

# **Fuel Efficiency: Relationship between Bore-Stroke Ratio and MPG City**

Regression Theory-261A  
11-29-2017

By:  
Kevin Wachs  
Noe Vidales  
Travis Barton  
Yuval Vardi

## Motivation:

Consumers in recent years have become extremely sensitive to fuel prices, and as a result fuel efficiency has become a definitive consideration when purchasing a new vehicle. With the advent of hybrid and electric vehicles, manufacturers have been forced to find innovative ways to make the traditional combustion engine competitive amongst electric/hybrid alternatives. This tradeoff between fuel efficiency and the traditional combustion engine design motivated our study and our research question. In particular, our research question is to determine if there is a statistically significant relationship between the Bore/Stroke Ratio of an engine and city miles per gallon.

The theory relating fuel efficiency and Bore/Stroke Ratio is well established(e.g., Filipi et al., 2000)<sup>1</sup>, our research interest is in the magnitude of this relationship. The relationship between Bore/Stroke Ratio and fuel efficiency is best demonstrated by the Figure 1.

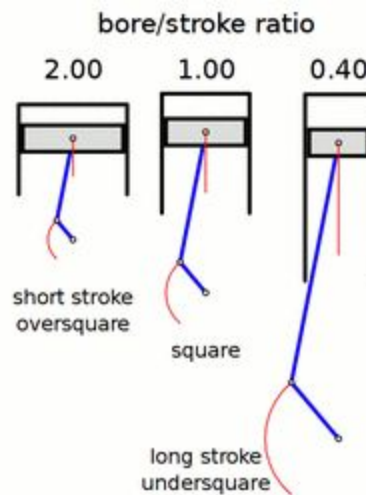


Figure 1: The relationship between Bore/Stroke Ratio and fuel efficiency

A smaller Bore/Stroke Ratio will increase the compression ratio.<sup>2</sup> Within the combustion chamber a ratio of fuel and air is mixed, a piston compresses the mixture, a spark plug ignites the mixture thrusting the piston down and turning the crankshaft, producing heat as a byproduct. A larger compression ratio(smaller Bore/Stroke Ratio) will reduce heat transfer (escape) and conserve energy directly increasing efficiency. Ultimately, more work is performed with less fuel; thus, increasing fuel efficiency. By discerning the magnitude of this tradeoff, manufactures will

<sup>1</sup> Z. S. Filipi, D. N. Assanis The effect of the stroke-to-bore ratio on combustion, heat transfer and efficiency of a homogeneous charge spark ignition engine of given displacement *International Journal of Engine Research* Vol 1, Issue 2, pp. 191 - 208 First Published April 1, 2000 <https://doi.org/10.1243/1468087001545137>

<sup>2</sup>"Stroke-to-Bore Ratio: A Key to Engine Efficiency." Achates, 22 Apr. 2015, [achatespower.com/stroke-to-bore/](http://achatespower.com/stroke-to-bore/).

be able to make decisions on the design of the cylinder bore, and be better informed about the limitations that Bore/Stroke Ratio can produce in fuel efficiency.

## Data Description:

The data that was used to start the regression process was obtained from the University of California Irvine's Machine Learning Repository and was uploaded on May 19th, 1987. It consists of 205 observations each with 26 variables. Four observations (56,57,58,59) were removed from the dataset because they contained missing values. Of the 26, only 4 variables were used in the final model (Figure 2). 1 additional variable in the model was created by dividing the bore dimension of a cylinder head by the stroke length of the cylinder lever. Because it has been shown that the shorter the distance the exhaust has to travel inside of the piston, the more energy it retains, we believe that Bore/Stroke Ratio may be a contributory factor in estimating a vehicle's city miles per gallon<sup>3</sup>.

Name	Range	Type	Notes
fuel-type	diesel, gas	categorical	Included in the final model
aspiration	std, turbo	categorical	Included in the final model
engine-size	continuous from 61 to 326	quantitative	Included in the final model
compression-ratio	continuous from 7 to 23	quantitative	Included in the final model
bore-stroke-ratio	continuous from 0.7723343 to 1.579909	quantitative	Included in the final model

Figure 2: Variables included in the final model

## Model Selection:

The original dataset included a total of 26 variables of which 13 were chosen as the regressors for our model. The selection of the regressors was based solely on vetted theory. The pool of regressors was as follows (names coincide with variable names in code): fuel.type, aspiration, body.style, drive.wheels, eng.loc, eng.type, num.cyl, eng.size, fuel.system, compression.ratio, Bore/Stroke Ratio.<sup>3</sup> A backward model selection was performed with a significance level of .05 to determine the set of regressors better suited to predict MPG City(dependent variable).

