FINAL REPORT

Statistical Modeling and Analysis Results for the City-cycle Fuel Consumption

(Math 161B Final Project)

Submitted to:

Dr. Andrea Gottlieb

Department of Mathematics and Statistics

San Jose State University

Report Prepared By:

Thien Tran

Jason Kong

Jin Qian

Kayley Buell

May 21, 2021

Table of Content

Executive Summary

We are Group J and this report aims to analyze city-cycle fuel consumption. City-cycle fuel consumption is a concern for many car owners. Although the average city-cycle fuel consumption can be found based on the information given to us by car manufacturers, we want to dive deeper into the details of the fuel consumption. Before that, we decided to answer two statistical questions. What variables of a car effectively determine the city-cycle fuel consumption (in miles per gallon)? Do the categorical variables affect the response from our dataset? First we adjusted the dataset to better address these questions and have removed the "car names" variable because we believe it is not necessary. Instead what we have done is replaced it with a variable called "brand" which will later be explained in the report. In the end, the modified dataset has 1 response "mpg" and 8 predictors with 392 observations.

Section 1 presents an introduction of the dataset we were working on. It gives specific details about where we got the dataset, how many observations it has, how many variables it contains as well as how each variable is represented (data types).

Section 2 presents the summary statistics of each variable and how we explored the dataset in the first look. Before we went ahead with applying a statistical model, we looked at the distributions of the variables. The variables "cylinder", "origin", and "brand" are categorical and the distributions of these are determined by the factor levels of each individual variable. Cylinder has 5 factor levels distinguishing from 3-cylinders, 4-cylinders, 5-cylinders, 6-cylinders, and 8-cylinders with the most common being 4-cylinders. Origin has 3 factor levels which are distinguished by 1, 2, and 3 which represent regions of the world. Brands have 13 factor levels where each individual factor level represents a car brand. Overall the distributions of these categorical variables are as expected with only 392 observations. The distributions of the continuous variables tell a different story. There are four continuous variables, "displacement", "horse power", "weight", and "acceleration". Looking at the distributions of these variables, 3 out of the 4 variables have a skewed distribution. In particular, "displacement", "horse power", and "weight" have a slight skew to the right. "Acceleration" however has a normal distribution.

Section 3 presents the analysis we did to answer our first statistical question: "Can we build a linear regression model to predict city-cycle fuel consumption based on other predictors?" The first model we decided to use is the linear regression model. For our preliminary step, we treated the variable "model year" as a continuous variable. After, we

received an R^2 of 0.84 and the p-value for the acceleration was 0.78 which was peculiar. The QQ plot for the studentized residuals of predicted mpg is a normal quantile plot and the normality assumption is fulfilled. The residuals plot of the predicted mpg is more of a triangle shape which needed further investigation. Then we tried to fit a new model with "model year" as a categorical variable and got an R^2 of 0.86 with the p-value of the acceleration variable being 0.9669. The consensus was to keep "model year" as a continuous variable for the next linear regression model. After applying a transformation for the response mpg, we got an R^2 of 0.89 and the QQ plots for both studentized residuals and residuals plot look relatively normal.

Section 4 presents the analysis for our second statistical question about how categorical variables effect on the fuel consumption and whether they affect each other on the mean mpg. The model we fitted the dataset with is the two-factor ANOVA model. We first looked at the interaction plots of the variables and discovered that the variables "model year" and "origin" are the effects of the model. Then we ran a hypothesis test with null hypotheses "There is no model year effect", "There is no origin effect", and "The model year does not depend on the origin". For both of the factors, we received a p-value of less than $0.0001 < 0.05$, which means there is an effect from these two factors. The p-value for the interaction effect is $0.4292 < 0.05$ and conclude that there is no interaction effect between "model year" and "origin".

Section 5 presents our results and conclusion after we analyzed the dataset.

Section 6 presents what we would do if we have more time to work on this dataset.

1. Introduction.

This report uses the Auto-MPG dataset from the StaLib library maintained at Carnegie Mellon University on July 7, 1993 to build a prediction model for city-cycle fuel consumption. The main purpose is to analyze the dataset and document the implemented model along with all corresponding statistical analyses, results as well as techniques used to obtain the final model.

The Auto-MPG dataset was used in the 1983 American Statistical Association Exposition. It is not the original version provided in the StaLib library but a slightly modified version. We took the dataset from the UCI Machine Learning Repository. The relevant information about this dataset is introduced specifically:

> "In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original".
> "The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)"".

There are total 398 observations with 9 attributes (Table 1.1):

Table 1.1 Original Dataset Information

Attribute	Data Type	Variable Type
mpg	continuous	**quantitative response**
cylinders	multi-valued discrete	categorical
displacement	continuous	quantitative
horsepower	continuous	quantitative
weight	continuous	quantitative
acceleration	continuous	quantitative
model year	multi-valued discrete	quantitative / categorical
origin	multi-valued discrete	categorical
car name	string (unique for each observation)	categorical

In this dataset, the population is all cars with model years from 1970 to 1982. Each car is an observational unit and we have a total of 398 units. Even though all variables are numeric, 2

of them are categorical variables: “cylinders” and “origin” (nominal since they don’t have meaningfully natural order). We will first look at distributions of them to get some senses.

