Analysis of the Car93 Dataset Using Linear Regression Models

Tsu-Hao Fu

1. Summary

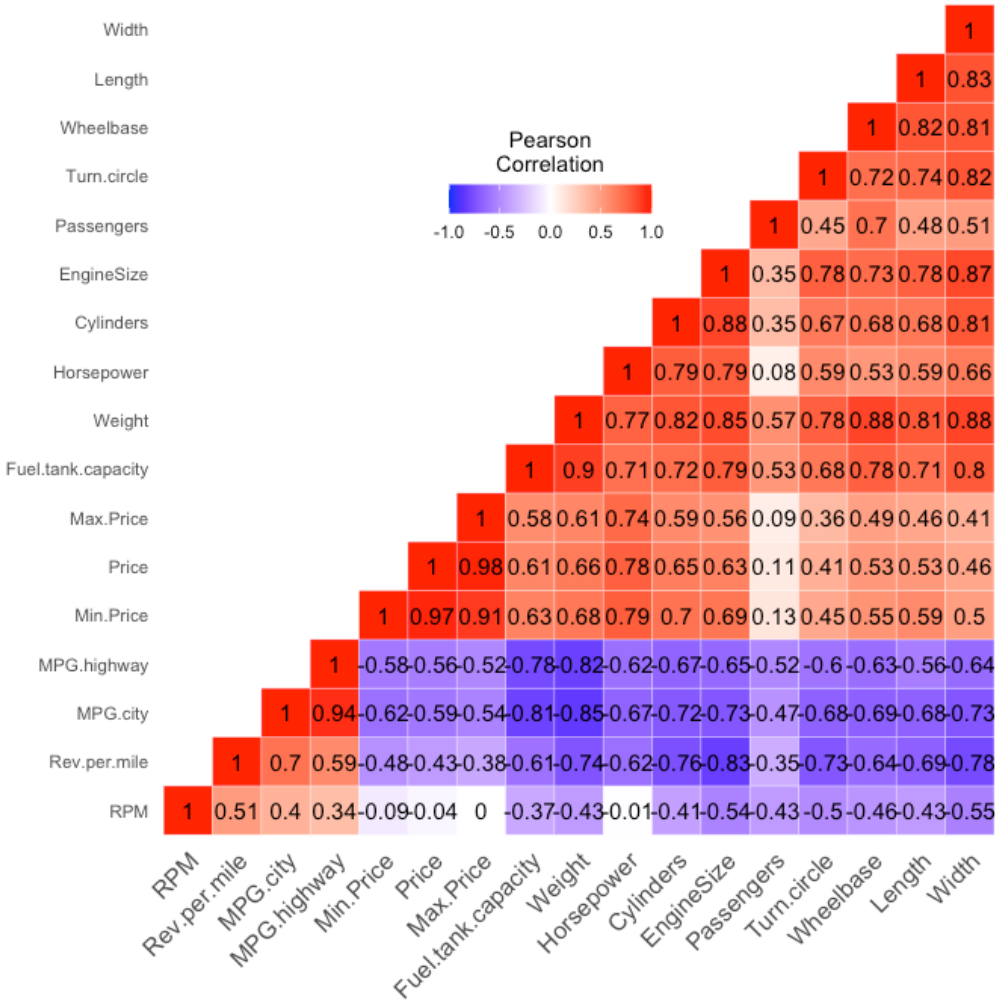
Due to the high correlation between “MPG.city” and “MPG.highway”, I decided to use the average of MPG.city and MPG.highway as the response variables to find out the most important variables for predicting the fuel consumption. To acquire my final model, I utilized the techniques of Exploratory Data Analysis (EDA), Box-Cox transformation, stepwise regression and the VIF test for multicollinearity.

My final fitted model includes “Weight”, “Length”, “Fuel.tank.capacity” and “Origin” as the predictor variables which all have significant influences on the average MPG. First of all, I found that “Weight” and “Fuel.tank.capacity” contrast with the fuel consumption. I think it is reasonable, since heavy-weight cars usually have bigger tanks, but their fuel economy are bad. Secondly, the length has positive coefficient in the model, however, it is negatively correlated with the average MPG. Therefore, even if I have already done the test for multicollinearity of the final model, I believe there still exist some problems of multicollinearity. Some other methods of variable selection, such as Lasso regression or random forest, might be needed for further research. Last but not least, it is interested that the model shows that the cars of non-USA company origins have better fuel economy. Most of the non-USA cars in this dataset are Japanese cars, therefore, I think it aligns with the public impression toward Japanese cars. They usually have better fuel economy.