At first attempt, the 13 regressors were included and systematically reduced one-by-one using the highest p-value as our discriminant. After every iteration, it became apparent that our estimated betas changed dramatically depending on the inclusion and exclusion of other regressors. Upon further examination, our dataset was found to have multicollinearity (Figure 3). All continuous numeric variables were compared for correlation:

---

<sup>3</sup> Description of variable names can be found in Appendix I

	mpg.city.inv	eng.size	compression.ratio	bore.stroke.ratio	curbweight	horsepower
mpg.city.inv	1.0000000	0.82122097	-0.29855555	0.3167180	0.8226105	0.8908362
eng.size	0.8212210	1.00000000	0.02489413	0.1988756	0.8589503	0.8290723
compression.ratio	-0.2985556	0.02489413	1.00000000	-0.1445429	0.1503502	-0.2057593
bore.stroke.ratio	0.3167180	0.19887562	-0.14454291	1.00000000	0.2661243	0.2719020
curbweight	0.8226105	0.85895033	0.15035019	0.2661243	1.00000000	0.7528630
horsepower	0.8908362	0.82907232	-0.20575928	0.2719020	0.7528630	1.00000000

Figure 3: Multicollinearity assessment through repetition

The regressors “Engine Size”, “Curb-Weight”, and “Horsepower” presented the highest correlations coefficients. Horsepower and Curb-Weight were removed from the list of potential regressors, reducing our pool to 11. A backward model selection, using all 11 regressors, was reperformed at the same significance level. The pilot model was as follows:

$$MPG = (Fuel\ Type) + (Aspiration) + (Engine\ Size) + (Compression\ Ratio) + (Bore/Stroke\ Ratio)$$

## Residual Analysis and Variable Transformations

After running our data through several model and variable selection techniques, we settled on a model that predicted Fuel Efficiency (measured by MPG City) as a function of the type of fuel that provided to the car, aspiration, the Engine Size, compression ratio, and Bore/Stroke Ratio. While the model has good prediction power and the variables included have statistical significance, it appears that some of the assumptions on our residuals (model errors) were not met. More specifically, it appears the residuals had a quadratic form and may be dependent on the mean, and non-normally distributed errors were found (Figure 4).

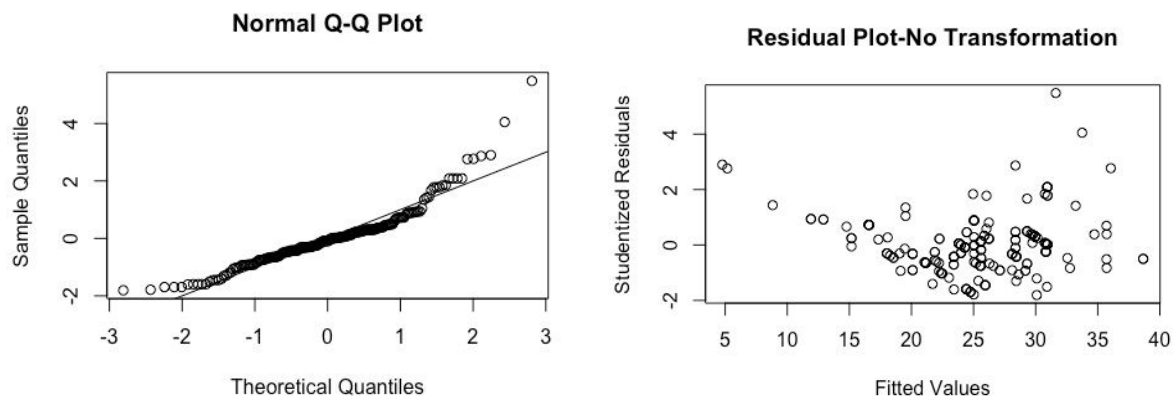


Figure 4: The Normal QQ (left) and Residual vs. Fitted (right) plots suggest non-normal data

As a result, we felt that a variable transformation on the response variable would allow us to maintain the integrity and simplicity of the model, while satisfying our assumptions on the errors. In order to find the best transformation, we performed a Box-Cox test, which produced a confidence interval of the recommended transformation on the response (in our case, MPG City) to maximize the log-likelihood function.

Applying the Box-Cox procedure suggested that the best transformation would be to change our response to 1/MPG City.

After applying the transformation, we ran the model again, this time with 1/MPG City as our response. The result was a model that kept the statistical significance of the predictors as before, but this time our model assumptions were met, with the normality of the residuals being true and the non-constant variance mitigated (Figure 5).