Before deeply diving into the process of building a good model for fuel consumption by the Auto-MPG dataset, we realized that the last variable “car name” is not meaningful at all for building a linear regression model since it’s unique for each observation and the name of a car obviously does not have any effect on fuel consumption. We decided to remove the last attribute column. Furthermore, we use the “car name” column to make a new column named “brand” which contains the brand names for each car. Now, instead of a “car name” column with no meaning, we had a new column “brand” (a new categorical variables) with 13 levels. We would use this column for an ANOVA model later to address the question whether the mpg means are different between different car brands.

In addition, we also figured out that there are 6 observations with missing data (for “horsepower” attribute). Therefore, we determined to remove those observations out of the original dataset to obtain a new dataset for addressing our statistical question: “Can we predict the city-cycle consumption based on 7 predictors: number of cylinders, displacement, horsepower, weight, model year and origin?

In short, the final dataset we are working on containing 392 observations with 8 variables: 1 quantitative response (“mpg”) and other 7 predictors. The question about whether to treat “model year” as a continuous variable or categorical variable will be addressed later in Section 3 when we actually tried two ways.

Table 1.2 Final Dataset Information

Attribute	Data Type	Variable Type
mpg	continuous	**continuous response**
cylinders	multi-valued discrete	categorical
displacement	continuous	continuous
horsepower	continuous	continuous
weight	continuous	continuous
acceleration	continuous	continuous
model year	multi-valued discrete	continuous / categorical
origin	multi-valued discrete	categorical
brand	string	categorical

2. Summary statistics.

With 392 observations and 9 variables, we at first explored each variable in the dataset.

Look at the histograms, we can see that, “cylinders” has 5 levels: 3, 4, 5, 6 and 8 with approximately half of the data are 4-cylinder cars. In addition, “origin” has 3 levels (which represent for 3 different regions in the world but the dataset doesn’t include specific additional information about it): 1, 2 and 3 with approximately 60% of cars are from region 1. The last categorical variable is “brand” with 13 levels.

For continuous variables, we looked at both their distributions and summary statistics. Even though we realized that each variable has its own range, we did not normalize to make all predictors falling in the same range from (0, 1) because it’s not a must-to-do for a linear regression model. We by then just go ahead and fit the model in JMP.

The general look at each distribution gave us a conclusion that data for “displacement”, “weight” and “horsepower” are slightly skewed right, whereas that of “acceleration” is approximately normal.

Figure 2.1 displays the distributions of 2 categorical variables: “cylinders” and “origin”.

Figure 2.2 displays the distribution of categorical variable “brand”.

Figure 2.3 summarizes statistics of 4 continuous variables.

Figure 2.4 displays the distributions of 4 continuous variables.

Figure 2.1 Distributions of “cylinders” and “origin” variables

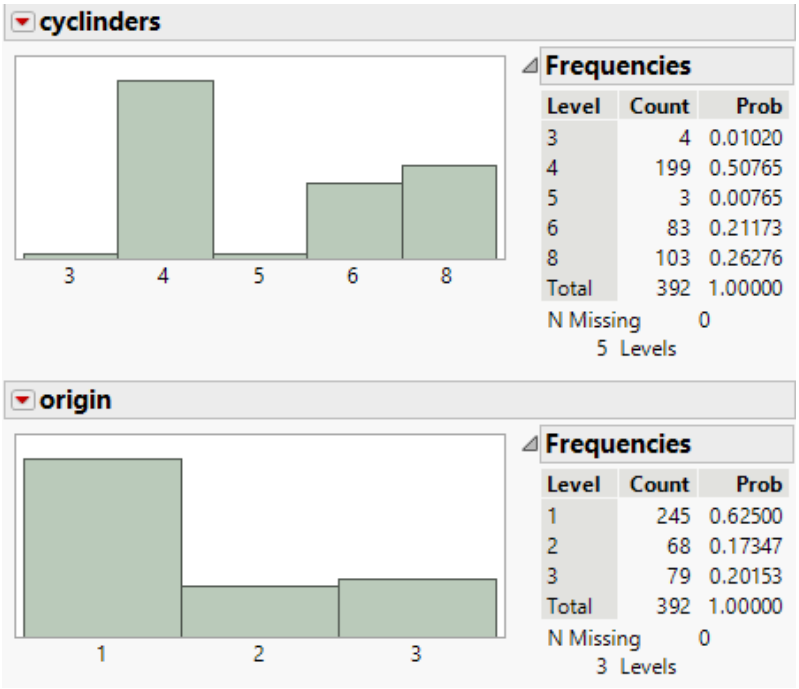


Figure 2.2 Distribution of “brand”