2. Analysis

    1. Explore Data Analysis

        - Preliminary Variable Selection



Correlation matrix

According to the correlation matrix, I found that “Min.Price”, “Max.Price” and “Price” are highly correlated. Therefore, I decided to just use “Price” as one of my predictors. Also, “Cylinders”, the number of cylinders, has a unique value for Mazda RX-7 which has a rotary engine, therefore, I decided to remove this observation and convert the type of “Cylinders” from categorical to numeric. Furthermore, there are other 8 categorical variables. Since some categorical variables have large numbers of categories, I only used the 5 out of 8 categorical variables

which are “Type”, “AirBags”, “DriveTrain”, “Man.trans.avail”, “Origin” and transformed them to dummy variables. These dummy variables and the numeric variables are my initial set of predictors.

- Missing Values

There are several missing values in the dataset. Since all the 11 missing values occur in two columns, “Rear.seat.room” and “Luggage.room”, and our sample size is already very small. Also, these two columns are just slightly correlated to “MPG.city” and “MPG.highway”, therefore, I decided to remove these two variables.

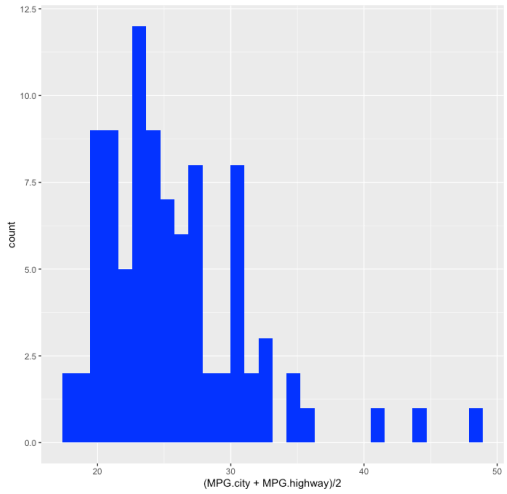
- Multicollinearity

Based on the correlation matrix, several variables are highly correlated to each other, such as “Weight” and “Fuel.tank.capacity”, “EngineSize” and “Width”, “Wheelbase” and “Length”, etc. These phenomena might result in less reliable statistical inferences. Hence, I would apply the test for multicollinearity to better interpret the fitted models.

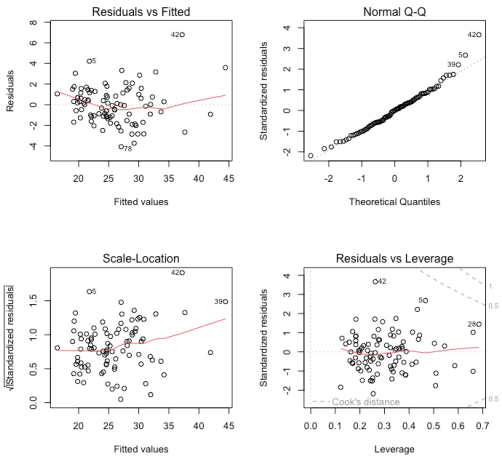
2. Linear model analysis

- Transformation

From the histograms of the response variable, it is right-skewed, so I believe some transformation would be needed to ensure the normality.



