Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

Lab 3: Modeling the Fuel Efficiency of Cars with R

Introduction

The goal of this lab is to examine the data set Cars93 which is included in R, then modeling the fuel efficiency of cars. The data contains 93 cars, with 27 variables per car.

There are two MPG variables in the data, one indicates the fuel efficiency of cars when driving in city, and the other indicates the fuel efficiency of cars when driving on highway. Therefore, we develop two models, one for each MPG variable.

Analysis of Data

We classify the 25 other variables as:
Quantitative: Min.Price, Price, Max.Price, Cylinders, EngineSize, Horsepower, RPM, Rev.per.mile, Fuel.tank.capacity, Passengers, Length, Wheelbase, Width, Turn.circle, Rear.seat.room, Luggage.room, Weight; and
Categorical: Manufacturer, Model, Type, AirBags, DriveTrain, Man.trans.avail, Origin, Make.

There are two quantitative variables that have missing values: Rear.seat.room (2 missing values) and Luggage.room (11 missing values). The missing values in Rear.seat.room occur in Chevrolet Corvette and Mazda RX-7, which both are sporty cars with no rear seats. Both cases indicate that the missing values are not “randomly missing”.

We also notice that Mazda RX-7's has a value “rotary” in Cylinders. This is because it has an unusual rotary engine. Therefore, it is reasonable to turn this value into NA.

Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

Analysis of Relationships

To construct the models, we first examine the relationships between the MPG's and each other variable.

For relationships between the MPG's and each quantitative variable, we determine the correlations. Notice that a NA value produces a NA correlation; with only slight influences in the results, we delete the rows which contain the NA value.

	Min.Price	Price	Max.Price	Cylinders
MPG.city	-0.62287544	-0.594562163	-0.54781090	-0.687222
MPG.highway	-0.57996581	-0.560680362	-0.52256074	-0.6361739

	EngineSize	Horsepower	RPM	Rev.per.mile
MPG.city	-0.7100032	-0.672636151	0.363045129	0.6958570
MPG.highway	-0.6267946	-0.619043685	0.313468728	0.5874968

	Fuel.tank.capacity	Passengers	Length	Wheelbase
MPG.city	-0.8131444	-0.416855859	-0.6662390	-0.6671076
MPG.highway	-0.7860386	-0.466385827	-0.5428974	-0.6153842

	Width	Turn.circle	Rear.seat.room	Luggage.room
MPG.city	-0.7205344	-0.6663889	-0.3843469	-0.4948936
MPG.highway	-0.6403592	-0.5936833	-0.3666844	-0.3716291

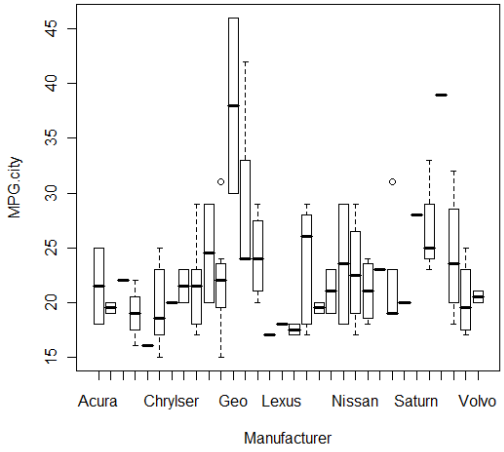
	Weight
MPG.city	-0.8431385
MPG.highway	-0.8106581

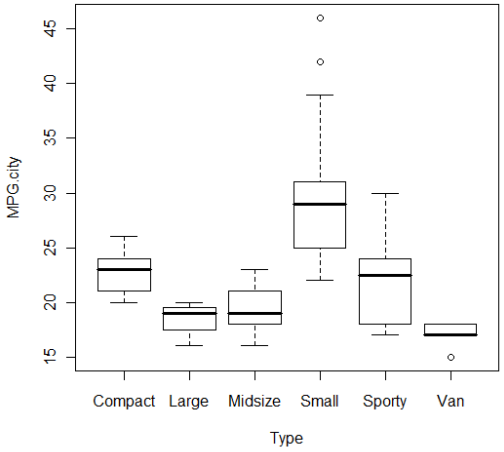
The correlation table above indicates that the MPG's are more or less related to each quantitative variable. The most and least related quantitative variables for both MPG's are Weight and RPM, respectively.

Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

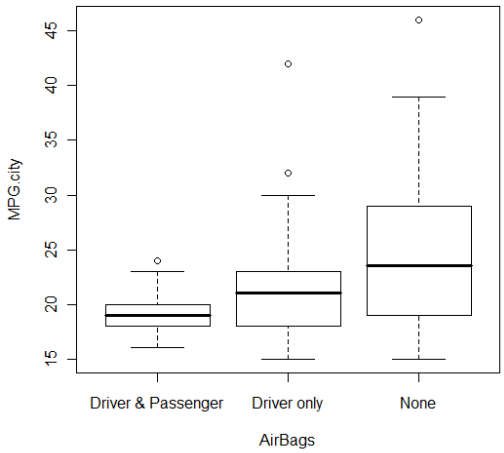
For relationships between the MPG's and each categorical variable, we construct boxplots. However, for variables Model and Make, due to the diversity of their values, examining their relationships with the MPG's would be worthless.

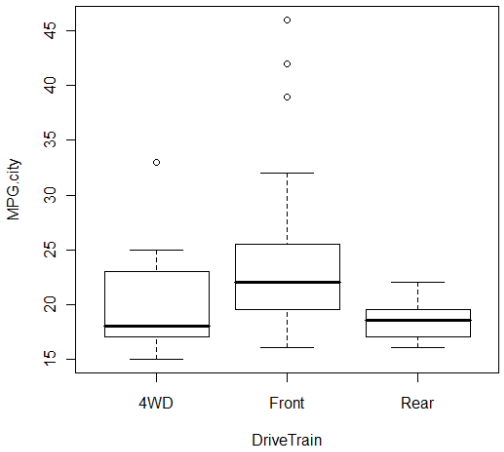
MPG.city vs. Categorical variables





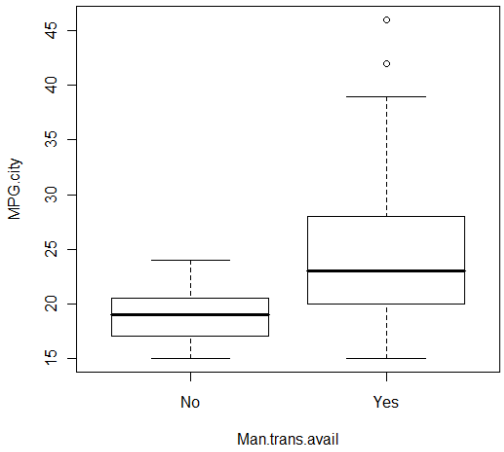
The distributions of MPG.city over Manufacturer are roughly the same, except for Geo and Suzuki. There might be two reasons: first, Geo and Suzuki both contain high MPG.city values, which are in the 39th observation & 83th; second, the sample size of Geo and Suzuki and are too small. Therefore there is no obvious relationship between MPG.city and Manufacturer.
The boxplot of MPG.city over Type shows signs that small cars generally have larger MPG.city values.

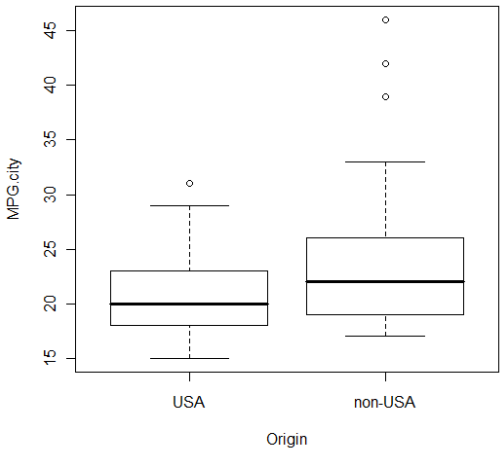




Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

No obvious relationship between MPG.city and AirBags, as well as between MPG.city and DriveTrain.