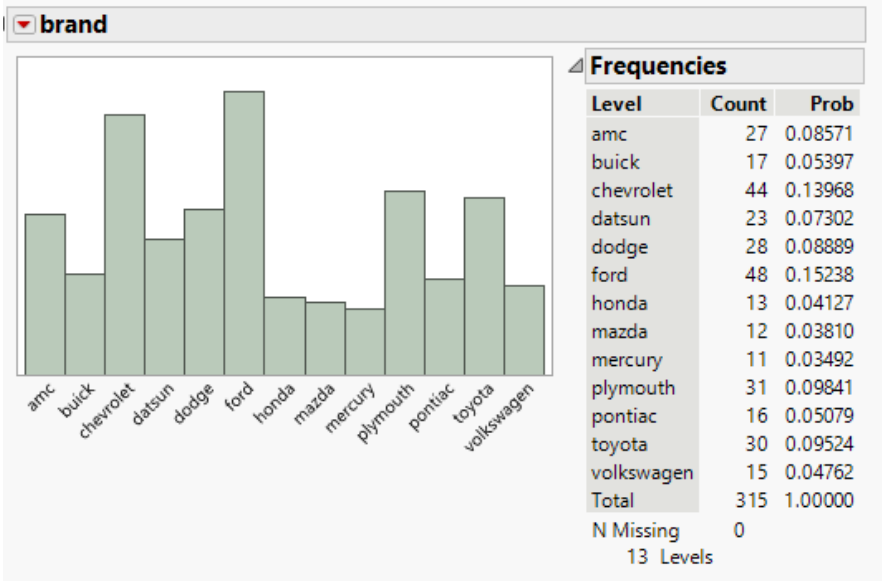
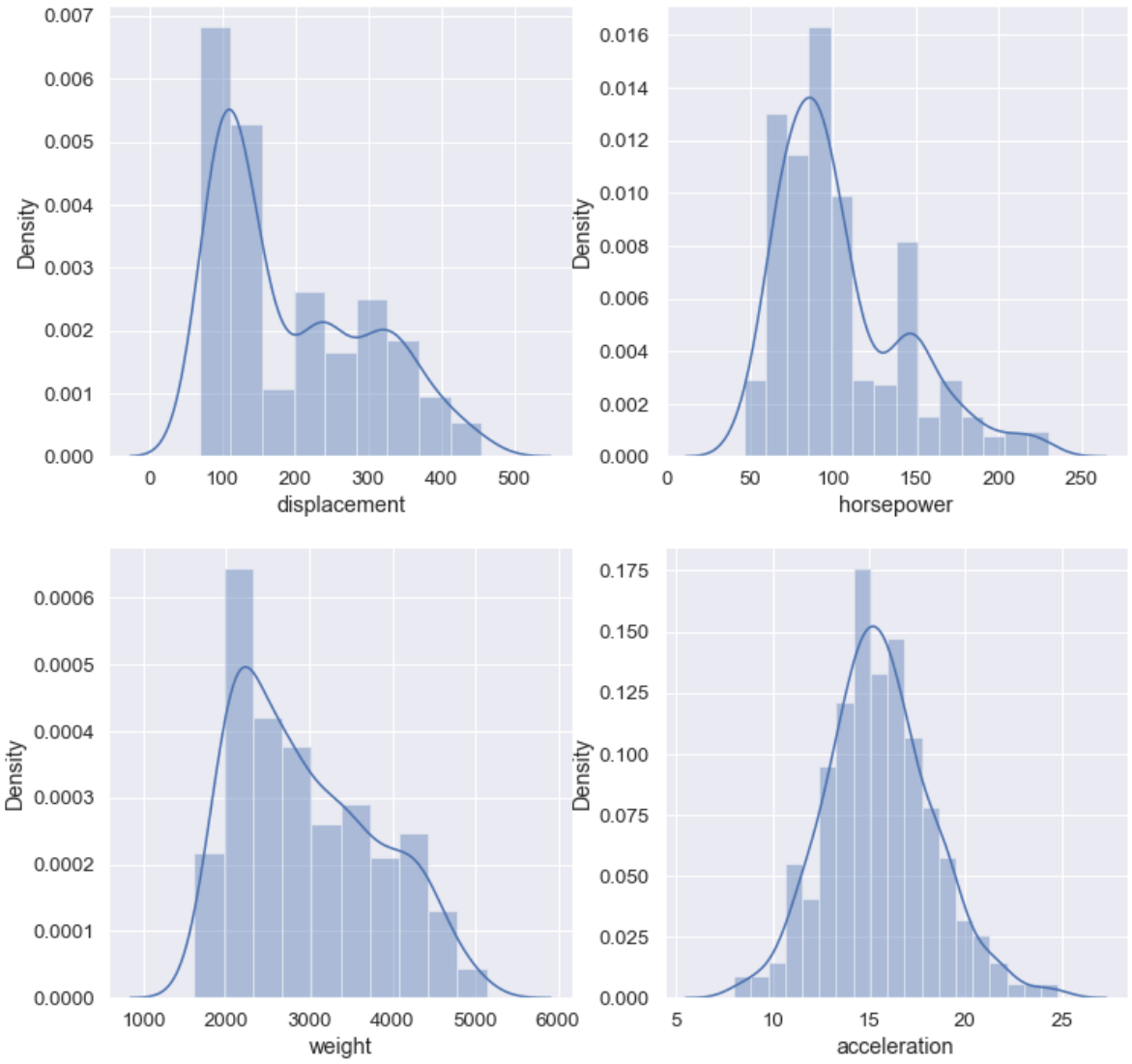


Figure 2.3 Summary statistics for 4 continuous variables

	displacement	horsepower	weight	acceleration
count	392.000000	392.000000	392.000000	392.000000
mean	194.411990	104.469388	2977.584184	15.541327
std	104.644004	38.491160	849.402560	2.758864
min	68.000000	46.000000	1613.000000	8.000000
25%	105.000000	75.000000	2225.250000	13.775000
50%	151.000000	93.500000	2803.500000	15.500000
75%	275.750000	126.000000	3614.750000	17.025000
max	455.000000	230.000000	5140.000000	24.800000

Figure 2.4 Distributions for continuous variables



3. Linear Regression Analysis.

3.1 Fit a full model (treat "model year" as a continuous variable).

We want to predict the city-cycle fuel consumption based on 7 attributes: number of cylinders, displacement, horsepower, weight, acceleration, model year and origin. The model we choose to fit here is a multiple linear regression model.