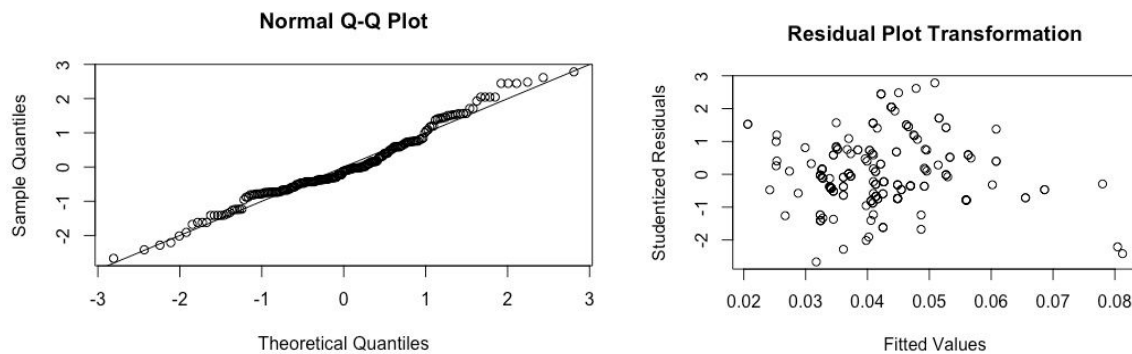


Figure 5: The Normal QQ (left) and Residual vs. Fitted (right) plots after the transform show that the assumptions no longer imply non-normal data

No observation was found to be an outlier using a cutoff of 3 and studentized residuals (Appendix II). Similarly, no observation was of concern when measured by “Cook’s Distance.” Many observations were found to have high leverage, but since none were coupled with an unusual outlier they were not found to be influential. Leverage is a measurement of distance of an observation from the centroid of the X-space. The further an observation is from the X-Space it increases its potential of overly influencing the fitted regression line. Observations 70,71,108,110,135,142, 179, 181, and 189 were found to be potentially influential, because they produced an unusual DFFIT value(cutoff .345) and a leverage (cutoff .059). DFFIT is a measurement of the standard deviations that a fitted observation changes as a result from the exclusion of the respective observation. A large DFFIT measurement indicates a potentially influential observation because it implies that the observation is pulling the fitted regression surface towards itself. There are a total of  $2^9 = 512$  potential combinations for subsets of the original dataset so due to time constraints, a manual systematic sampling would not be feasible to assess whether these observations are influential. Random subsets of size five were attempted to identify any variation in the fitted coefficients. After 100 iterations, Figure 6 displays the summary statistics produced on the fitted coefficients. There is some variation in the coefficients depending on which of the observations is removed, but not enough to warrant removal. The coefficients were plotted on a histogram (Appendix III)

Intercept.Coefficient	Fuel.Type.Coeff	Aspiration.Coefficient	Eng.Size.Coefficient	Compression.Ratio.Coeff	Bore.Stroke.Ratio.Coeff
Min. :0.04442	Min. : -0.02398	Min. :0.003976	Min. :0.0001989	Min. : -0.002761	Min. :0.006974
1st Qu.:0.04737	1st Qu.: -0.02085	1st Qu.:0.004490	1st Qu.:0.0002041	1st Qu.: -0.002541	1st Qu.:0.008377
Median :0.04874	Median : -0.02027	Median :0.004713	Median :0.0002068	Median : -0.002503	Median :0.008914
Mean :0.04860	Mean : -0.02016	Mean :0.004708	Mean :0.0002073	Mean : -0.002481	Mean :0.009051
3rd Qu.:0.04985	3rd Qu.: -0.01918	3rd Qu.:0.004879	3rd Qu.:0.0002124	3rd Qu.: -0.002406	3rd Qu.:0.009929
Max. :0.05340	Max. : -0.01668	Max. :0.005284	Max. :0.0002155	Max. : -0.002251	Max. :0.011104

Figure 6: summary statistics produced on the fitted coefficients after 100 iterations

## Sensitivity Analysis:

The three quantitative regressors (Engine Size, Compression Ratio, Bore/Stroke Ratio) were plotted against both MPG City and 1/MPG City to assess functional form (Appendix IV). The only regressor that appears to have a strong linear relationship with 1/MPG City is Engine Size. A correlation matrix validates this observation with Engine Size and 1/MPG city having a correlation coefficient of .82. Compression Ratio appears to have two clusters, but this is to be expected given that most gas combustion engines have a compression ratio below 10 and diesel engines typically have a compression ratio above 20. Bore/Stroke Ratio does not appear to have a strong linear relationship with 1/MPG City (correlation coefficient .31). As the plot for Bore/Stroke Ratio demonstrates there is high variability in the data; thus, causing a low correlation coefficient. A simple regression--using only the Bore/Stroke Ratio as the regressor--provided validated statistical significance, but produced an R2 of 9.8%(Appendix V). A confidence interval was created for the simple linear regression at the 95% confidence level (0.012958231 0.03195616) with an estimated coefficient of 0.0005043256. The wide margin of the confidence interval demonstrates the variability in the fitted coefficient for Bore/Stroke Ratio. While, the lacking of a relationship between Bore/Stroke Ratio and 1/MPG City is of concern, the full model has an R2 of 83% which is much stronger. The statistical significance of Bore/Stroke Ratio still implies a relationship with 1/MPG City, while not a strong one.