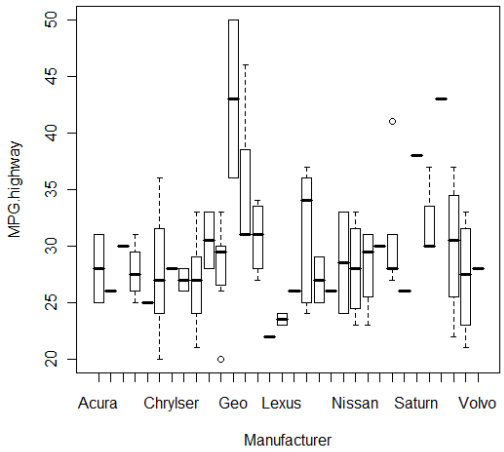


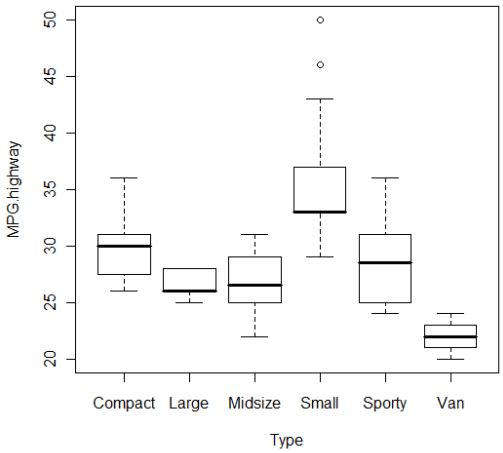


Evidence shows that the distribution of MPG.city over Man.trans.avail(Yes) is wider than the distribution of MPG.city over Man.trans.avail(No).

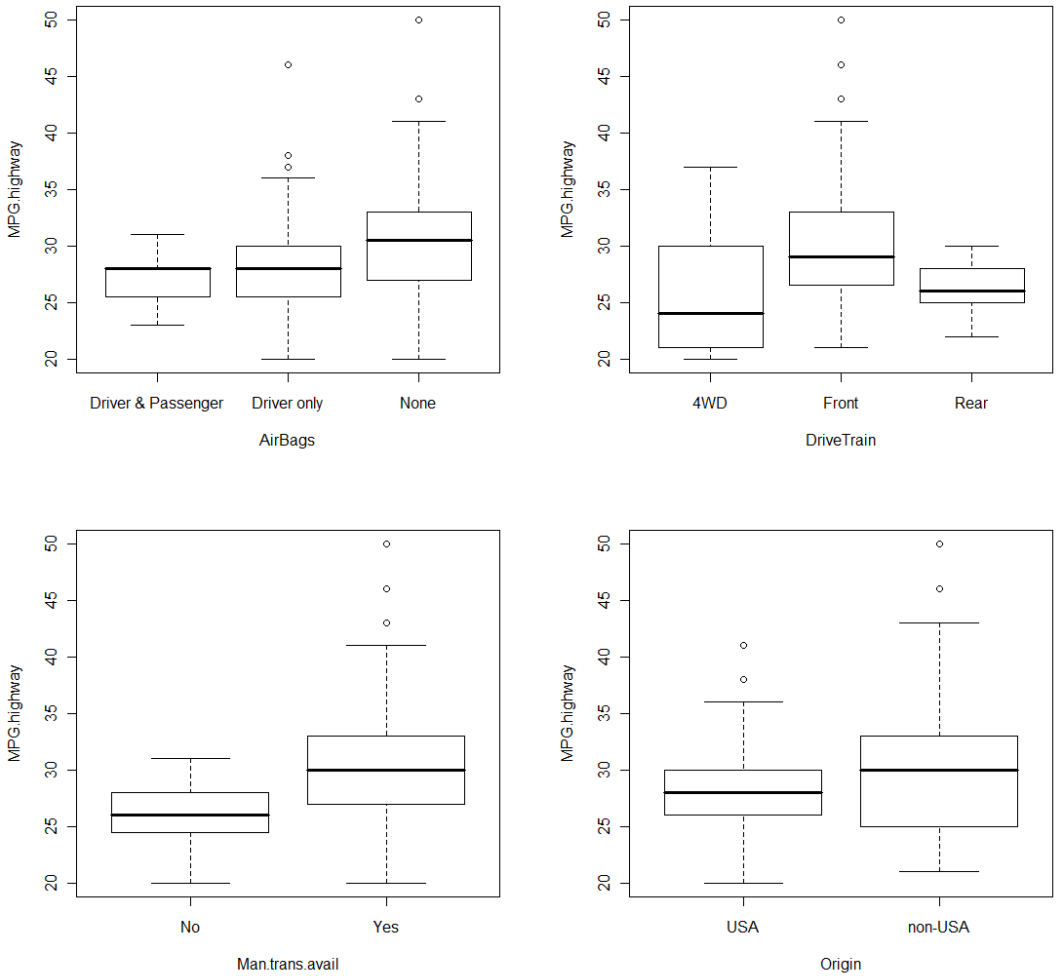
Again, there are no obvious relationship between MPG.city and Origin.