First, treat "model year" as a continuous variable to make a simple model and see if it represents well for the dataset. It's plausible to that since "model year" is numeric. We tried to fit a full model with all predictors and without interaction terms:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \varepsilon$$

Model interpretation is in Table 3.1.1, 3.1.2 and 3.1.3

Model assumption: error terms ε_i's are independent to each other and normally distributed with mean 0 and constant variance.

We got a linear regression model with the R^2 adjusted is 0.84 (in Figure). There are a couple of comments we drew from the results table. When we look at the "Effect Tests" table (Figure), we see that the p-value for the slope of "acceleration" is 0.78, which is higher than our significant level ($\alpha = 0.05$). This means that β_4 is not significantly different from 0. We took a note here and would try to fit a new model later without "acceleration" prediction to see if the model performs worse or not.

After that, we looked at two plots: the QQ-plot of studentized residuals and the residuals plot by predicted mpg. In the QQ-Plot of studentized residuals (Figure), most of the dots fall on the 45 degree line. There are a few dots have bigger variance (up right corner), but in general, this is a good normal quantile plot (for a not-small dataset with number of observations is 392). We all agreed that the normality assumption is reasonably satisfied here. However, the residuals plot (Figure) by predicted mpg is not good. It looks like a triangular shape, which implies that the variance becomes bigger as the response variable is bigger, or the constant variance assumption here for linear regression model is violated. We decided that applying a transformation on fuel consumption is reasonable option here.

Table 3.1.1 Model interpretation

Variable	Interpretation
$\boldsymbol{Y}$	City-cycle fuel consumption (in mpg)
X_1	Piston displacement (in cc)
X_2	Horsepower (in hp)
X_3	Weight (in lbs)
X_4	Acceleration (in mph)
X_5	Model year
$X_6 - X_9$	Dummy variables for cylinders (defined in Table 3.1.2)
$X_{10} - X_{11}$	Dummy variables for origin (defined in Table 3.1.3)

Table 3.1.2

Cylinders	X_6	X_7	X_8	X_9
3	1	0	0	0
4	0	1	0	0
5	0	0	1	0
6	0	0	0	1
8	0	0	0	0

Table 3.1.3

Origin	X_{10}	X_{11}
1	1	0
2	0	1
3	0	0

Figure 3.1.1 Full model results ("model year" as continuous variable)

Summary of Fit

RSquare	0.846916
RSquare Adj	0.842484
Root Mean Square Error	3.09767
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	20172.680	1833.88	191.1175
Error	380	3646.313	9.60	Prob > F
C. Total	391	23818.993		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
cyclinders	4	4	566.4874	14.7591	<.0001*
displacement	1	1	64.3670	6.7080	0.0100*
horsepower	1	1	66.8224	6.9639	0.0087*
weight	1	1	804.0465	83.7936	<.0001*
acceleration	1	1	0.7482	0.0780	0.7802
model year	1	1	2177.3481	226.9120	<.0001*
origin	2	2	243.9769	12.7130	<.0001*

Figure 3.1.2 Residuals plot by predicted response

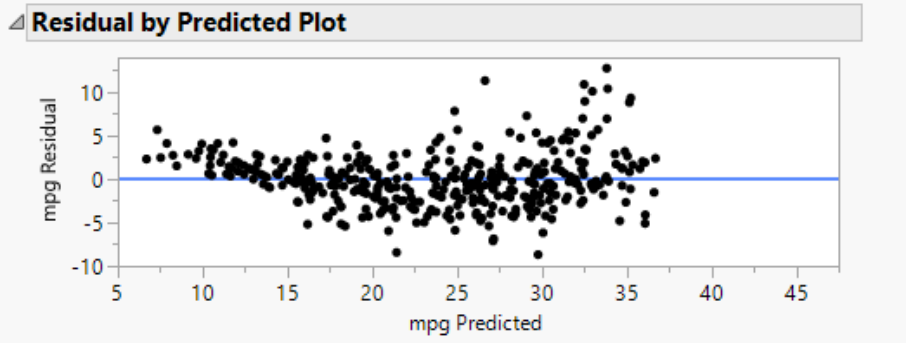
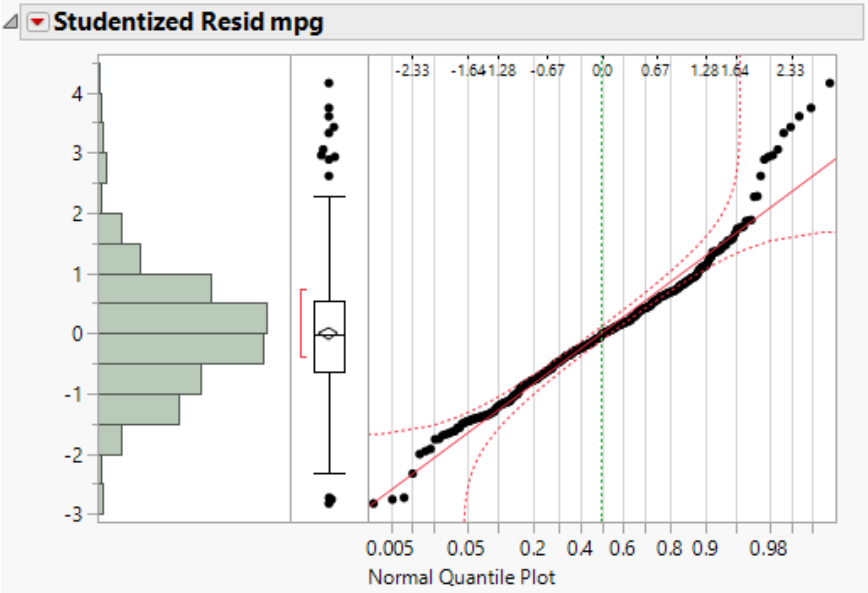


Figure 3.1.3 QQ-Plot of Studentized residuals



3.2 Fit a full model (treat "model year" as a categorical variable).

Next, we treated "model year" as a categorical variable and try to fit a new multiple linear regression model to see if there is a positively huge change for the new model. Since we have one more categorical with 13 levels (1970 to 1982), we need to create 12 dummy variables and this increase number of slopes in the new model by 12 compared to the number of slopes of the full model we did in 3.1.

