

Prediction of car price

**Abstract**

Using panel data for the car type in the US market, this paper empirically examines the variation of car prices under the effect of many other variables. Unlike most previous studies, sixteen different variables are used to construct a comprehensive development model that captures the variation in car prices. The results indicate that car prices are heavily impacted by engine size, horsepower, peak Rpm, fuel type, car width, curb Weight, stroke, city mpg, and finally highway Mpg.

**Introduction**

A **car** (or **automobile**) is a wheeled [motor vehicle](https://en.wikipedia.org/wiki/Motor_vehicle) that is used for [transportation](https://en.wikipedia.org/wiki/Transportation). Most definitions of cars say that they run primarily on [roads](https://en.wikipedia.org/wiki/Road), [seat](https://en.wikipedia.org/wiki/Car_seat) one to eight people, have four [wheels](https://en.wikipedia.org/wiki/Wheels), and mainly transport [people](https://en.wikipedia.org/wiki/Private_transport#Personal_transport) instead of [goods](https://en.wikipedia.org/wiki/Cargo).

Cars have controls for [driving](https://en.wikipedia.org/wiki/Driving), [parking](https://en.wikipedia.org/wiki/Parking), [passenger](https://en.wikipedia.org/wiki/Passenger) comfort, and a variety of lights. Over the decades, additional features and controls have been added to vehicles, making them progressively more complex. These include rear-reversing cameras, [air conditioning](https://en.wikipedia.org/wiki/Automobile_air_conditioning), [navigation systems](https://en.wikipedia.org/wiki/Automotive_navigation_system), and [in-car entertainment](https://en.wikipedia.org/wiki/In-car_entertainment). Most cars in use in the early 2020s are propelled by an [internal combustion engine](https://en.wikipedia.org/wiki/Internal_combustion_engine), fueled by the [combustion](https://en.wikipedia.org/wiki/Combustion) of [fossil fuels](https://en.wikipedia.org/wiki/Fossil_fuel). [Electric cars](https://en.wikipedia.org/wiki/Electric_car), which were invented early in the [history of the car](https://en.wikipedia.org/wiki/History_of_the_automobile), became commercially available in the 2000s and are predicted to cost less to buy than gasoline cars before 2025. The transition from fossil fuels to electric cars features prominently in most [climate change mitigation scenarios](https://en.wikipedia.org/wiki/Climate_change_mitigation_scenarios), such as [Project Drawdown](https://en.wikipedia.org/wiki/Project_Drawdown)'s 100 actionable solutions for climate change.

There are costs and benefits to car use. The costs to the individual include acquiring the vehicle, interest payments (if the car is financed), repairs and [maintenance](https://en.wikipedia.org/wiki/Auto_maintenance), fuel, [depreciation](https://en.wikipedia.org/wiki/Depreciation), driving time, parking fees, taxes, and [insurance](https://en.wikipedia.org/wiki/Vehicle_insurance). The costs to society include maintaining roads, [land use](https://en.wikipedia.org/wiki/Land_use), [road congestion](https://en.wikipedia.org/wiki/Road_congestion), [air pollution](https://en.wikipedia.org/wiki/Air_pollution), [public health](https://en.wikipedia.org/wiki/Public_health), [healthcare](https://en.wikipedia.org/wiki/Health_care), and [disposing of the vehicle at the end of its life](https://en.wikipedia.org/wiki/Vehicle_recycling). [Traffic collisions](https://en.wikipedia.org/wiki/Traffic_collisions) are the largest cause of injury-related deaths worldwide.

Personal benefits include on-demand transportation, mobility, independence, and convenience. Societal benefits include economic benefits, such as job and wealth creation from the [automotive industry](https://en.wikipedia.org/wiki/Automotive_industry), transportation provision, societal well-being from leisure and travel opportunities, and revenue generation from [taxes](https://en.wikipedia.org/wiki/Category:Vehicle_taxes). People's ability to move flexibly from place to place has far-reaching implications for the nature of societies. There are around one billion cars in use worldwide. Car usage is increasing rapidly, especially in [China](https://en.wikipedia.org/wiki/Automotive_industry_in_China), [India](https://en.wikipedia.org/wiki/Automotive_industry_in_India), and other [newly industrialized countries](https://en.wikipedia.org/wiki/Newly_industrialized_country).

The global automobile industry is a multi-trillion-dollar enterprise. There are millions of vehicles sold worldwide, roughly 74 to 78 million new vehicles.