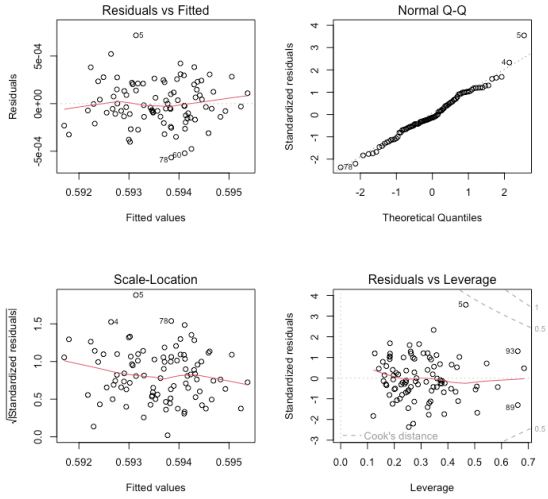
Histogram of the avg MPG



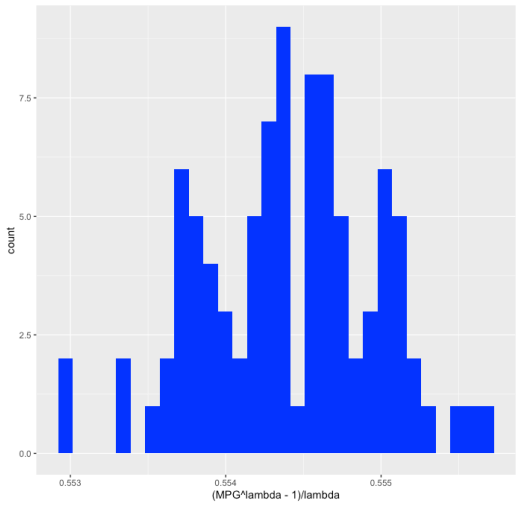
Residual plots of the initial model

After fitting all the predictors to the initial linear model, surprisingly, the adjusted R-squared value is 0.795 which is very high. However, in order to prevent the overfitting and improve the explainability, I still have to select the most significant predictors. Also, according to the normal Q-Q plot and residual plots, they indicate that there are departures from the normality. The right tail of normal Q-Q plot deviates the 45-degree line. Also, the residual curve from the Residuals vs Fitted values plot is slightly U-shaped, therefore, there might be a non-linear relationship in the model.

By the Box-Cox transformation, I found that the adjusted R-squared value of the transformed model is improved to 0.889 when we set λ to be -1.798. Furthermore, the normal Q-Q plot and residual plots also showed that the new model conforms to the normality and has linear relationship between the response variable and predictors. Also, the histogram of the transformed average MPG is basically bell-shaped.



Residual plots of the transformed model

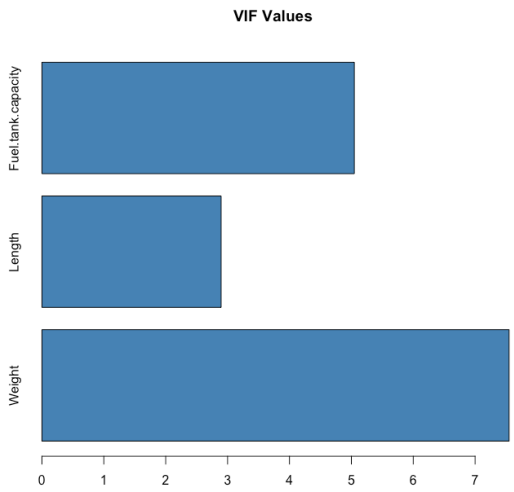


Histogram of the transformed avg MPG

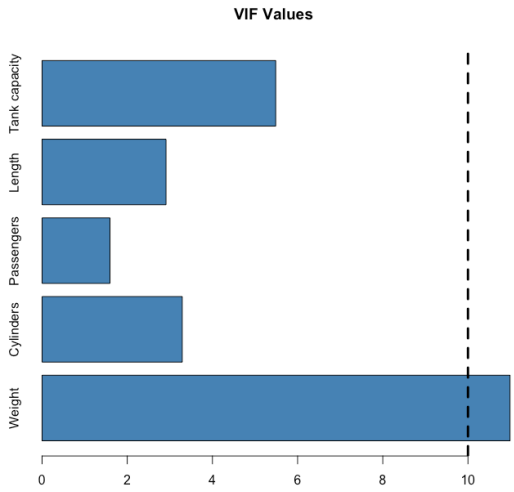
- Stepwise regression

I applied the backward elimination and forward selection to the initial model and used AIC and BIC to provide measures of model performance that account for goodness of fit and model complexity. Hence, there will be 4 different ways of selection process. The plots in the appendix are the models after stepwise regression. As we can see, the Forward and Backward AIC, which are actually identical, have better adjusted R-squared values (0.894), however, not all the variables are significant to the response variable. On the other hand, even if the Forward and Backward BIC models have lower adjusted R-squared values (0.848 and 0.871), all of their predictors are significant.