MPG.highway vs. Categorical variables





Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com



Notice that the boxplots of MPG.highway over categorical variables are highly consistent with the boxplots of MPG.city over categorical variables.

We may conclude that there might be relationships between the MPG's and Type, between the MPG's and Man.trans.avail. As evidences show that small cars generally have larger MPG's values; and distributions of MPG's over Man.trans.avail(Yes) is wider than distributions of MPG's over Man.trans.avail(No).

There are no obvious relationship between the MPG's and other categorical variables. Although the unusual high MPG's values in the 39^{th} & 83^{th} observation as well as the lack of samples for Geo and Suzuki generate abnormal distributions of MPG's over Manufacturer for Geo and Suzuki.

Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

Model Selections

Model selection is intended to explain the data in the simplest way.
To select appropriate linear models for MPG.city and MPG.highway, identifying suitable variables as predictors is necessary.

Categorical variables are not considered here as they have no predictive values. The correlations between the MPG's and quantitative variables illustrate that there are certain kinds of relationship between the MPG's and each quantitative variable. So we choose the linear models with all quantitative variables as predictors as our full models.

i. Linear model for MPG.city

First we perform a criterion-based procedure — minimize AIC.

The linear model we obtained is:

MPG.city ~ Max.Price + Price + Min.Price + RPM + EngineSize + Wheelbase + Weight + Rev.per.mile + Fuel.tank.capacity

Coefficients:

(Intercept)	Max.Price	Price	Min.Price	RPM	EngineSize	Wheelbase
11.974017	-7.304880	14.702794	-7.530689	0.001195	2.305197	0.200422
Weight	Rev.per.mile	Fuel.tank.capacity				
-0.006343	0.003771	-0.606712				

As we can see, there are still 9 predictors in this model. To insure we get the simplest model which fits the data, we then perform the backward elimination on the above model. At each step, we remove the predictor with highest p-value greater than 10%.

The linear model we obtained is:

MPG.city ~ EngineSize + Wheelbase + Weight + Rev.per.mile + Fuel.tank.capacity

Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

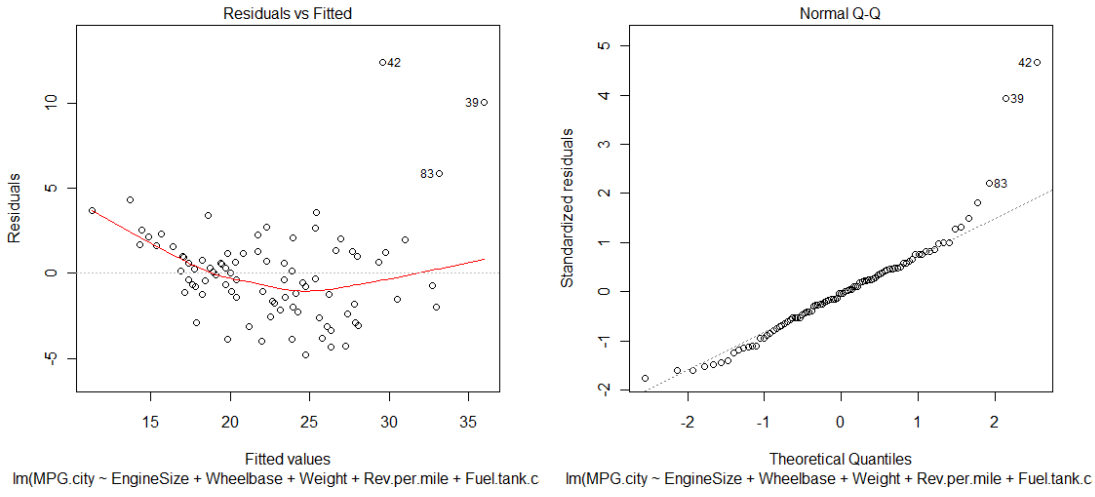
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.750161	7.498097	2.767	0.00690 **
EngineSize	1.245283	0.631130	1.973	0.05166 .
Wheelbase	0.207006	0.086661	2.389	0.01907 *
Weight	-0.006938	0.001685	-4.118	8.66e-05 ***
Rev.per.mile	0.003396	0.001053	3.224	0.00178 **
Fuel.tank.capacity	-0.589453	0.201029	-2.932	0.00430 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