There are many factors behind this high number of cars sold globally. There’s more disposable income among consumers and changing driving habits due to urbanization. Also, there’s an increasing tendency for consumers to trade vehicles every two or three years for a newer model.

Although hybrid and electric cars only account for 3.4% of new car sales, the car market is growing. The reason for this growth is due to consumer demand. Most people buy these cars because they want to help the environment. Also, it’s not because it’s mandated by their employer or government.

Many brands are introducing new models with longer battery life and better performance every few months. Industry experts expect to see further growth in this segment over the next few years if technological advancements continue to be made and gas prices remain low.

Many car buyers prefer buying new cars from dealerships or car showrooms. It’s because they can test drive the vehicle and get a chance to inspect it before buying. Another reason is that car dealerships offer various services and spare parts for their cars.

Also, dealers find and deliver the car to their customers. This way, car buyers don’t have to travel far distances to get their desired car. Thus, customers can enjoy a hassle-free experience when buying from a dealership.

People looking to buy a new car will need to accept the higher average cost. In the US, the average price for a new car is around $46,000, which is a record high. Analysts in the automotive industry state that this is an 8% increase based on the previous year.

**Objective**

A relevant question is what determines the price of used cars in the American Market. Various factors such as engine size, horsepower, peak Rpm, and fuel type determine this. Based on these various factors, our objective is to determine the price of used cars in the USA.

Ten determinants of the car’s specs are commonly used, and so will be used in this study: fuel type, aspiration, Wheelbase, car length, car width, car height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg.

Data Description:

This dataset is from Kaggle:

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

Variables conceptualization:

* Fuel type: energy sources
* Aspiration: an engine with the mechanics of internal combustion.
* Wheelbase: distance between the midpoints of the front of your vehicle and rear wheels
* car length: the measurement of a vehicle from the front tip to the rear
* car width: how wide the car is, excluding the mirrors.
* car height: it’s the measurement from a car’s lowest point to the best
* curb weight: vehicle's weight with all the standard equipment and amenities
* engine size: the volume of fuel and air that can be pushed through a car's cylinders
* bore ratio: the ratio of bore to stroke
* stroke: Each movement of the piston
* compression ratio:  the ratio between the volume of the cylinder and combustion chamber in an internal combustion engine at their maximum and minimum
* horsepower: the power an engine produces.
* peak rpm: revolutions per minute
* city mpg: Driving with occasional stopping and braking, simulating the conditions you're likely to run into while driving on city streets.
* highway mpg:more continuous acceleration
* Price: price of the cars

NB: MPG refers to how many miles your car can travel on a single gallon of gas.

Inspiration:

The model will be a good way for management to understand the pricing dynamics of a new market.

<https://housegrail.com/new-car-statistics/#:~:text=Between%2074%20and%2078%20million%20new%20cars%20are%20sold%20every%20year%20worldwide.&text=The%20global%20automobile%20industry%20is,number%20of%20cars%20sold%20globally>.

<https://en.wikipedia.org/wiki/Car#:~:text=The%20year%201886%20is%20regarded,by%20the%20Ford%20Motor%20Company>

Study Methodology

Study period and data collection:

This study attempts to clarify the impact of many variables on car’s prices in the USA. The data source is Kaggle Dataset. In this examination, cross sectional data was used as a method to assess the relationship between variables.

Study design:

1. Descriptive statistics: They provide simple summaries about the sample and the measures. They are the appropriate tests to address the study questions.
2. The first six rows in the data:

*fueltype aspiration wheelbase carlength carwidth carheight curbweight enginesize*

