**Automobile Statistical Analysis with R**

Introduction: This study is designed to look at data to determine whether are not there is sufficient information to perform analysis. If the data is sufficient, the testing will inform us. If the data is not sufficient this will also be apparent in our analysis. Using R Studio to do this analysis, gives us the ability to use a variety of tests and linear models to provide us with the necessary information to report back to our stakeholders.

Problem Summary: Multiple Linear Regression Analysis

Using multiple linear regression, design a linear model that predicts the mpg of MechaCar prototypes using a number of variables within the MechaCar mpg dataset. Create a separate text file called MechaCarWriteUp.txt. In the text file, provide a small writeup of your interpretation of the multiple linear regression results. Be sure to include the following details:

* Which variables/coefficients provided a non-random amount of variance to the mpg values in the dataset?
* Is the slope of the linear model considered to be zero? Why or why not?
* Does this linear model predict mpg of MechaCar prototypes effectively? Why or why not?

In this analysis, we used R and RStudio to review the information, analyze the data and summarize the results. A review of the sample vehicle data revealed information about the 50 vehicles being studied.

The coding and linear module equation is as follows:

* mecha\_data <- read.csv("MechaCar\_mpg.csv",stringsAsFactors = F,check.names = F)
* mecha\_lm <- lm(mpg ~ vehicle\_length + vehicle\_weight + spoiler\_angle + ground\_clearance + AWD,data=mecha\_data)
* summary(mecha\_lm)

The mecha\_data variable was established to read in the vehicle sample information. This variable will also be utilized to perform our tests.

Results: The mecha\_lm variable holds our linear module information, as listed below:

* coefficients
* residuals
* effects
* rank
* fitted values
* qr
* call
* terms
* model

Those results are displayed as follows:

**Call:**

lm(formula = mpg ~ vehicle\_length + vehicle\_weight + spoiler\_angle +

ground\_clearance + AWD, data = mecha\_data)

**Residuals:**

Min: -19.4701

1Q 4.4994

Median --0.0692

3Q 5.4433

Max 18.5849

**Coefficients:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient Table | Estimate Std. | Error | t value | Pr(>|t|) |
| (Intercept) | -1.04E+02 | 1.59E+01 | -6.559 | 5.08e-08 \*\*\* |
| vehicle\_length | 6.27E+00 | 6.55E-01 | 9.563 | 2.60e-12 \*\*\* |
| vehicle\_weight | 1.25E-03 | 6.89E-04 | 1.807 | 0.0776 . |
| spoiler\_angle | 6.88E-02 | 6.65E-02 | 1.034 | 0.3069 |
| ground\_clearance | 3.55E+00 | 5.41E-01 | 6.551 | 5.21e-08 \*\*\* |
| AWD | -3.41E+00 | 2.54E+00 | -1.346 | 0.1852 |

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

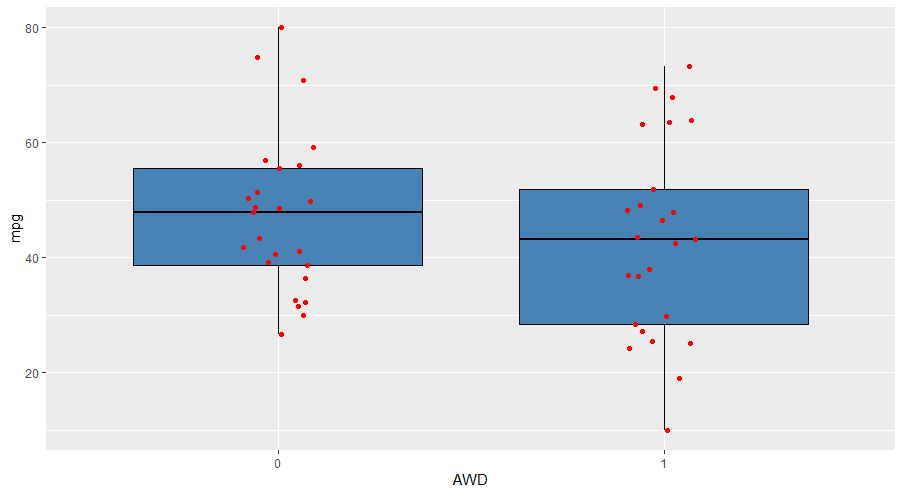
Residual standard error: 8.774 on 44 degrees of freedom