- Test for multicollinearity



VIF plot of Forward BIC



VIF plot of Backward BIC

However, the problem of multicollinearity is still needed to be addressed. I applied to the Variance Inflation Factor (VIF) to measure how much the variance of an estimated regression coefficient is increased because of collinearity. Since the VIF test is not well-defined for categorical variables, I only used it for the numeric ones. Luckily, all the VIF values of the Forward BIC model are below 10 and the largest one is about 7.5. It indicates that the problem of multicollinearity of it did not pose a strong influence on our analysis. Consequently, I decided to use the Forward BIC model as my final fitted model which has “Weight”, “Length”, “Fuel.tank.capacity” and “Origin” as predictors for predicting the fuel consumption of the cars.

3. Appendix

```
lm(formula = MPG ~ . - MPG.highway - MPG.city, data = data_cont)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4222 -1.2830  0.0654  1.1995  8.1688

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.5047181 18.6413091   0.349 0.728228
TypeLarge         0.4419386  1.7528912   0.252 0.801720
TypeMidsize      -1.1991643  1.1572311  -1.036 0.303818
TypeSmall         1.9149406  1.2933823   1.481 0.143409
TypeSporty       -2.5213981  1.4850689  -1.698 0.094180 .
TypeVan           2.0555993  2.5857366   0.795 0.429435
Price            -0.0385462  0.0656998  -0.587 0.559375
AirBagsDriver only  0.1795555  0.8411349   0.213 0.831610
AirBagsNone      -0.7720304  1.1365039  -0.679 0.499286
DriveTrainFront  -0.3575388  1.0859650  -0.329 0.743004
DriveTrainRear   -1.5744602  1.5056147  -1.046 0.299446
Cylinders        -0.8747673  0.6245565  -1.401 0.165942
EngineSize       -0.0636694  1.2794157  -0.050 0.960458
Horsepower        0.0104353  0.0242204   0.431 0.667961
RPM              -0.0007057  0.0011515  -0.613 0.542015
Rev.per.mile      0.0018911  0.0010746   1.760 0.083012 .
Man.trans.availYes -2.7515638  1.0015435  -2.747 0.007711 **
Fuel.tank.capacity -0.3570104  0.2418774  -1.476 0.144629
Passengers       -2.6457378  0.7633030  -3.466 0.000926 ***
Length            0.0270184  0.0547384   0.494 0.623208
Wheelbase         0.2366973  0.1278408   1.851 0.068505 .
Width             0.5367665  0.2783786   1.928 0.058071 .
Turn.circle      -0.0471006  0.1843297  -0.256 0.799101
Weight           -0.0076048  0.0023538  -3.231 0.001915 **
Originnon-USA     1.8522754  0.7782000   2.380 0.020155 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.449 on 67 degrees of freedom
Multiple R-squared:  0.8488,    Adjusted R-squared:  0.7947
F-statistic: 15.68 on 24 and 67 DF,  p-value: < 2.2e-16
```