*1 gas std 88.6 168.8 64.1 48.8 2548 130*

*2 gas std 88.6 168.8 64.1 48.8 2548 130*

*3 gas std 94.5 171.2 65.5 52.4 2823 152*

*4 gas std 99.8 176.6 66.2 54.3 2337 109*

*5 gas std 99.4 176.6 66.4 54.3 2824 136*

*6 gas std 99.8 177.3 66.3 53.1 2507 136*

*boreratio stroke compressionratio horsepower peakrpm citympg highwaympg price*

*1 3.47 2.68 9.0 111 5000 21 27 13495*

*2 3.47 2.68 9.0 111 5000 21 27 16500*

*3 2.68 3.47 9.0 154 5000 19 26 16500*

*4 3.19 3.40 10.0 102 5500 24 30 13950*

*5 3.19 3.40 8.0 115 5500 18 22 17450*

*6 3.19 3.40 8.5 110 5500 19 25 15250*

1. The dimension of the data is: 205 rows and 16 columns.
2. The columns names are:

*"fueltype" "aspiration" "wheelbase" "carlength"*

*"carwidth" "carheight" "curbweight" "enginesize"*

*"boreratio" "stroke" "compressionratio" "horsepower"*

*"peakrpm" "citympg" "highwaympg" "price"*

1. The summary of data:

*fueltype aspiration wheelbase carlength carwidth carheight*

*diesel: 20 std :168 Min. : 86.60 Min. :141.1 Min. :60.30 Min. :47.80*

*gas :185 turbo: 37 1st Qu.: 94.50 1st Qu.:166.3 1st Qu.:64.10 1st Qu.:52.00*

*Median : 97.00 Median :173.2 Median :65.50 Median :54.10*

*Mean : 98.76 Mean :174.0 Mean :65.91 Mean :53.72*

*3rd Qu.:102.40 3rd Qu.:183.1 3rd Qu.:66.90 3rd Qu.:55.50*

*Max. :120.90 Max. :208.1 Max. :72.30 Max. :59.80*

*curbweight enginesize boreratio stroke compressionratio*

*Min. :1488 Min. : 61.0 Min. :2.54 Min. :2.070 Min. : 7.00*

*1st Qu.:2145 1st Qu.: 97.0 1st Qu.:3.15 1st Qu.:3.110 1st Qu.: 8.60*

*Median :2414 Median :120.0 Median :3.31 Median :3.290 Median : 9.00*

*Mean :2556 Mean :126.9 Mean :3.33 Mean :3.255 Mean :10.14*

*3rd Qu.:2935 3rd Qu.:141.0 3rd Qu.:3.58 3rd Qu.:3.410 3rd Qu.: 9.40*

*Max. :4066 Max. :326.0 Max. :3.94 Max. :4.170 Max. :23.00*

*horsepower peakrpm citympg highwaympg price*

*Min. : 48.0 Min. :4150 Min. :13.00 Min. :16.00 Min. : 5118*

*1st Qu.: 70.0 1st Qu.:4800 1st Qu.:19.00 1st Qu.:25.00 1st Qu.: 7788*

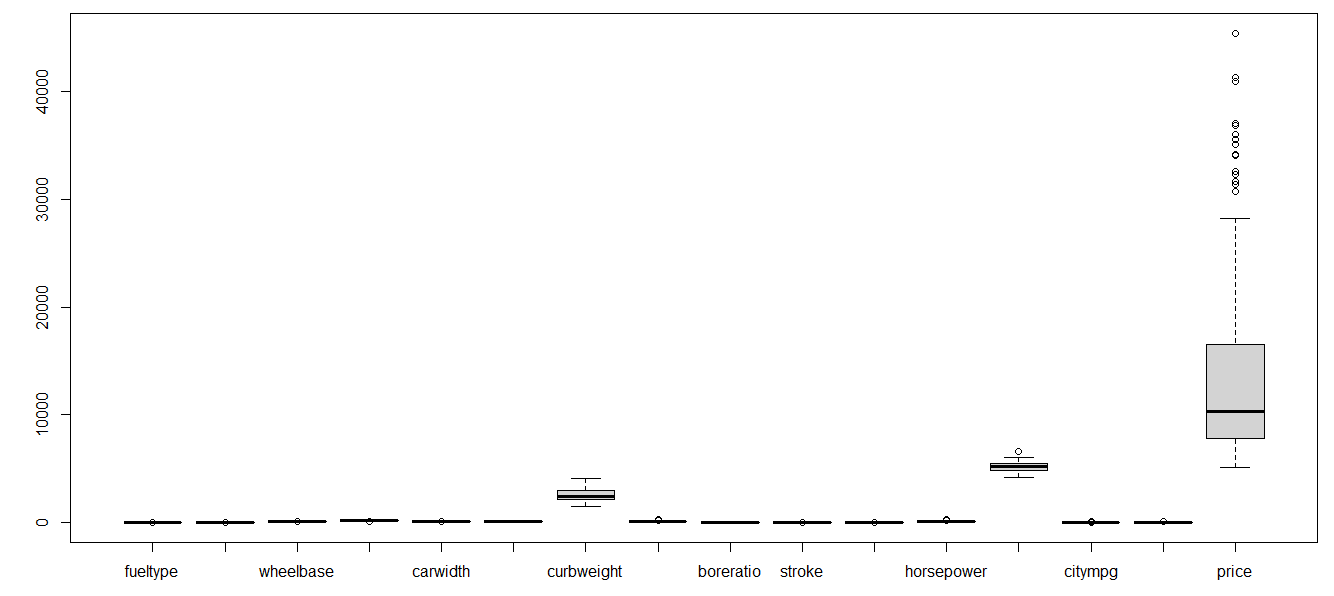
*Median : 95.0 Median :5200 Median :24.00 Median :30.00 Median :10295*