The resulted R^2 adjusted for this model is 0.86 (Figure 3.2.1), which is slightly bigger than 0.84 of the previous model. Since R^2 does not increase significantly, we tend to gravitate toward considering "model year" as continuous variable since it will result in a simpler model (11 slopes rather than 23 slopes of second model).

Also, when we looked at the residuals plot, it looks pretty much the same as the model in 3.1 (triangular spread), which implies a violation of constant variance assumption. Additionally, p-value for the slope of "displacement" is 0.0818 > 0.05 and that for the slope of "acceleration" is 0.9669 >> 0.05, those two slopes are not significantly different from 0 (Figure 3.2.1).

In short, after trying fitting model with "model year" as categorical variable, we all agreed that treating "model year" as a continuous predictor is much more reasonable for model simplicity.

Figure 3.2.1 Full model results ("model year" as categorical variable)

Summary of Fit

RSquare	0.874383
RSquare Adj	0.866894
Root Mean Square Error	2.847554
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
cyclinders	4	4	472.7706	14.5763	<.0001*
displacement	1	1	24.6961	3.0457	0.0818
horsepower	1	1	73.4463	9.0579	0.0028*
weight	1	1	558.6040	68.8906	<.0001*
acceleration	1	1	0.0140	0.0017	0.9669
model year	12	12	2831.6004	29.1009	<.0001*
origin	2	2	183.2089	11.2972	<.0001*

3.3 Apply a transformation for the response. Fit a new model for ln(mgp).

The results from 3.1 and 3.2 both do not satisfy the constant variance assumption for residuals. Therefore, we decided to apply a transformation for the response (city-cycle fuel consumption) and fit the following multiple linear regression model:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \varepsilon$$

Where Y, X_i's are same interpretation as in Table 3.1.1, Table 3.1.2 and Table 3.1.3. The R^2 adjusted is 0.89 (Figure 3.3.1), which is the highest R^2 we have got so far. This means that 89% variation of the fuel consumption in the Auto-MPG dataset is explained through the resulted linear regression model.

Look at the residuals plot by predicted ln(mpg) (Figure 3.3.2), it looks pretty good when the dots spread in a rectangular region with no discernible pattern. Also, the QQ-plot of studentized residuals plot (Figure 3.3.3) also looks really good. Both plots can help us be confident that all the assumptions for the linear regression model are reasonably satisfied. With the high R^2 adjusted, we are more confident to believe that the linear model (after transformation the response) is absolutely reasonable to help us predict city-cycle fuel consumption based on 7 predictors.

However, we wanted to make the model become as simple as possible. We looked at the "Effect Tests" table (Figure 3.3.4), the p-value for the slope of "acceleration" is 0.2823 > 0.05. It implies that β_4 is not significantly different from 0. We would want to try fitting the last linear regression model by removing "acceleration" predictors to see if R^2 significantly decreases or not.

Figure 3.3.1 R^2 summary for the model of ln(mpg)

Summary of Fit

RSquare	0.893861
RSquare Adj	0.890789
Root Mean Square Error	0.112373
Mean of Response	3.098313
Observations (or Sum Wgts)	392

Figure 3.3.2 Residual plot by predicted ln(mpg)