Initial linear model

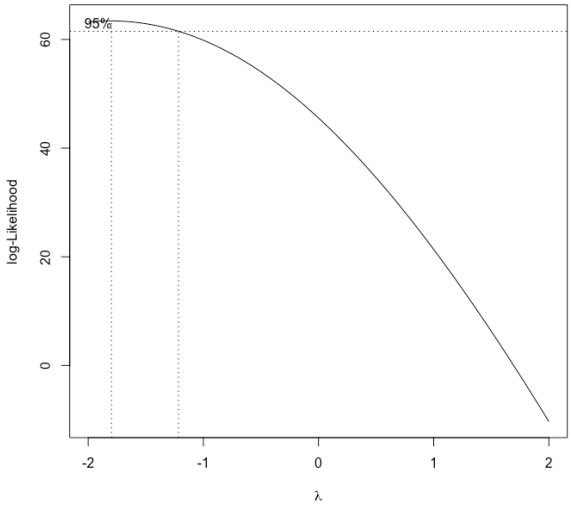
```
lm(formula = ((MPG^lambda - 1)/lambda) ~ . - MPG.highway - MPG.city,
    data = data_cont)

Residuals:
       Min         1Q     Median         3Q        Max
-3.800e-04 -8.884e-05 -2.171e-05  1.296e-04  4.678e-04

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.552e-01  1.402e-03 396.071  < 2e-16 ***
TypeLarge          6.156e-05  1.318e-04   0.467  0.64199
TypeMidsize       -1.202e-04  8.702e-05  -1.381  0.17188
TypeSmall          8.244e-05  9.726e-05   0.848  0.39964
TypeSporty        -1.947e-04  1.117e-04  -1.743  0.08590 .
TypeVan           -1.413e-04  1.944e-04  -0.727  0.47001
Price             -7.201e-06  4.940e-06  -1.458  0.14964
AirBagsDriver only -5.032e-05  6.325e-05  -0.796  0.42911
AirBagsNone       -1.187e-04  8.546e-05  -1.389  0.16947
DriveTrainFront    6.642e-05  8.166e-05   0.813  0.41889
DriveTrainRear    -4.484e-05  1.132e-04  -0.396  0.69331
Cylinders         -1.600e-04  4.697e-05  -3.406  0.00112 **
EngineSize        -9.994e-05  9.621e-05  -1.039  0.30263
Horsepower         1.661e-06  1.821e-06   0.912  0.36498
RPM               -9.091e-08  8.659e-08  -1.050  0.29754
Rev.per.mile      -1.830e-08  8.081e-08  -0.226  0.82157
Man.trans.availYes -2.490e-04  7.531e-05  -3.306  0.00153 **
Fuel.tank.capacity -3.778e-05  1.819e-05  -2.077  0.04165 *
Passengers        -1.977e-04  5.740e-05  -3.445  0.00099 ***
Length             7.263e-06  4.116e-06   1.764  0.08221 .
Wheelbase          6.983e-06  9.613e-06   0.726  0.47011
Width              3.264e-05  2.093e-05   1.559  0.12363
Turn.circle       -1.689e-05  1.386e-05  -1.219  0.22726
Weight            -4.839e-07  1.770e-07  -2.734  0.00800 **
Originnon-USA      1.333e-04  5.852e-05   2.278  0.02595 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001842 on 67 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8894
F-statistic: 31.51 on 24 and 67 DF,  p-value: < 2.2e-16
```

Transformed linear model’



Box-Cox λ and log-likelihood plot

```
lm(formula = ((MPG^lambda - 1)/lambda) ~ Type + Cylinders + EngineSize +
    Man.trans.avail + Fuel.tank.capacity + Passengers + Length +
    Width + Weight + Origin, data = data_cont)

Residuals:
       Min         1Q     Median         3Q        Max
-4.392e-04 -9.482e-05  5.710e-06  9.258e-05  4.830e-04

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.538e-01  8.423e-04 657.511  < 2e-16 ***
TypeLarge           7.046e-05  1.134e-04   0.621 0.536397
TypeMidsize        -1.344e-04  7.897e-05  -1.702 0.092720 .
TypeSmall           2.545e-05  8.372e-05   0.304 0.761938
TypeSporty         -2.519e-04  9.877e-05  -2.550 0.012751 *
TypeVan            -2.443e-04  1.551e-04  -1.575 0.119312
Cylinders          -1.460e-04  3.990e-05  -3.659 0.000461 ***
EngineSize         -1.050e-04  6.288e-05  -1.670 0.098980 .
Man.trans.availYes -2.218e-04  7.107e-05  -3.121 0.002536 **
Fuel.tank.capacity -4.860e-05  1.570e-05  -3.095 0.002746 **
Passengers         -1.713e-04  4.988e-05  -3.435 0.000959 ***
Length              4.830e-06  3.607e-06   1.339 0.184586
Width               5.195e-05  1.687e-05   3.079 0.002875 **
Weight             -4.430e-07  1.301e-07  -3.405 0.001055 **
Originnon-USA       1.238e-04  5.194e-05   2.383 0.019638 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001803 on 77 degrees of freedom
Multiple R-squared:  0.9103,    Adjusted R-squared:  0.894
F-statistic: 55.83 on 14 and 77 DF,  p-value: < 2.2e-16
```

