111 0.036420 *
isuzu             1.734e-01  9.601e-02   1.806 0.072865 .
jaguar            1.015e-01  7.033e-02   1.444 0.150911
volkswagen       -1.280e-01  1.180e-01  -1.085 0.279848
mazda             1.550e-01  4.539e-02   3.414 0.000823 ***
buick            -1.575e-01  6.369e-02  -2.473 0.014534 *
Other             5.462e-02  4.445e-02   1.229 0.221043
mitsubishi        3.514e-02  7.146e-02   0.492 0.623590
nissan           -3.347e-01  1.701e-01  -1.967 0.051016 .
peugeot          -1.200e-01  6.261e-02  -1.917 0.057138 .
plymouth          2.582e-01  1.064e-01   2.427 0.016401 *
porsche          -9.090e-02  1.034e-01  -0.879 0.380980
renault           2.664e-01  8.687e-02   3.067 0.002567 **
toyota            3.933e-02  5.123e-02   0.768 0.443886
volvo             1.159e-01  6.647e-02   1.744 0.083226 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1126 on 150 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9501
F-statistic:  72.9 on 54 and 150 DF,  p-value: < 2.2e-16
```

Figure 10

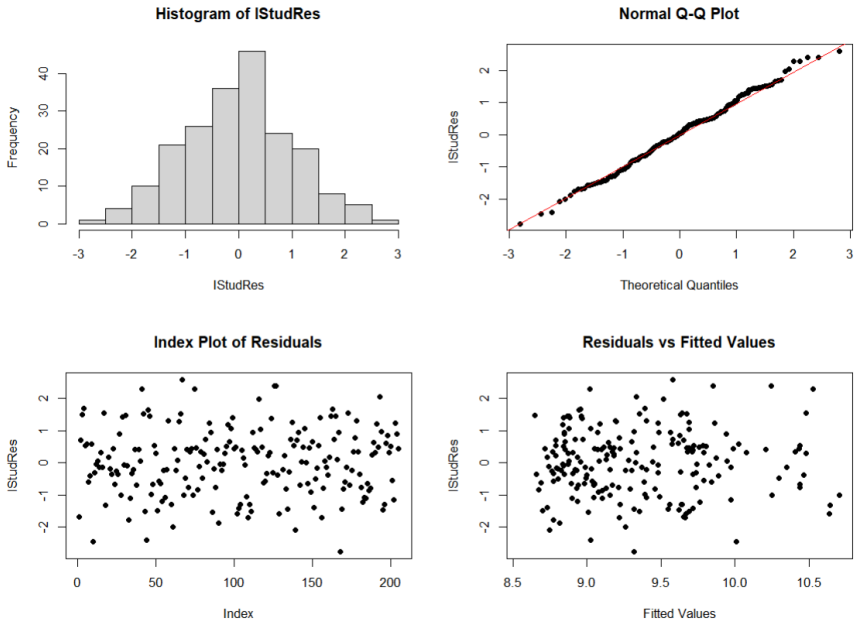