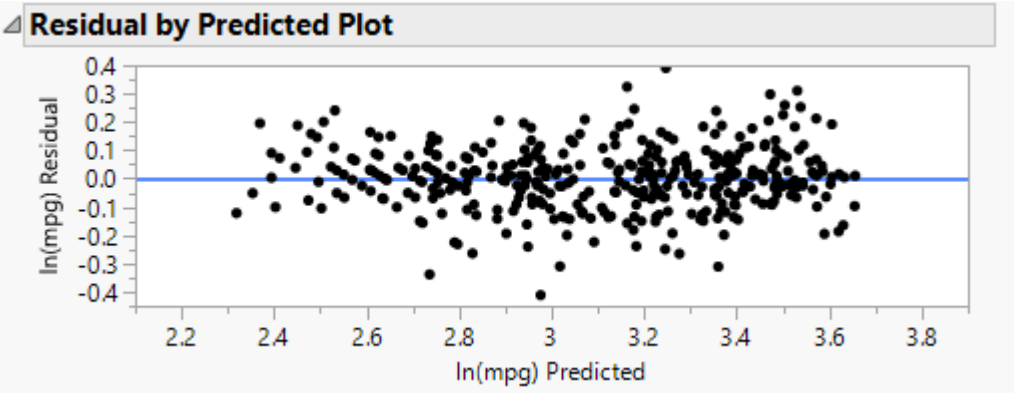


Figure 3.3.3 QQ-Plot for Studentized residuals plot

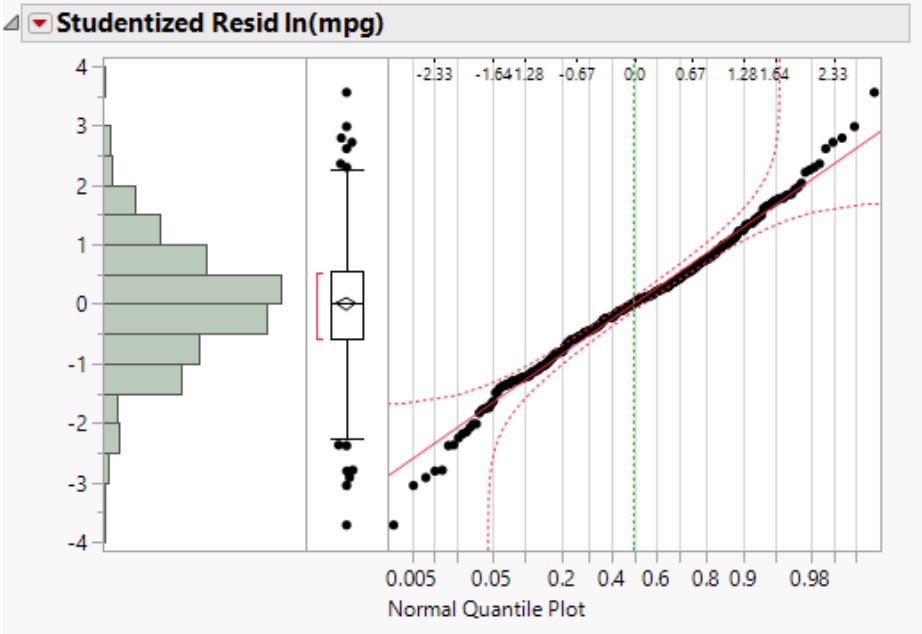


Figure 3.3.4 “Effect Tests” results for the model of ln(mpg)

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
cyclinders	4	4	0.6417894	12.7060	<.0001*
displacement	1	1	0.0676180	5.3547	0.0212*
horsepower	1	1	0.2061584	16.3259	<.0001*
weight	1	1	1.3121552	103.9108	<.0001*
acceleration	1	1	0.0146383	1.1592	0.2823
model year	1	1	3.4144942	270.3970	<.0001*
origin	2	2	0.2218573	8.7845	0.0002*

3.4 Fit the last linear regression model (with 6 predictors).

Having discussed in 3.3, we decided to build the following model:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \varepsilon$$

Model interpretation is in Table 3.4.1, 3.4.2 and 3.4.3

Model assumption: error terms ε_i's are independent to each other and normally distributed with mean 0 and constant variance.

Table 3.4.1 Final model interpretation

Variable	Interpretation
$\boldsymbol{Y}$	City-cycle fuel consumption (in mpg)
X_1	Piston displacement (in cc)
X_2	Horsepower (in hp)
X_3	Weight (in lbs)
X_4	Model year
$X_5 - X_8$	Dummy variables for cylinders (defined in Table 3.4.2)
$X_9 - X_{10}$	Dummy variables for origin (defined in Table 3.4.3)

Table 3.4.2

Cylinders	X_5	X_6	X_7	X_8
3	1	0	0	0
4	0	1	0	0
5	0	0	1	0
6	0	0	0	1
8	0	0	0	0

Table 3.4.3

Origin	X_9	X_{10}
1	1	0
2	0	1
3	0	0

We got a R^2 adjusted is 0.89, the same as the model in Section 3.3. The residuals plot as well as the QQ-plot for Studentized residuals are also similar to the model in Section 3.3. Therefore, all assumptions are satisfied and the final model represents well for the variation of ln(mpg) in the Auto-MPG dataset (89% variation).

4. ANOVA Analysis.

4.1 Single Factor ANOVA Model.

In this section, we are going to analyze whether the mpg mean differs by car brand with the Single ANOVA model. By observing the original data set, the variable "car model" does not have an effective meaning for statistical analysis. Therefore, we modify the "car model" column to a new variable "car brand": we omit the car brands that have a size under ten, and thus we are left with 315 individual observation units and 13 different car brands in this portion. The notation is presenting as the following:

The ANOVA model is:

$$X_{ij} = \mu + a_i + \epsilon_{ij}$$

Where:

μ_1: The mean of the mpg for AMC's car.

μ_2: The mean of the mpg for Buick's car.

μ_3: The mean of the mpg for Chevrolet's car.

μ_4: The mean of the mpg for Datsun's car.

μ_5: The mean of the mpg for Dodge's car.

μ_6: The mean of the mpg for Ford's car.

μ_7: The mean of the mpg for Honda's car.

μ_8: The mean of the mpg for Mazda's car.

μ_9: The mean of the mpg for Mercury's car.

μ_{10}: The mean of the mpg for Plymouth's car.

μ_{11}: The mean of the mpg for Pontiac's car.

μ_{12}: The mean of the mpg for Toyota's car.

μ_{13}: The mean of the mpg for Volkswagen's car.

Model assumption:

ϵ_{ij}'s are independent, normally distributed with mean 0 and variance σ^2.

We have (J_i: sample size of the car in group i^{th}):

$J_1 = 27$ $J_2 = 17$ $J_3 = 44$ $J_4 = 23$

$J_5 = 28$ $J_6 = 48$ $J_7 = 13$ $J_8 = 12$

$J_9 = 11$ $J_{10} = 31$ $J_{11} = 16$ $J_{12} = 30$ $J_{13} = 15$

I = 13 = the number of groups being compared

Then we conduct a hypothesis test to inspect whether there is a car brand effect.