There are 5 predictors in this model, each with a p-value less than 10%. This result is consistent with the correlations we obtained before: each predictor has a relatively large correlation. Therefore we have reason to believe this model is appropriate.
However, we are not able to draw a precise conclusion without running a series of tests.

Error Analysis

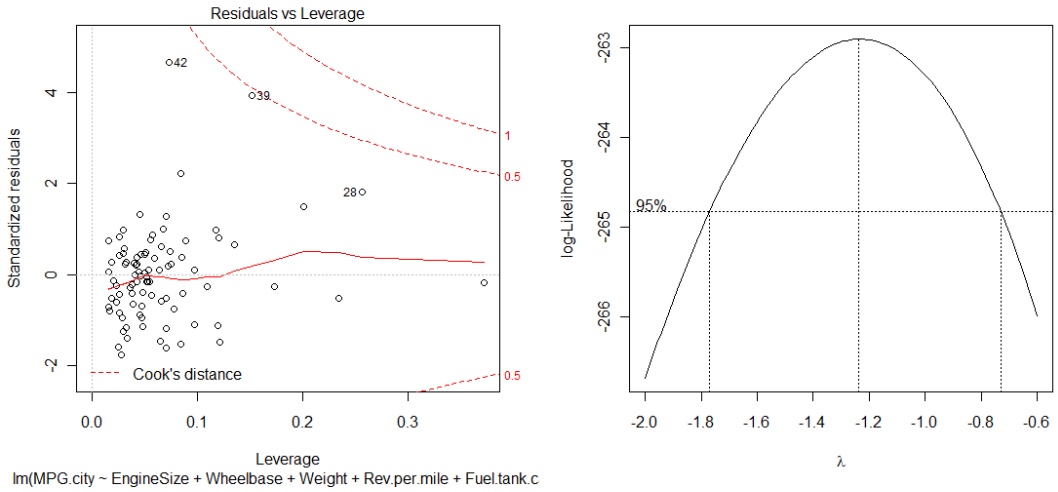


We first look at the residual plot. The scatters are roughly symmetric vertically about the 0 line; which suggests that the variance of the model is constant. Therefore transformation for variables is not needed.

Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

Also, we are able to identify the three noticeable outliers. They are the 42th (of value 42), 39th (of value 46) and 83th (of value 39).

Next, we check the Q-Q plot for normality assumption. The residuals follow the line approximately, except for the three outliers we identified earlier. Therefore the normality assumption is retained.



Last but not least, we check the leverage plot as well as the Cook's distance. The Cook's distance identifies the influential points. There are three noticeable points with large Cook's distance, which means they are influential points in this model. They are the 42th, 39th and 28th. The 42th has leverage less than 0.1, the 39th has leverage of approximately 0.15, while the 28th has a large leverage of 0.26.

Last, we check the Log-Likelihood plot to determine that whether a Box-Cox transformation is needed. The 95% confidence interval for λ runs from about -1.8 to -0.7. Therefore a Box-Cox transformation with λ = -1 may be considered.

ii. Linear model for MPG.highway

We follow exactly the same procedures as selecting the appropriate linear model for MPG.city.

The linear model obtained by performing criterion-based procedure — minimize AIC is:

Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

MPG.highway ~ Min.Price + Price + Max.Price + EngineSize + RPM + Rev.per.mile +
Fuel.tank.capacity + Passengers + Wheelbase + Luggage.room + Weight

Coefficients:

(Intercept)	Min.Price	Price	Max.Price	EngineSize	RPM
5.915984	-9.072432	17.684633	-8.771892	2.269392	0.001340
Rev.per.mile	Fuel.tank.capacity	Passengers	Wheelbase	Luggage.room	Weight
0.002454	-0.682375	-1.795119	0.444507	0.370491	-0.007884

The linear model obtained from the above model by performing the backward elimination is:

MPG.highway ~ Fuel.tank.capacity + Passengers + Wheelbase + Luggage.room + Weight

Coefficients:

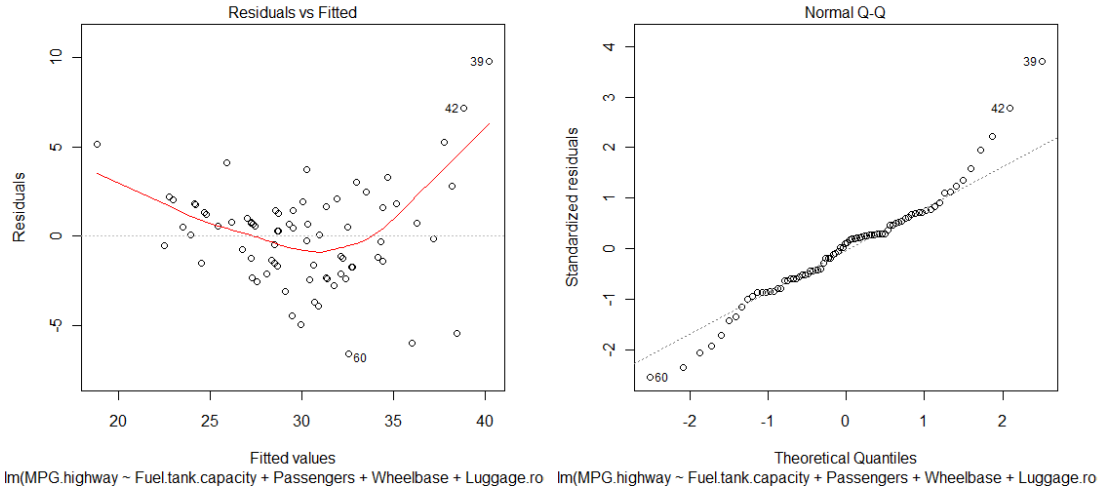
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.957065	8.367130	2.744	0.007575 **
Fuel.tank.capacity	-0.454275	0.238684	-1.903	0.060795 .
Passengers	-1.967672	0.621566	-3.166	0.002227 **
Wheelbase	0.451142	0.120719	3.737	0.000358 ***
Luggage.room	0.349530	0.163638	2.136	0.035899 *
Weight	-0.009156	0.001613	-5.677	2.39e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

There are 5 predictors in this model, each with a p-value less than 10%. This result is consistent with the correlations we obtained before: each predictor has a relatively large correlation. Therefore we have reason to believe this model is appropriate.
However, we are not able to draw a precise conclusion without running a series of tests.

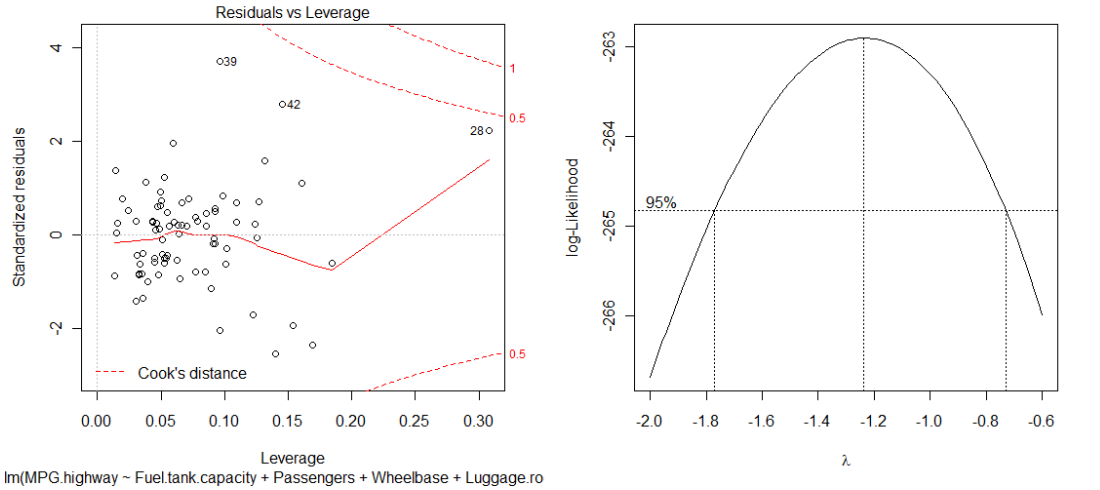
Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

Error Analysis



The scatters in the residual plot are roughly symmetric vertically about the 0 line; which suggests that the variance of the model is constant. Therefore transformation for variables is not needed. Also, there are two noticeable outliers in the plot. They are the 42th (of value 42), 39th (of value 46).