Figure 11

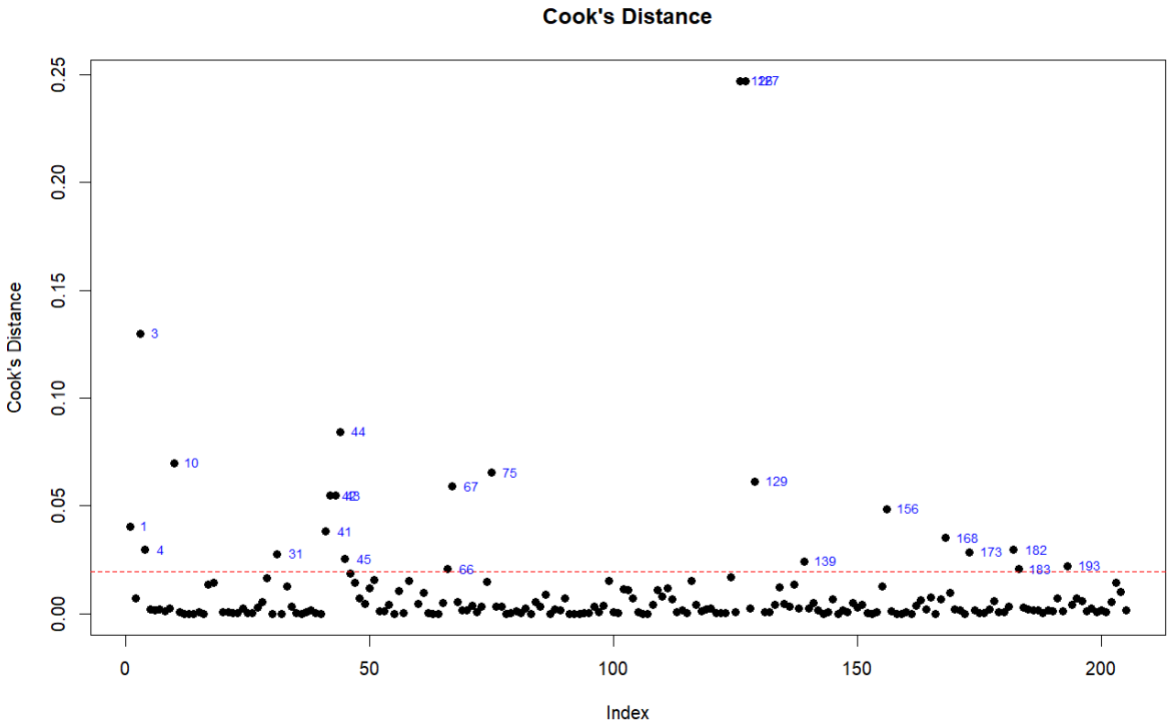


Figure 12

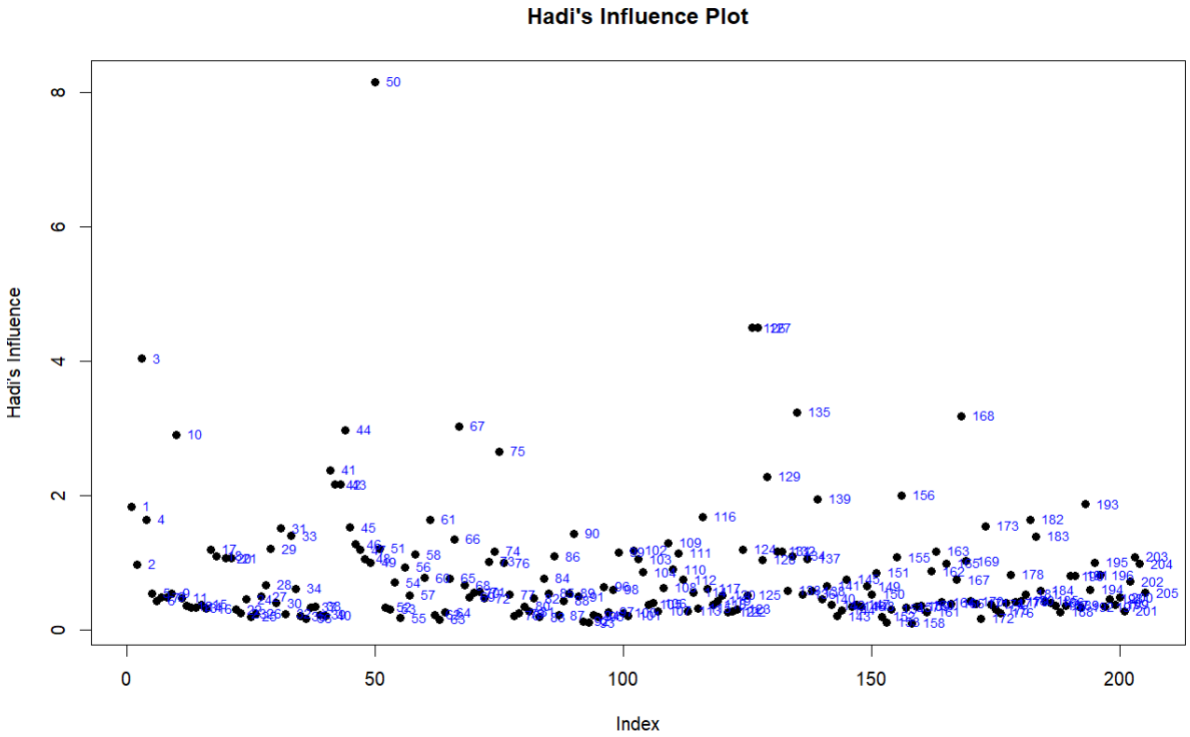


Figure 13

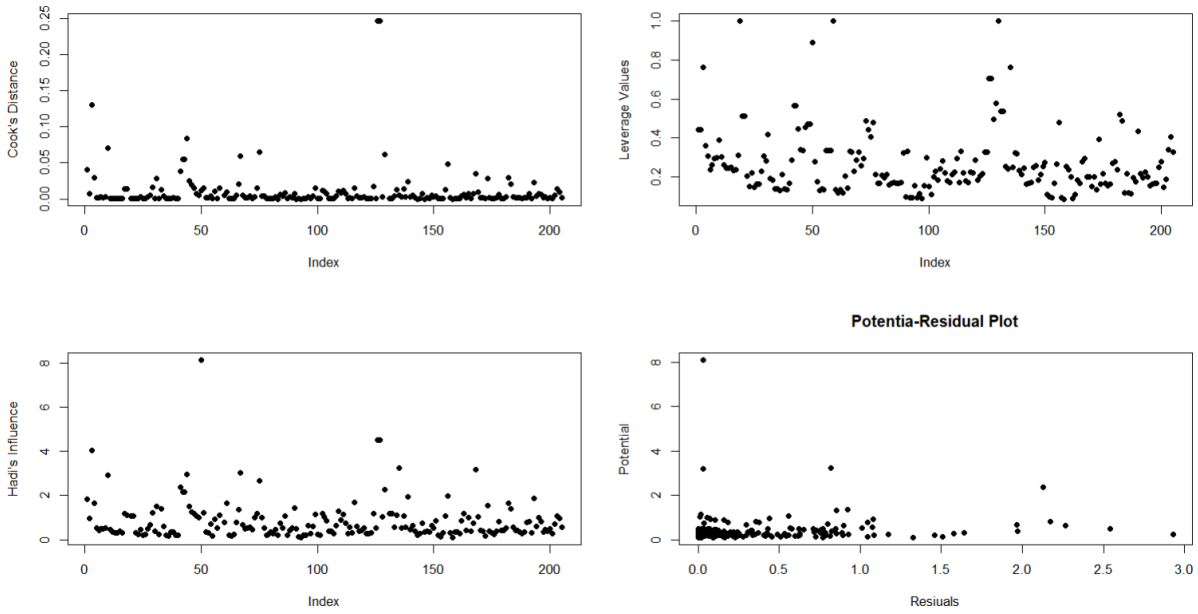


Figure 14

```
Call:
lm(formula = df_new2$price ~ W_significant)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81055 -0.21332 -0.04496  0.25211  1.00347

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.567e-15  2.444e-02   0.000        1
W_significant1 -2.975e-01  8.034e-03 -37.030  < 2e-16 ***
W_significant2  9.584e-02  1.161e-02   8.257 2.28e-14 ***