*Mean :104.1 Mean :5125 Mean :25.22 Mean :30.75 Mean :13277*

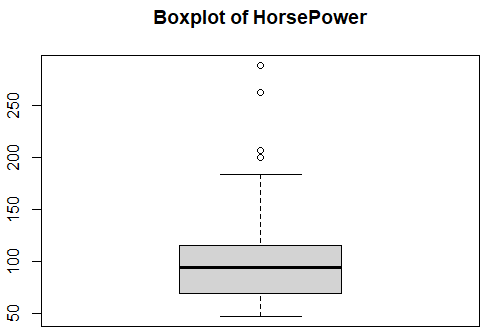
*3rd Qu.:116.0 3rd Qu.:5500 3rd Qu.:30.00 3rd Qu.:34.00 3rd Qu.:16503*

*Max. :288.0 Max. :6600 Max. :49.00 Max. :54.00 Max. :45400*

1. The standard deviation of the dependent variable (salary) is 7988.852
2. the mean of the price is 13277, we divide the sd(salary) by the mean(salary) = 7988.852/13277=0.6017193 so the coefficient of variation is 60% > 30% and our distribution is heterogeneous, and we have high dispersion around the mean.
3. Boxplot of the data:



1. Boxplot of some variables



Chart, box and whisker chart

Description automatically generated

The boxplot of horsepower and peak Rpm is normally distributed with outliers.

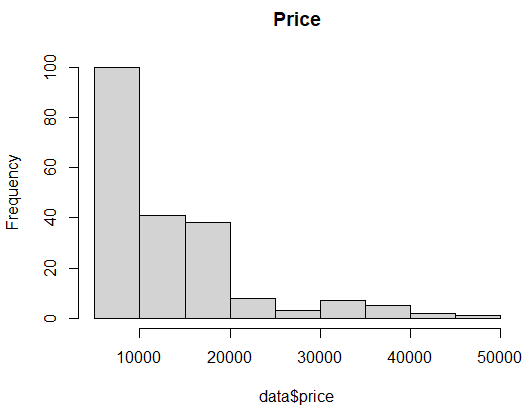
1. Chart, box and whisker chart

   Description automatically generatedChart, box and whisker chart

   Description automatically generatedDistribution of the independent variable: Price

Chart, histogram

Description automatically generatedAfter the transformation between the price and the log of the price we can see that the distribution of the data seems to be normal.