Linear model of Backward AIC

```
lm(formula = ((MPG^lambda - 1)/lambda) ~ Cylinders + Man.trans.avail +
    Fuel.tank.capacity + Passengers + Length + Weight + Origin,
    data = data_cont)

Residuals:
       Min         1Q     Median         3Q        Max
-5.600e-04 -1.271e-04  5.860e-06  1.498e-04  5.089e-04

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         5.557e-01  4.209e-04 1320.339  < 2e-16 ***
Cylinders          -1.194e-04  3.533e-05   -3.379 0.001104 **
Man.trans.availYes -1.853e-04  6.840e-05   -2.710 0.008163 **
Fuel.tank.capacity -5.638e-05  1.518e-05   -3.715 0.000366 ***
Passengers         -1.041e-04  2.975e-05   -3.499 0.000750 ***
Length              1.324e-05  2.642e-06    5.010 2.98e-06 ***
Weight             -5.883e-07  1.178e-07   -4.994 3.17e-06 ***
Originnon-USA       1.094e-04  4.801e-05    2.278 0.025274 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000199 on 84 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.871
F-statistic: 88.74 on 7 and 84 DF,  p-value: < 2.2e-16
```

Linear model of Backward BIC

```
lm(formula = ((MPG^lambda - 1)/lambda) ~ Weight + Length + Fuel.tank.capacity +
    Origin + Cylinders + Type + Width + Passengers + Man.trans.avail +
    EngineSize, data = data_cont)

Residuals:
       Min         1Q     Median         3Q        Max
-4.392e-04 -9.482e-05  5.710e-06  9.258e-05  4.830e-04

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.538e-01  8.423e-04 657.511  < 2e-16 ***
Weight             -4.430e-07  1.301e-07  -3.405 0.001055 **
Length              4.830e-06  3.607e-06   1.339 0.184586
Fuel.tank.capacity -4.860e-05  1.570e-05  -3.095 0.002746 **
Originnon-USA       1.238e-04  5.194e-05   2.383 0.019638 *
Cylinders          -1.460e-04  3.990e-05  -3.659 0.000461 ***
TypeLarge           7.046e-05  1.134e-04   0.621 0.536397
TypeMidsize        -1.344e-04  7.897e-05  -1.702 0.092720 .
TypeSmall           2.545e-05  8.372e-05   0.304 0.761938
TypeSporty         -2.519e-04  9.877e-05  -2.550 0.012751 *
TypeVan            -2.443e-04  1.551e-04  -1.575 0.119312
Width               5.195e-05  1.687e-05   3.079 0.002875 **
Passengers         -1.713e-04  4.988e-05  -3.435 0.000959 ***
Man.trans.availYes -2.218e-04  7.107e-05  -3.121 0.002536 **
EngineSize         -1.050e-04  6.288e-05  -1.670 0.098980 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001803 on 77 degrees of freedom
Multiple R-squared:  0.9103,    Adjusted R-squared:  0.894
F-statistic: 55.83 on 14 and 77 DF,  p-value: < 2.2e-16
```

Linear model of Forward AIC

```
lm(formula = ((MPG^lambda - 1)/lambda) ~ Weight + Length + Fuel.tank.capacity +
    Origin, data = data_cont)

Residuals:
       Min         1Q     Median         3Q        Max
-5.238e-04 -1.362e-04 -3.040e-06  1.441e-04  7.979e-04

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         5.551e-01  3.787e-04 1465.564  < 2e-16 ***
Weight             -7.853e-07  1.067e-07   -7.361 9.63e-11 ***
Length              1.514e-05  2.777e-06    5.453 4.57e-07 ***
Fuel.tank.capacity -6.346e-05  1.624e-05   -3.908 0.000184 ***
Originnon-USA       1.238e-04  4.903e-05    2.524 0.013411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000216 on 87 degrees of freedom
Multiple R-squared:  0.8546,    Adjusted R-squared:  0.8479
F-statistic: 127.8 on 4 and 87 DF,  p-value: < 2.2e-16
```

Linear model of Forward BIC