Another concern is the drastic change (60% reduction) in the estimated coefficient for Bore/Stroke Ratio when including control variables (.008943) and a simple linear regression (.022457). A combination of control regressors were attempted in order to determine if the coefficient for Bore/Stroke Ratio was dependent on other regressors (Figure 7). Inclusion or exclusion of Engine Size appears to be the culprit for the variation. The low correlation between the two variables(.19) was unable to explain this observation. Engine Size is the most important variable in the model, with it being able to explain 55% of the variation in 1/MPG City in a simple linear regression. As the figure below verifies, its removal from the model varies all fitted coefficients drastically. The correlation of eng.size with the remaining regressors is not alarming and a VIF of 1.11 does not warrant further consideration. Confounding variables may be affecting our estimates and an exploratory study may help in identifying them.

(Intercept)	aspirationturbo	eng.size	compression.ratio	bore.stroke.ratio
0.0161434572	0.0064009308	0.0002045965	-0.0009935354	0.0082905520
(Intercept)	fuel.typegas	eng.size	compression.ratio	bore.stroke.ratio
0.0827482050	-0.0411544659	0.0001977213	-0.0038341752	0.0096113830
(Intercept)	fuel.typegas	aspirationturbo	eng.size	bore.stroke.ratio
-0.006242093	0.013359699	0.007469845	0.000207490	0.007997070
(Intercept)	fuel.typegas	aspirationturbo	compression.ratio	bore.stroke.ratio
0.135186466	-0.065392963	0.002472317	-0.005579613	0.021182154

Figure 7: The combination of control variables tests dependencies on other variables

## Predictive Power:

Through a PRESS Residual analysis, the predictive coefficient of determination was found to be .829. This implies that just short of 83% of variability that would be seen in predictions can be explained by the model. This is a high number in terms of satisfaction and provides excellent evidence to support the claim that the model is a good predictor of 1/MPG City.

A cross-validation test was done to examine the potential instability in the model's predictive power. Five points were randomly selected to be removed from the data and then were replaced with predicted values from our model. The predicted values and original values were then compared. This process was repeated 1000 times and the resulting predictive coefficients of determination for each subset of data was plotted on a histogram (Figure 8). The figure shows that over 85% of the predictive coefficients of determination lie between .5 and 1, and over 95% are greater than 0. This supplies ample reason to believe that our model is stable when predicting new observations. Further evidence of stability can be seen the collective mean squared error for our 1000 iterated models. While the majority of mean squared error values did increase above the mean of our original model, all had a sigma value far below .0001 with the spread ranging in the tens of millionths.

Overall, between the high predictive coefficient of determination, and the stability of the model parameters, the model shows a strong ability to predict new observations with accuracy.