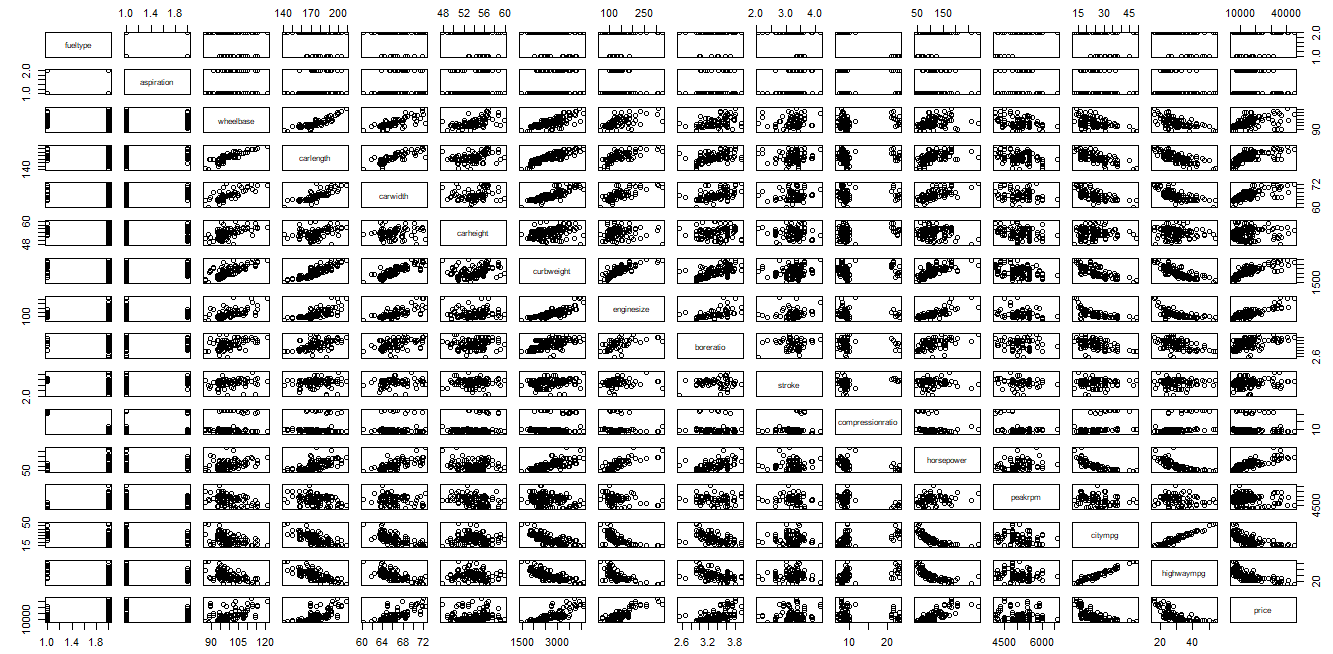
The histogram of the same transformation, we can assure what we saw in the boxplot for that we will use the log (Price).

1. Correlation analysis

In terms of market research this means that, correlation analysis is used to analyze quantitative data gathered from research methods such as surveys and polls, to identify whether there is any significant connections, patterns, or trends between the two.

Essentially, correlation analysis is used for spotting patterns within datasets. A positive correlation result means that both variables increase in relation to each other, while a negative correlation means that as one variable decreases, the other increases.

* 1. Matrix of scatterplots



1. Multiple linear regression (MLR):

MLR is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know, how strong the relationship is between two or more independent variables and one dependent variable and the value of the dependent variable at a certain value of the independent variables. The assumptions are:

1. **Homogeneity of variance (homoscedasticity)**: the size of the error in our prediction doesn’t change significantly across the values of the independent variable.
2. **Independence of observations**: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables. In multiple linear regression, it is possible that some of the independent variables are correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.
3. **Normality**: The data follows a [normal distribution](https://www.scribbr.com/statistics/normal-distribution/).
4. **Linearity**: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor
5. The Akaike information criterion: (AIC)

[Estimator](https://en.wikipedia.org/wiki/Estimator) of prediction error and thereby relative quality of [statistical models](https://en.wikipedia.org/wiki/Statistical_model) for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for [model selection](https://en.wikipedia.org/wiki/Model_selection).

**Study model**

Based on testing six models, a relationship exists between many factors affecting the price of used cars. Thus, we use the following models to describe the relationship between the variables, model A and model B. The log transformation for the independent variable Y (Price) exists only in model B. And after checking the diagnostics, we will choose the best-fitted model.

1. Model A – before the log transformation of Y:

modelA <- lm(price ~ fueltype +carwidth+ carheight+ enginesize+ stroke+

+horsepower +peakrpm, data = data)

AIC = 3898.923

* 1. Summary of the model

*Call:*

*lm(formula = price ~ fueltype + carwidth + carheight + enginesize +*

*stroke + +horsepower + peakrpm, data = data)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-10617.0 -1775.4 -86.2 1519.8 13657.3*