Next, we look into the Q-Q plot for normality assumption. The residuals follow the line approximately, except for the two outliers we identified earlier. Therefore the normality assumption is retained.



Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

The leverage plot as well as the Cook's distance are then checked. The influential points in this model (points with large Cook's distance) are the same as the influential points in the model of MPG.city. That is, the 42th, 39th and 28th. However, their leverages are different. The 39th has leverage of approximately 0.1, the 42th has leverage of approximately 0.15, while the 28th has a large leverage of 0.3.

Last, we check the Log-Likelihood plot to determine that whether a Box-Cox transformation is needed. The 95% confidence interval for λ runs from about -1.8 to -0.7. Therefore a Box-Cox transformation with λ = -1 may be considered.

Conclusion

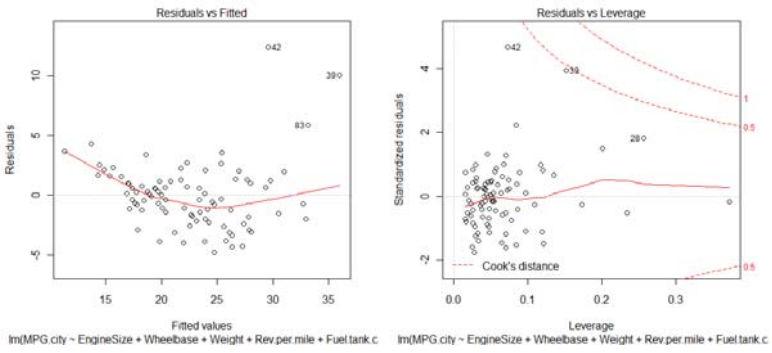
We firstly examined the relationships between the MPG's and each other variable besides Model and Make. We determined the correlations for quantitative variables and plotted the boxplots for categorical variables. We identified the most and least related quantitative variables for both MPG's are Weight and RPM, respectively. We indicated the existence of relationships between the MPG's and Type, and between the MPG's and Man.trans.avail.

Then we selected two appropriate linear models first through criterion-based procedure (minimize AIC), then followed by the backward elimination.

The linear model for MPG.city is:

MPG.city = 1.245283*EngineSize + 0.207006*Wheelbase - 0.006938*Weight + 0.003396*Rev.per.mile - 0.589453*Fuel.tank.capacity

With the residual plot and leverage plot:

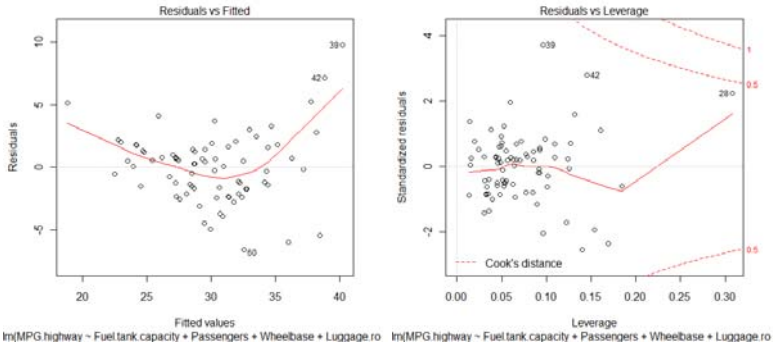


Hong Zhou
Regression Analysis Fall 2011
felicievip@gmail.com

The linear model for MPG.highway is:

MPG.highway = - 0.454275*Fuel.tank.capacity - 1.967672*Passengers + 0.451142*Wheelbase + 0.349530*Luggage.room - 0.009156*Weight

With the residual plot and leverage plot:



In the error analysis followed after the model selection, we checked the constant variance assumption and normality assumption, identified all outliers and influential points with their leverages, we also considered the possibility of using Box-Cox transformation.

Based on our findings and the consistency of all the analysis, we identify the two models above as our strongly endorsed models.