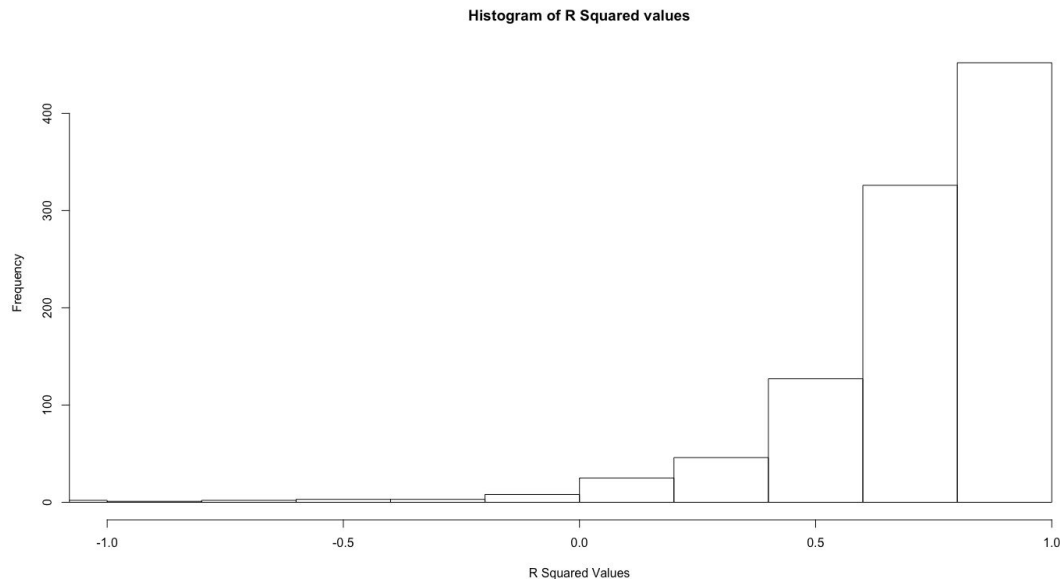


Figure 8: Coefficients of determination from 1000 iterations of testing plotted



## Findings:

Our original research question appears to be validated with a statistically significant coefficient for Bore/Stroke Ratio at the 10%, 5% and 1% significance level. Our model predicts an increase of .00894 in 1/MPG City for a unit increase in Bore/Stroke Ratio, resulting in a decrease in MPG City. According to our model Bore/Stroke Ratio and MPG City should have a negative relationship, which is as theorized. Because of the transformation of the dependent variable only the direction of the coefficient can be related to MPG City and the magnitude requires theory outside the scope of this paper. According to our model, a bigger engine, which usually is found in larger, less efficient cars, leads to a higher 1/MPG City value, which in return, leads to a lower MPG City. We believe that the most interesting finding is that a gas engine is more efficient on average than a diesel engine, holding all else constant (see future research suggestions section for more). This finding is not validated by accepted theory. Diesel engines have larger compression ratios as demonstrated by the graph plotting MPG City and Compression Ratio (Appendix IV) and as such should yield higher fuel efficiency. Furthermore, our model predicts a reduction in 1/MPG City for a unit increase of Compression Ratio which validates common theory. An increase in compression ratio should result in an increase in MPG City thus decreasing 1/MPG City.

Contrary to common theory our model predicts that a turbo-charged engine increases 1/MPG City, which decreases MPG City. A turbo engine is designed to force air into the combustion chamber; thus, delivering more power (efficiency) using less fuel and increasing MPG.