*Coefficients:*

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

*(Intercept) -5.821e+04 1.144e+04 -5.087 8.45e-07 \*\*\**

*fueltypegas -3.861e+03 9.332e+02 -4.137 5.21e-05 \*\*\**

*carwidth 6.354e+02 1.757e+02 3.617 0.000379 \*\*\**

*carheight 1.800e+02 1.059e+02 1.699 0.090936 .*

*enginesize 1.206e+02 1.238e+01 9.736 < 2e-16 \*\*\**

*stroke -2.989e+03 7.677e+02 -3.893 0.000135 \*\*\**

*horsepower 4.211e+01 1.264e+01 3.331 0.001035 \*\**

*peakrpm 2.628e+00 6.416e-01 4.096 6.14e-05 \*\*\**

*---*

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

*Residual standard error: 3186 on 197 degrees of freedom*

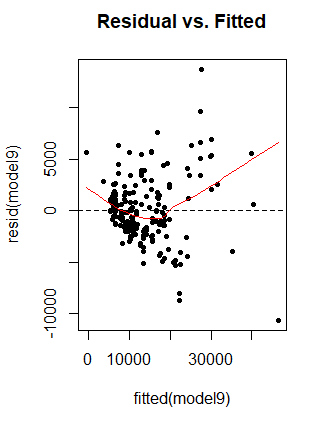
*Multiple R-squared: 0.8464, Adjusted R-squared: 0.8409*

*F-statistic: 155.1 on 7 and 197 DF, p-value: < 2.2e-16*

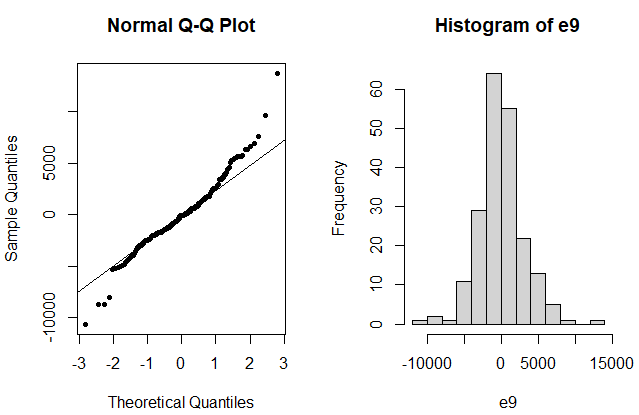
This model is obtained by stepwise backward method and by doing this backward method manually without ‘citympg’ and ‘highwaympg’.

Starting the diagnostics:

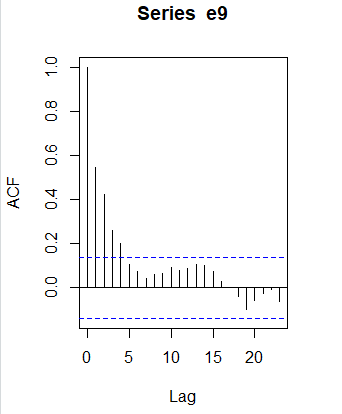
* 1. The mean of residuals is 3.096147e-13 which is very close to zero. So, we don’t have a problem with the mean.
  2. By plotting the residuals vs fitted values, we can say that we don’t have equality of variances. Firstly, the variance was small, but it increases later.



* 1. The normal QQ-plot and the histogram of residuals show us that we have a normal distribution with outliers.



* 1. For the autocorrelation, we have a problem with the residuals



1. Model B – after the log transformation of Y:

modelB <- lm(log(price) ~ fueltype + carwidth + curbweight + enginesize + stroke + horsepower + peakrpm + citympg + highwaympg, data = data)

AIC = -109.6065

* 1. Summary of the model

*Call:*

*lm(formula = log(price) ~ fueltype + carwidth + curbweight +*

*enginesize + stroke + horsepower + peakrpm + citympg + highwaympg,*

*data = data)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-0.55712 -0.11526 -0.02500 0.09788 0.47217*

*Coefficients:*

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

*(Intercept) 6.544e+00 7.696e-01 8.502 4.81e-15 \*\*\**

*fueltypegas -2.985e-01 6.323e-02 -4.720 4.49e-06 \*\*\**

*carwidth 2.699e-02 1.190e-02 2.269 0.024392 \**

*curbweight 3.240e-04 8.467e-05 3.827 0.000175 \*\*\**

*enginesize 2.810e-03 7.820e-04 3.593 0.000414 \*\*\**

*stroke -1.360e-01 4.315e-02 -3.152 0.001876 \*\**

*horsepower 2.464e-03 8.164e-04 3.019 0.002878 \*\**

*peakrpm 1.083e-04 3.682e-05 2.942 0.003658 \*\**

*citympg -3.129e-02 9.441e-03 -3.315 0.001094 \*\**

*highwaympg 1.748e-02 8.743e-03 1.999 0.046972 \**

*---*

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

*Residual standard error: 0.18 on 195 degrees of freedom*

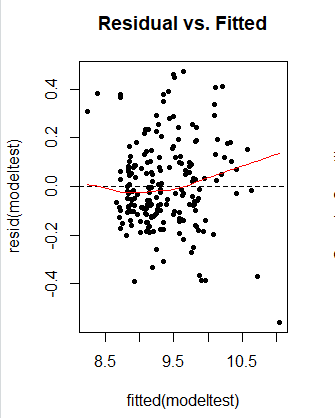
*Multiple R-squared: 0.878, Adjusted R-squared: 0.8724*

*F-statistic: 155.9 on 9 and 195 DF, p-value: < 2.2e-16*

This model is obtained by stepwise backward method and by doing this backward method manually with all the variables which are significant.