Multiple R-squared: 0.7149, Adjusted R-squared: 0.6825

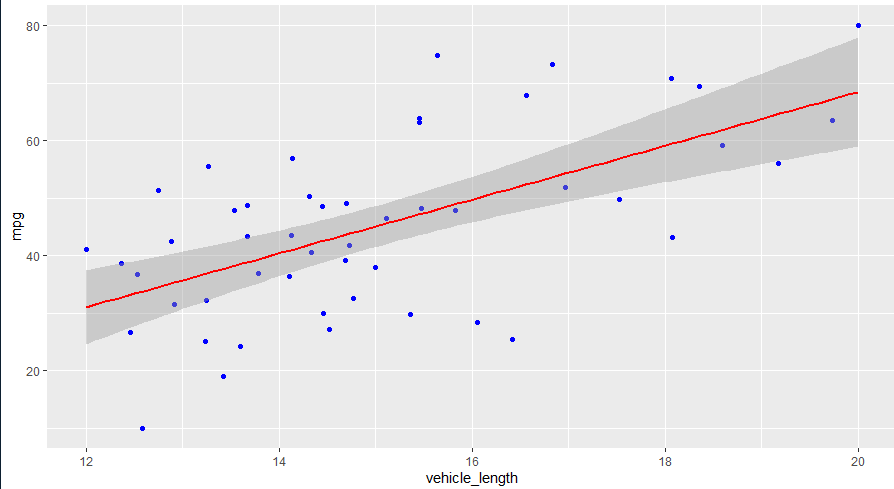
F-statistic: 22.07 on 5 and 44 DF, p-value: 5.35e-11

In viewing these data, I created a few plots to determine whether or not the was enough information about the vehicles to make decisions about the study.

AWD Box & Whisker Plot: The Box and Whisker plot views the Miles Per Gallon (mpg) for cars that have All Wheel Drive (AWD) and those that don’t. What we learn from this analysis is that the majority of the cars fall within 50% of their MPG category, whether they have AWD or not. There are outliers in both areas. The outliers for vehicles with no AWD are greater in 1QR. The outliers for vehicles with AWD are greater in 3QR. <https://github.com/GodIsLove4U/AutosRUs/blob/master/AWD-MPG.PNG>



Scatter Plot: Doing a scatter plot, we can look at our dependent variable, MPG. Then we look at our independent variable, vehicle length. These data tell us that the MPG improves with the length of the vehicle being larger. Smaller vehicles get less MPG

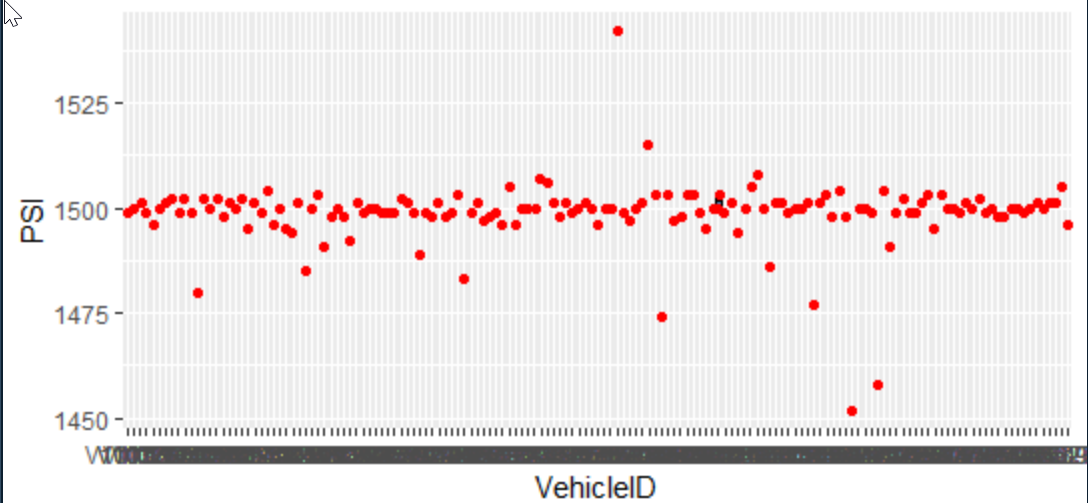


Conclusion

The AWD and vehicle length provide a non-random amount of variance with the MPG values in the dataset. The slope in the linear model indicates a positive slope. The p-value tells us that the date set is not statistically significant and there is strong evidence for the null hypothesis <https://www.graphpad.com/quickcalcs/pvalue1.cfm> and <https://www.khanacademy.org/math/ap-statistics/tests-significance-ap/one-sample-t-test-mean/v/calculating-p-value-from-t-statistic>.

**Suspension Coil Summary**

There is not a significant amount of variation on the PSI for the vehicles analyzed in the suspension coil analysis. There are 150 vehicles, most fall in the 1500 PSI spectrum. There are a few outliers above and below 1500.