The null hypothesis:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_{13}$: All car brands have the same mean.

The alternative hypothesis:

H_a: At least two of the μ_i's are different.

We got the result as Figure 4.1.1.

$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}) = 18815.919$, $SSTr = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{i.} - \bar{X}) = 7362.923$

$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - X_{i.}) = 11452.996$

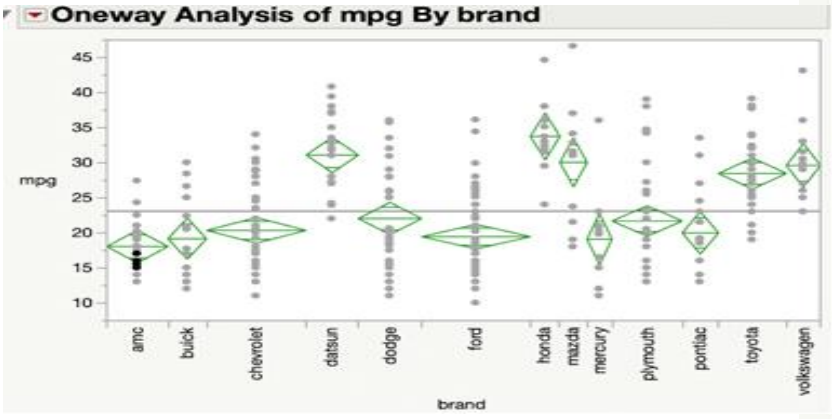
df(SSTr) = I – 1 = 12, df(n – 1) = 315 – 1 = 314

Test statistic: $f = \frac{MSTr}{MSE} = \frac{SSTr}{SSe} = 16.1792$ follow F distribution: $F_{\alpha,df1=11,df2=314}$.

Since p-value is 0.0001 < α = 0.05, we reject H_0.

We conclude that the mean mpg is different by car brand.

Figure 4.1.1 Single Factor ANOVA model on "brand"



Oneway Anova

Summary of Fit

Rsquare	0.391313
Adj Rsquare	0.367127
Root Mean Square Error	6.158233
Mean of Response	23.11143
Observations (or Sum Wgts)	315

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
brand	12	7362.923	613.577	16.1792	<.0001*
Error	302	11452.996	37.924		
C. Total	314	18815.919			

Means for Oneway Anova

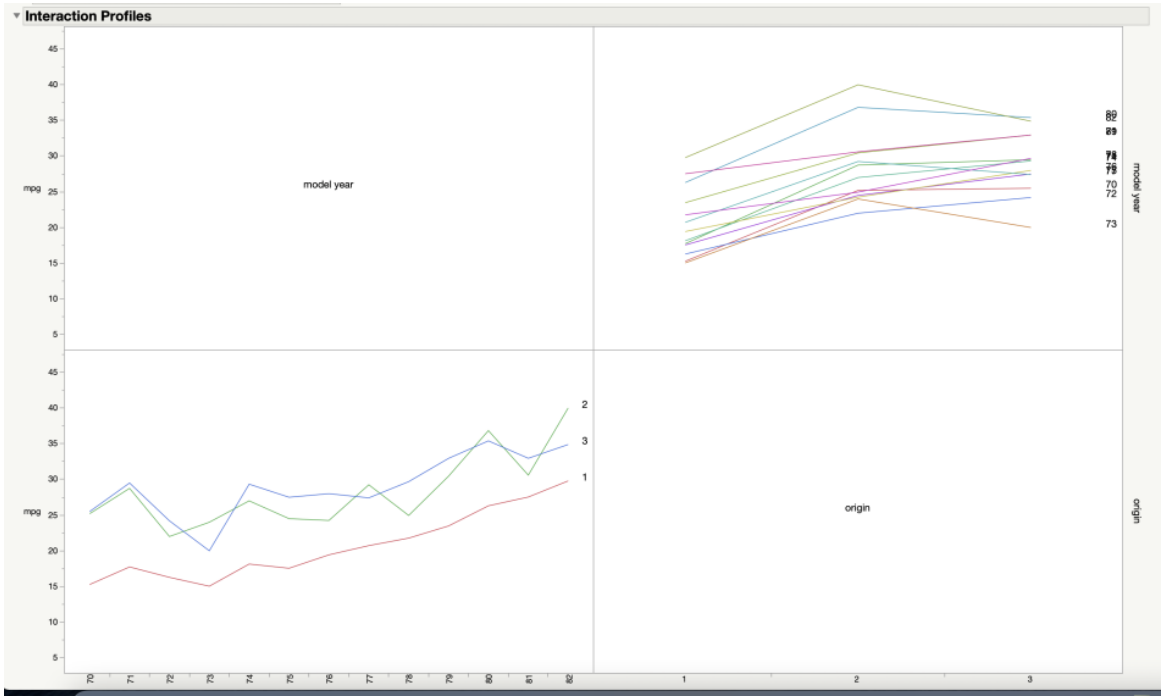
Level	Number	Mean	Std Error	Lower 95%	Upper 95%
amc	27	18.0704	1.1852	15.738	20.403
buick	17	19.1824	1.4936	16.243	22.122
chevrolet	44	20.3705	0.9284	18.544	22.197
datsun	23	31.1130	1.2841	28.586	33.640
dodge	28	22.0607	1.1638	19.771	24.351
ford	48	19.4750	0.8889	17.726	21.224
honda	13	33.7615	1.7080	30.400	37.123
mazda	12	30.0583	1.7777	26.560	33.557
mercury	11	19.1182	1.8568	15.464	22.772
plymouth	31	21.7032	1.1061	19.527	23.880
pontiac	16	20.0125	1.5396	16.983	23.042
toyota	30	28.4800	1.1243	26.267	30.693
volkswagen	15	29.6267	1.5900	26.498	32.756

Std Error uses a pooled estimate of error variance

4.2 Two-Factor ANOVA Model.

Furthermore, we want to test whether the model has an interaction effect between the origin and model year variables. Based on our interaction plot (Figure 4.2.1), we notice that the origin lines and the model year lines are not parallel, which means that there are origin and model year effects.