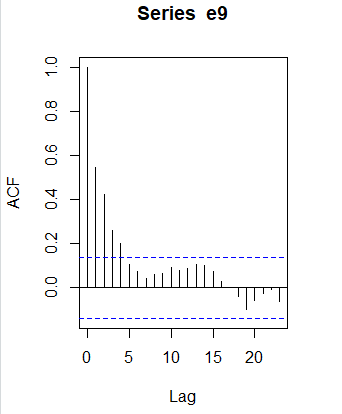
Starting the diagnostics:

1. The mean of residuals is -6.210098e-18 which is very close to zero. So, we don’t have a problem with the mean.
2. By plotting the residuals vs fitted values, we can say that we don’t have equality of variances. Firstly, the variance was small, but it increases later.



1. The normal QQ-plot and the histogram of residuals show us that we have a normal distribution with outliers.Chart, histogram

   Description automatically generated
2. For the autocorrelation, we have a problem with the residuals



To check the multicollinearity between the variables, we can use the variance inflation factor (VIF) function, and all the values we get are less than 10, which we can see that we have curbweight, citympg and highwaympg are multicollinearity.

*fueltypegas carwidth curbweight enginesize stroke horsepower*

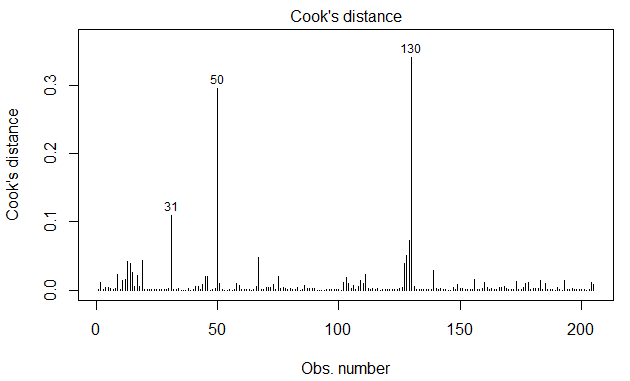
*2.227965 4.100948 12.241852 6.678641 1.153102 6.563645*

*peakrpm citympg highwaympg*

*1.942954 24.026150 22.828273*

Since we used al our data in our model B, and it has the lowest AIC we will continue our study using it.

We noticed that we have outliers in our data, so first we will check if we have influential points that can affect the estimation of parameters and later if we have outliers that can affect the equality of variances and the normality of residuals. And every influential point that is an outlier at the same time will be dropped out and check it by using cook’s distance

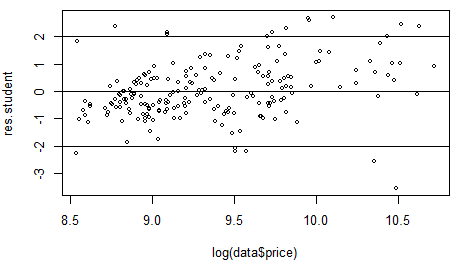


Every point has a cook’s distance greater than 0.5 is considered an influential point. But in our case, all the values are less than 0.5 so we don’t have influential points.

To detect outliers or atypical points, we have two methods.

* First method:

It is based on standardized residuals, and we plot the graph to show the outliers clearly.



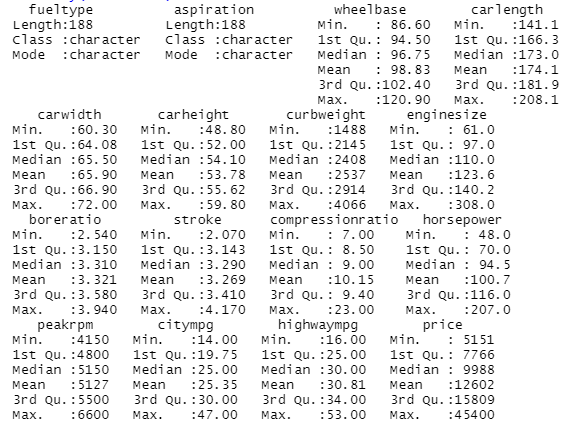
The points that are less than -2 and greater than 2 are considered outliers. We can see the outliers on this graph.