While our model demonstrates certain deficiencies, we believe that it is a great starting point and will require further analysis and data. Finally, we hope that this type of research would be able to educate consumers on what goes into the figure of fuel efficiency, and how to pick a car that would best serve an individual consumer's needs, while maximizing efficiency.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.693e-02  1.324e-02   3.545 0.000491 ***
fuel.typegas   -1.889e-02  7.982e-03  -2.366 0.018939 *
aspirationturbo 4.693e-03  1.102e-03   4.259 3.19e-05 ***
eng.size        2.004e-04  7.847e-06  25.533 < 2e-16 ***
compression.ratio -2.332e-03  5.713e-04  -4.082 6.52e-05 ***
bore.stroke.ratio 8.943e-03  2.126e-03   4.206 3.96e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.004351 on 195 degrees of freedom
Multiple R-squared:  0.8409,    Adjusted R-squared:  0.8368
F-statistic: 206.2 on 5 and 195 DF,  p-value: < 2.2e-16
```



## Limitations:

While the model produced statistically significant findings, the limitations of the study must be acknowledged. Our dataset consisted of 205 observations which does not substantiate a large enough cross-section of the automobile industry to make sweeping predictions. Similarly, the dataset is highly skewed towards international makes, with a limited representation of domestic vehicles, as demonstrated by Figure 9. Our model may be restricted in use for international makes. The strong dependence of the model on eng.size also requires further consideration.

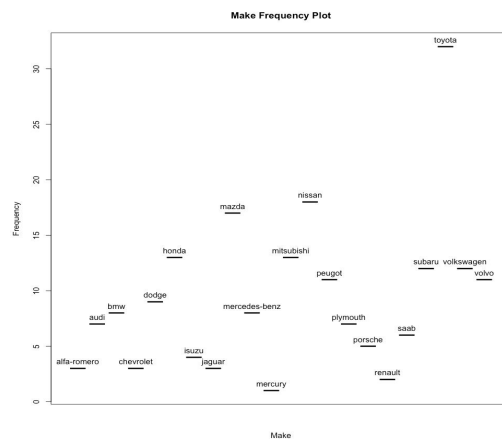


Figure 9: Plot of frequency of makes (exact figures in Appendix VII)

## Future Research Suggestions:

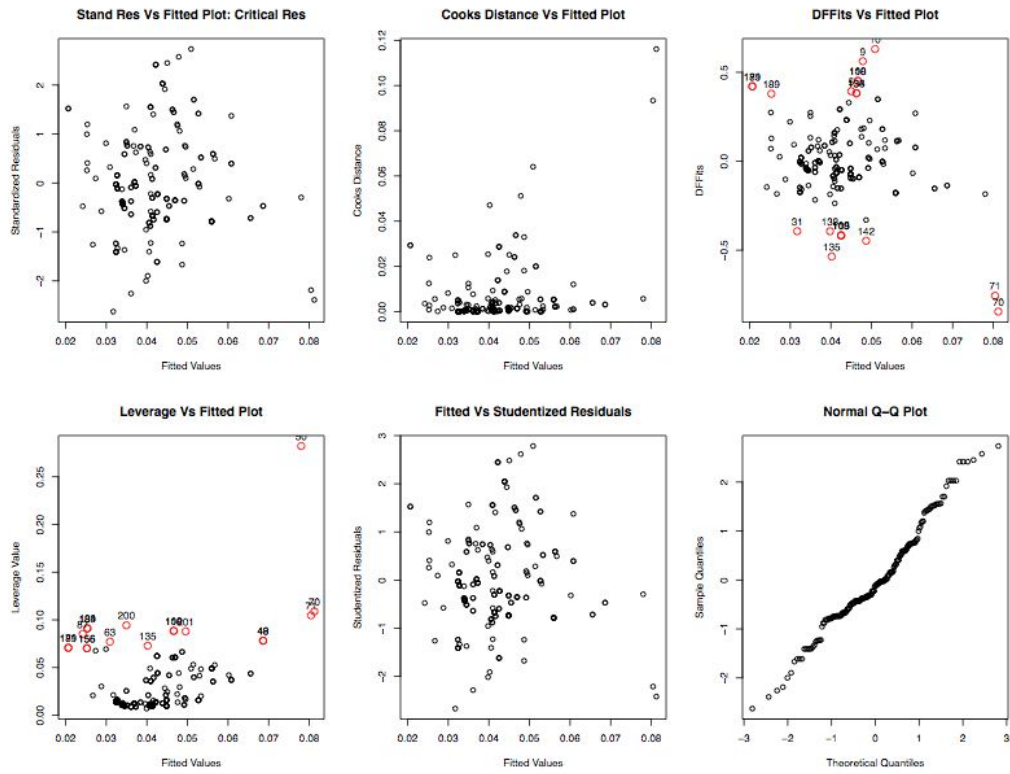
Since this regression analysis was conducted with a limited sample of 205, we recommend more data collection in order to fit an even better model that would be able to make predictions outside of the range of the current data. Further research may attempt to find the relationship between Bore/Stroke Ratio and the mpg of highway miles. Due to the strong dependency of our model