Figure 4.2.1 Interactions Plot between "model year" and "origin"



Next, we want to conduct a two-factor ANOVA for a hypothesis test to inspect whether the model has an interaction effect between the origin and model year variables.

The two-factor ANOVA model is:

$$X_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Where $i = 1,\ldots,12;\ j = 1,2;\ k = 1,\ldots,K$

Model assumption:

$\epsilon_{ijk}'s$ *are independent, normally distributed with mean 0 and variance* σ^2.

The factor A hypothesis test:

H_{0A}: $\alpha_1 = \ldots = \alpha_{12} = 0$: There is no model year effect.

H_{aA}: At least one $\alpha_i \neq 0$.

The factor B hypothesis test:

H_{0B}: $\beta_1 = \beta_2 = 0$: There is no origin effect.

H_{aB}: At least one $\beta_i \neq 0$.

The interaction hypothesis test:

H_{0AB}: $\gamma_{ij} = 0$ (i=1,…,12; j=1, 2): The model year does not depend on the origin.

H_{aB}: At least one $\gamma_{ij} \neq 0$.

The test statistic for factor A is 16.3926 (Figure 4.2.2) and the p-value is less than 0.0001 < 0.05, concluding that we reject H_{0A}, and so there is a model year effect.

The test statistic for factor B is 104.1169 (Figure 4.2.2) and the p-value is less than 0.0001 < 0.05, we reject H_{0B}, which means there is an origin effect.

Now we want to test whether the model has an interaction effect between the origin and model year. The test statistic for the interaction effect is 1.0276 (Figure 4.2.2) and the p-value is 0.4292 > 0.05, we fail to reject H_{0AB} and so there is no interaction effect between the model year effect and the origin effect.

Figure 4.2.2 Two-Factor ANOVA Table for "model year" and "origin"

Summary of Fit

RSquare	0.659747
RSquare Adj	0.62312
Root Mean Square Error	4.791539
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	38	15714.520	413.540	18.0122
Error	353	8104.473	22.959	Prob > F
C. Total	391	23818.993		<.0001*

Parameter Estimates

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
model year	12	12	4516.2684	16.3926	<.0001*
origin	2	2	4780.8101	104.1169	<.0001*
model year*origin	24	24	566.2211	1.0276	0.4292

5. Results and Conclusion.

5.*1 Fitted model to predict city-cycle consumption.*

In this study, we tried to determine the impact that the horsepower, piston displacement, weight, and the model year of a car had on the predicted city-cycle fuel consumption using a linear regression model first treating model year as a continuous variable and then as a categorical variable. The R2 adjusted value was high but some of our assumptions were disproven in both cases. We then created a transformed linear regression model which had a higher R2 value of 0.89. Moreover, our assumptions were confirmed to be true with this model and almost all of our predictors had a significant relationship with the predicted city-cycle fuel consumption of a car.

Therefore, our final fitted model was:

$$\ln(\hat{Y}) = \beta_0 + 0.0006375355 * X_1 - 0.001636411 * X_2 - 0.000245338 * X_3$$

Where:

- $\hat{Y}$: Predicted city-cycle fuel consumption of a car (in mpg)
- X_1: Piston displacement (in cc)
- X_2: Horsepower (in hp)
- X_3: Weight (in lbs)
- X_4: Model year (the last two digits)

The value of β_0 is specified in the following table (Table 5.1) since it depends on the number of cylinders and the origin of a car:

Table 5.1 The intercept values in the final predicted model.

β_0		*origin*		
		1	*2*	*3*
number of cylinders	*3*	1.42870013	1.48628314	1.5059442
	4	1.67732264	1.73490565	1.75456671
	5	1.70755663	1.76513964	1.7848007
	6	1.55424159	1.6118246	1.63148566
	8	1.56875874	1.62634175	1.64600281

5.2 ANOVA conclusion.

We were able to create a single factor ANOVA model from the data and analyze the effects that the car brand had on the city-cycle fuel consumption. After trimming our data to better fit the model we had. The single factor ANOVA analysis showed an R^2 adjusted value of 0.37 and the p-value being less than 0.0001 told us that the car brands have different mpg values.

Finally, we created a two-factor ANOVA model to test the origin and model year variables and determined that the origin and model year do have effects on the mpg. However, we found with a p-value of 0.4292 that there is no interaction effect between the model year and the origin effect.

6. Recommendation.

Given our final model for the predicted city-cycle fuel consumption, we can say that to decrease this number would be to change the horsepower of the cars being produced, the placement of the piston, and the weight of the car. All of these had significant relationships with the city-cycle fuel consumption. The world is already coming up with ways to decrease this number with alternative methods of transportation and alternative fuel sources. This data is slightly dated so another model could be made to factor in the people who drive electric cars as well as those who choose an alternative mode of transportation.

7. References.

[1] http://archive.ics.uci.edu/ml/datasets/Auto+MPG