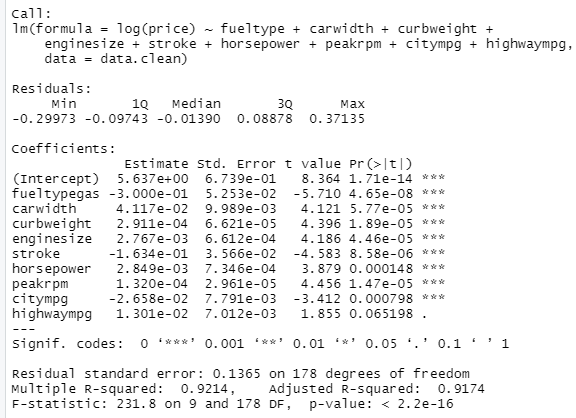
There exist 17 outliers using this method.

**Cleaning Data**

We will exclude the outliers from the data, so we keep all the rows except the outliers. And after that our data is clean.

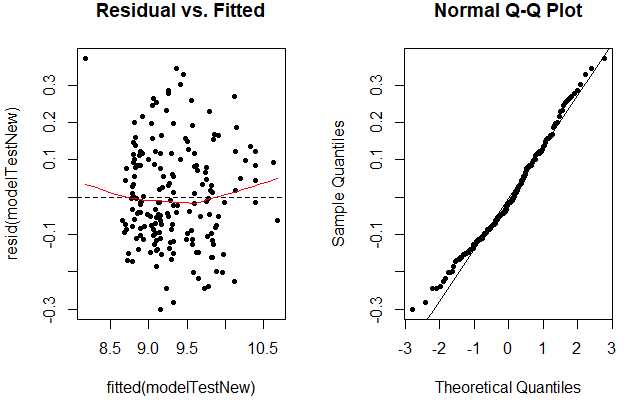
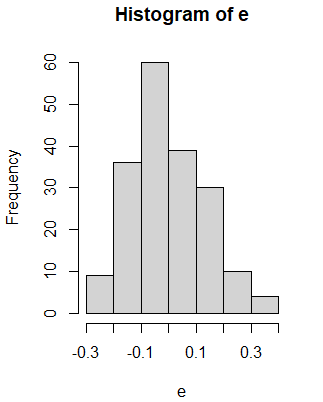
The dimension of the clean data is 188 rows and 16 columns.



* Summary of model B with clean data

With the clean data, the R squared increases and the AIC decreases to -203.4942 which is good for the model.

We will check the diagnostics for this model:

* The mean is always very close to zero (-2.099777e-18).
* The equality of variances and the normality is better now with some outliers.

**Prediction**

Using the testing data containing 47 rows, we will do the prediction of the price of used cars.

We use the predict function and we got these results. But we should take into consideration that we predict log(Price) so we will apply the exponential function to the results.

*fit lwr upr*

*1 8306.587 6551.810 10531.348*

*2 8617.622 6789.118 10938.595*

*3 6684.093 5272.274 8473.970*

*4 7423.213 5909.537 9324.603*

*5 7462.207 5941.034 9372.870*

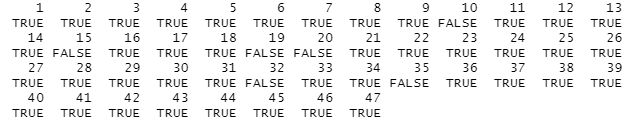
*6 7208.348 5738.666 9054.418*

This is the head of the predicted values with the confidence interval for every value.

The confidence interval is not considered large regarding the distribution of the data, this indicates that the prediction is precise.

**Cross-validation**

After prediction, we should compare the predicted values to the true values and then calculate the accuracy.



If the true value is in the confidence interval of the predicted value, it will give TRUE, and if not, it will give False.

We have only six False values from a total of 47 predictions, so the accuracy is 87.234% > 75% which indicates that this is a good model.

**Conclusion**

Based on the results of regression, 16 features: fuel type, aspiration, Wheelbase, car length, car width, car height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg has an impact on the price of cars.