W_significant3 -1.493e-01  2.628e-02  -5.683 4.80e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3439 on 194 degrees of freedom
Multiple R-squared:  0.8835,    Adjusted R-squared:  0.8817
F-statistic: 490.6 on 3 and 194 DF,  p-value: < 2.2e-16
```

Figure 15

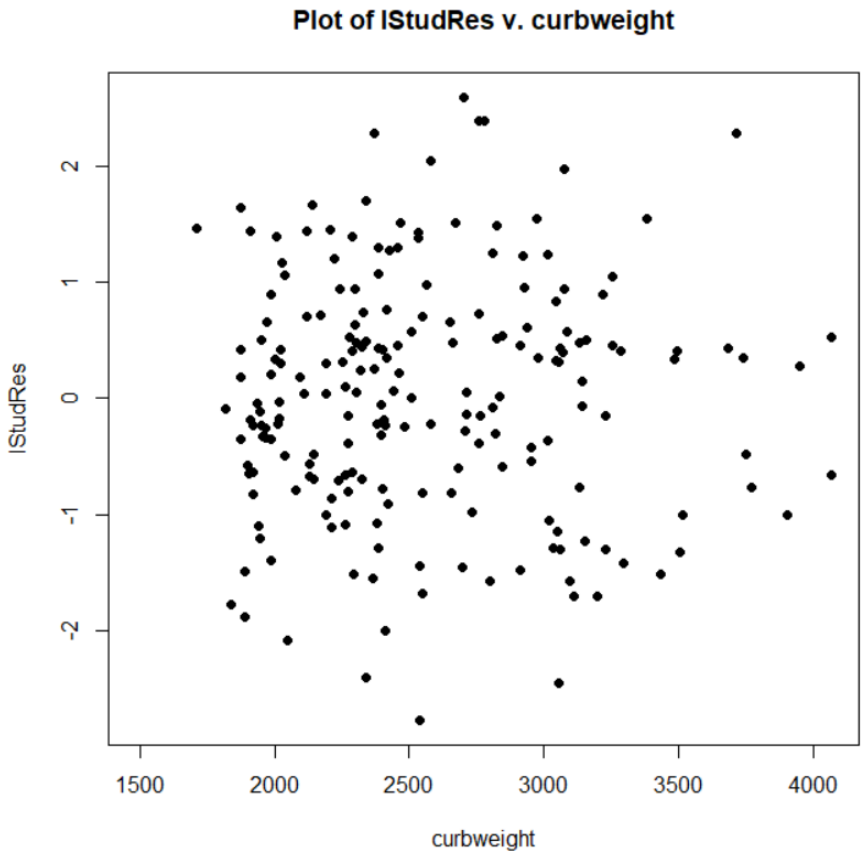
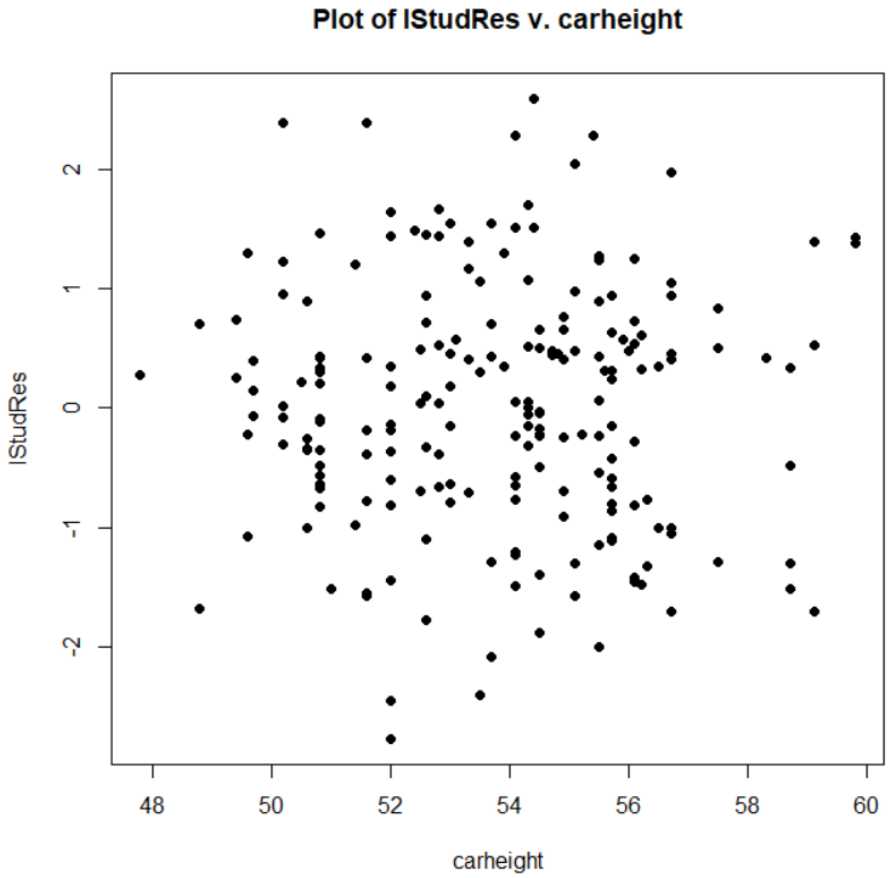

Plot of lStudRes v. curbweight
lStudRes
2
1
0
-1
-2
1500
2000
2500
3000
3500
4000
curbweight

Figure 16


Plot of lStudRes v. carheight
lStudRes
2
1
0
-1
-2
48
50
52
54
56
58
60
carheight