on Engine Size, there may exist confounding variables that may need to be controlled for. An exploratory study may help in finding such variables. We would also recommend looking into more variables, such as resale data, different gas types (more granular than gas vs. Diesel), and how often such cars need to be maintained to explore other interesting questions.

## Appendix

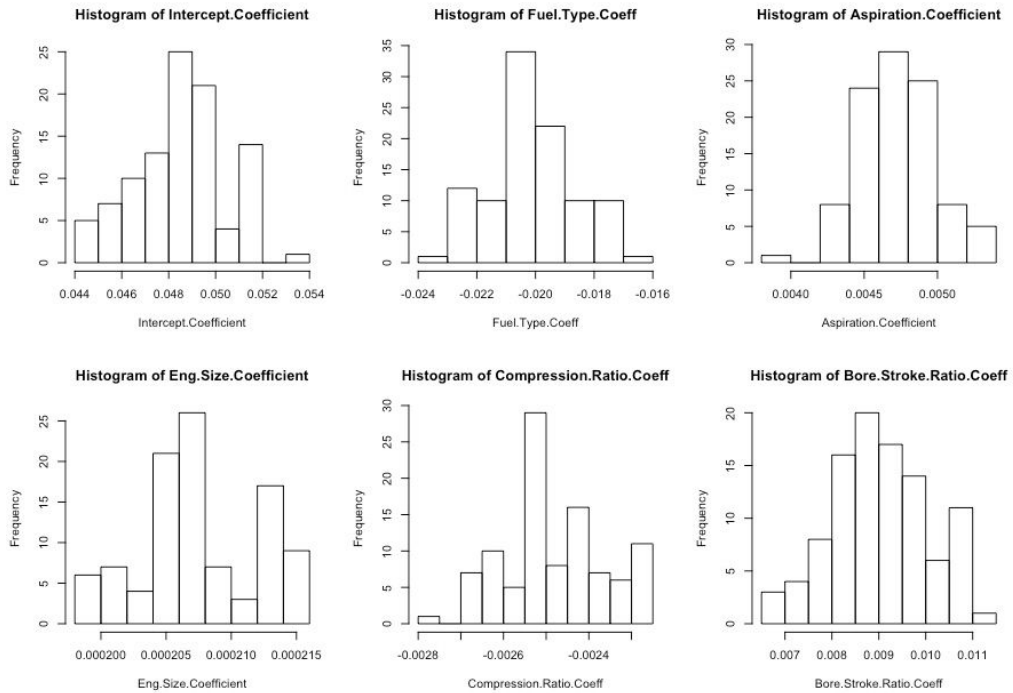
### I

Name	Range	Type	Notes
symboling	-3, -2, -1, 0, 1, 2, 3	categorical	
normalized-losses	continuous from 65 to 256	quantitative	
make	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo	categorical	
fuel-type	diesel, gas	categorical	Included in the final model
aspiration	std, turbo	categorical	Included in the final model
num-of-doors	four, two	categorical	
body-style	hardtop, wagon, sedan, hatchback, convertible	categorical	
drive-wheels	4wd, fwd, rwd	categorical	
engine-location	front, rear	categorical	
wheel-base	continuous from 86.6 to 120.9	quantitative	
length	continuous from 141.1 to 208.1	quantitative	
width	continuous from 60.3 to 72.3	quantitative	
height	continuous from 47.8 to 59.8	quantitative	
curb-weight	continuous from 1488 to 4066	quantitative	
engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor	categorical	
num-of-cylinders	eight, five, four, six, three, twelve, two	categorical	
engine-size	continuous from 61 to 326	quantitative	Included in the final model
fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi	categorical	
bore	continuous from 2.54 to 3.94	quantitative	
stroke	continuous from 2.07 to 4.17	quantitative	
compression-ratio	continuous from 7 to 23	quantitative	Included in the final model
horsepower	continuous from 48 to 288	quantitative	
peak-rpm	continuous from 4150 to 6600	quantitative	
city-mpg	continuous from 13 to 49	quantitative	
highway-mpg	continuous from 16 to 54	quantitative	
price	continuous from 5118 to 45400	quantitative	
bore-stroke-ratio	continuous from 0.7723343 to 1.579909	quantitative	Included in the final model

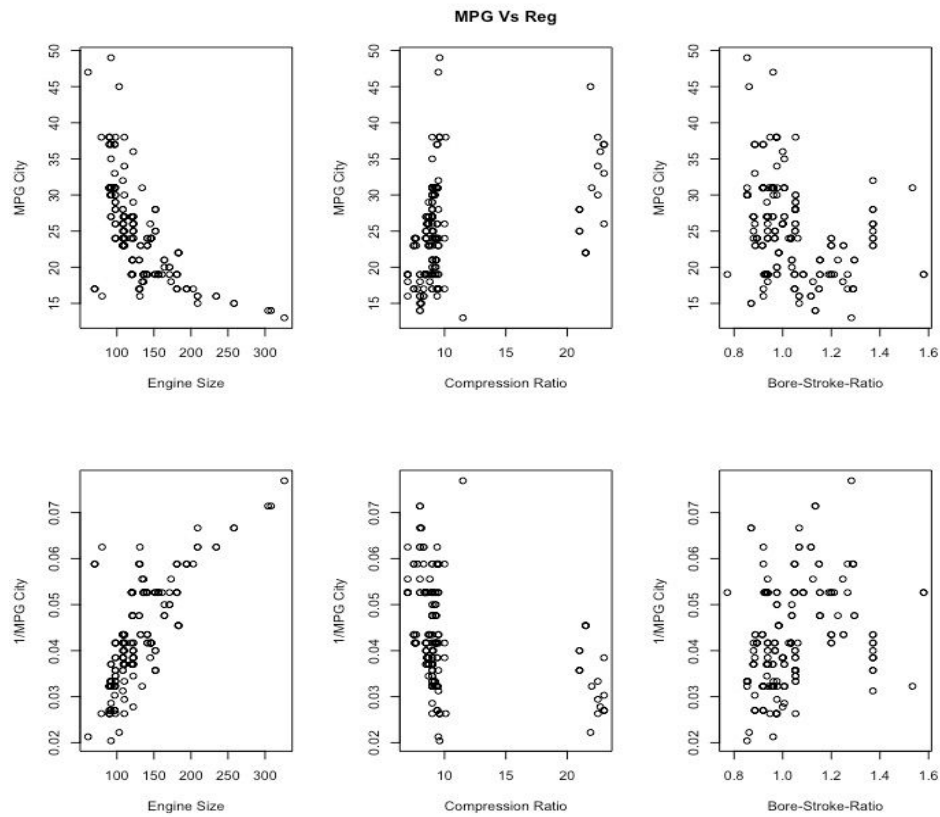
## II



## III



#### IV



#### V

Call:

```
lm(formula = mpg.city.inv ~ bore.stroke.ratio, data = data.auto.post.model.sel.y.tran)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.020952	-0.007666	-0.002233	0.006661	0.029357

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.018763	0.005035	3.727	0.000253 ***
bore.stroke.ratio	0.022457	0.004817	4.662	5.73e-06 ***

---

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

Residual standard error: 0.01025 on 199 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.09847, Adjusted R-squared: 0.09393

F-statistic: 21.73 on 1 and 199 DF, p-value: 5.73e-06

## VI

Make	Freq	Make	Freq	Make	Freq
alfa	3	mazda	17	renault	2
audi	7	mercedes	8	saab	6
bmw	8	mercury	1	subaru	12
chevy	3	mitsubishi	13	toyota	32
dodge	9	nissan	18	volkswagen	12
honda	13	peugot	11	volvo	11
isuzu	4	plymouth	7		
jaguar	3	porsche	5		

## VII

The following table list the observations that surpass the standard critical thresholds.  
The following thresholds were used:  
Standardized Residual > 3  
Cooks Distance > 1  
Leverage > 0.05970149  
Dffit > 0.3455474

## VIII

If an observation failed one of the thresholds all residual diagnostics are included for comparison.

obs.	Raw Res	Stud Res	Stand Res	Cooks Dist	Leverage	DFITT
5	0.0105312467	2.48256703	2.4503394	0.0251270172	0.008474614	0.39338776
9	0.0109655678	2.61586529	2.5775352	0.0511820706	0.040969724	0.56240002
10	0.0116106434	2.78276295	2.7358574	0.0640717017	0.036525119	0.63065452
31	-0.0113105129	-2.66817535	-2.6272715	0.0249623634	0.017991467	-0.39303198
48	-0.0019625452	-0.46880336	-0.4697441	0.0031196640	0.075066481	-0.13653970
49	-0.0019625452	-0.46880336	-0.4697441	0.0031196640	0.075066481	-0.13653970
50	-0.0010875634	-0.29430403	-0.2949957	0.0057022316	0.103140922	-0.18453491
60	0.0003960895	0.09402682	0.0942667	0.0001074115	0.066754940	0.02532180
63	0.0013556988	0.32356351	0.3243089	0.0014654384	0.063928883	0.09355352
64	-0.0014926506	-0.35313752	-0.3539328	0.0013495504	0.060655171	-0.08978283
65	-0.0014926506	-0.35313752	-0.3539328	0.0013495504	0.060655171	-0.08978283
66	-0.0014926506	-0.35313752	-0.3539328	0.0013495504	0.060655171	-0.08978283
67	-0.0014926506	-0.35313752	-0.3539328	0.0013495504	0.060655171	-0.08978283
70	-0.0098110464	-2.41787723	-2.3883814	0.1161391974	0.107278550	-0.84507568
71	-0.0090096108	-2.20992479	-2.1882409	0.0933787316	0.103010731	-0.75593068
87	-0.0019811053	-0.47508920	-0.4760352	0.0035238857	0.064499854	-0.14511843
105	-0.0067924030	-1.61849930	-1.6118198	0.0286905429	0.054751750	-0.41662089
108	0.0060075255	1.45009073	1.4460078	0.0338217917	0.083448502	0.45175033
109	-0.0067924030	-1.61849930	-1.6118198	0.0286905429	0.054751750	-0.41662089
110	0.0060075255	1.45009073	1.4460078	0.0338217917	0.083448502	0.45175033
113	-0.0067924030	-1.61849930	-1.6118198	0.0286905429	0.054751750	-0.41662089

133	0.0063302874	1.50566975	1.5008018	0.0241504010	0.039002066	0.38189518
134	0.0063302874	1.50566975	1.5008018	0.0241504010	0.039002066	0.38189518
135	-0.0079471772	-1.90952379	-1.8966974	0.0471078310	0.072815810	-0.53524078
138	-0.0085384320	-2.01483227	-1.9992095	0.0253776610	0.032929541	-0.39326202
142	-0.0070118918	-1.67539898	-1.6676900	0.0329658049	0.065843806	-0.44679715
146	-0.0052002976	-1.23851728	-1.2368252	0.0181321652	0.065843806	-0.33028908
155	0.0041815691	0.99658361	0.9966010	0.0125106884	0.066488741	0.27397355
156	0.0010855939	0.25811162	0.2587316	0.0008432132	0.066488741	0.07095818
171	0.0034103274	0.81167436	0.8123854	0.0081954451	0.061975844	0.22155504
179	0.0063796316	1.52603176	1.5208588	0.0293475481	0.070718310	0.42105246
181	0.0063796316	1.52603176	1.5208588	0.0293475481	0.070718310	0.42105246
184	0.0016868213	0.40572567	0.4065975	0.0027606660	0.070804370	0.12842521
189	0.0049628246	1.19758842	1.1962573	0.0238964583	0.070804370	0.37907522
195	0.0072560142	1.71009626	1.7017198	0.0200273204	0.038011689	0.34835299
196	0.0072560142	1.71009626	1.7017198	0.0200273204	0.038011689	0.34835299
200	0.0035041062	0.84552541	0.8461442	0.0124228713	0.066880643	0.27281542