Looking at the dplyr summary data, the minimum suspension captures .9796 of our PSI data. Based on our Mecha vehicle data, we determined that the pounds per inch range from 10-63. So, we are well with our weight parameters and criteria for the design specification.

> summary(susp\_sum)

Manufacturing\_Lot Mean

Length:3 Min. :1496

Class :character 1st Qu.:1498

Mode :character Median :1500

Mean :1499

3rd Qu.:1500

Max. :1500

Median Var

Min. :1498 Min. : 0.9796

1st Qu.:1499 1st Qu.: 4.2245

Median :1500 Median : 7.4694

Mean :1500 Mean : 59.5784

3rd Qu.:1500 3rd Qu.: 88.8778

Max. :1500 Max. :170.2861

SD

Min. : 0.9897

1st Qu.: 1.8614

Median : 2.7330

Mean : 5.5907

3rd Qu.: 7.8912

Max. :13.0494

# Total Summary

total\_summary <- susp\_data %>%

summarise(Mean=mean(PSI),Median=median(PSI),Var=var(PSI),SD=sd(PSI))

> summary(tot\_sum)

Mean Median

Min. :1499 Min. :1500

1st Qu.:1499 1st Qu.:1500

Median :1499 Median :1500

Mean :1499 Mean :1500

3rd Qu.:1499 3rd Qu.:1500

Max. :1499 Max. :1500

Var SD

Min. :62.29 Min. :7.893

1st Qu.:62.29 1st Qu.:7.893

Median :62.29 Median :7.893

Mean :62.29 Mean :7.893

3rd Qu.:62.29 3rd Qu.:7.893

Max. :62.29 Max. :7.893

**Suspension Coil T-Test**

Our t-test shows a very high confidence level just as our summary data showed with the suspension coil information we have. What we now know that we did not know before is that even though our pounds per inch fall within the specifications, the sample size of our vehicles informed us of several other pieces of information about the vehicle size, length and miles per gallon.

> t.test(subset(susp\_data2, Manufacturing\_Lot=="Lot1")$PSI,mu = 1500)

One Sample t-test

data: subset(susp\_data2, Manufacturing\_Lot == "Lot1")$PSI

t = 0, df = 49, p-value = 1

alternative hypothesis: true mean is not equal to 1500

95 percent confidence interval:

1499.719 1500.281

sample estimates:

mean of x

1500

> t.test(subset(susp\_data2, Manufacturing\_Lot=="Lot2")$PSI,mu = 1500)

One Sample t-test

data: subset(susp\_data2, Manufacturing\_Lot == "Lot2")$PSI

t = 0.51745, df = 49,

p-value = 0.6072

alternative hypothesis: true mean is not equal to 1500

95 percent confidence interval:

1499.423 1500.977

sample estimates:

mean of x

1500.2

> t.test(subset(susp\_data2, Manufacturing\_Lot=="Lot3")$PSI,mu = 1500)

One Sample t-test

data: subset(susp\_data2, Manufacturing\_Lot == "Lot3")$PSI

t = -2.0916, df = 49,

p-value = 0.04168

alternative hypothesis: true mean is not equal to 1500

95 percent confidence interval:

1492.431 1499.849

sample estimates:

mean of x

1496.14

**Design Your Own Study**

Our study revealed very important and telling information about our vehicles. I will summarize by discussion the following points:

* Vehicle Length

The size of the vehicle revealed that longer had a higher mileage per gallon. This is counter to vehicles that exists today. If MechaCar vehicles produce better mileage with bigger cars, this will greatly affect the market, pushing buyers to buy larger vehicles to save money on gas. This shift in auto buying size will also affect the driving, parking and storing vehicles. Larger vehicles on the road may cause more accidents. It may be more difficult to fin places to put the vehicles due to shrinking availability of space/road real estate. In addition, owners may not be able to fit vehicles in garage or parking spaces in their garage or near their home due to some of the reason aforementioned.

* Vehicle Weight

While vehicle weight meets our threshold criteria in our multiple linear regression model and our t-test, it is not the only factor for determining the feasibility of our vehicle prototypes.

* Miles Per Gallon

The most surprising part of this study is the miles per gallon. While a large group of vehicles were within our 3QR, the trend greatly favored larger vehicles meeting the fuel efficiency criteria.

Using a multiple linear regression and t-test, we can look at very aspects of our data. With the current dataset, either of these test would be useful for the purpose of our study. However, we would have to factor in the intended outcomes and remove any biases or subjective information we know about this market.

We should also collect a larger sample size to ensure we have the information necessary to be confident that our data is telling us what we need to know about our test vehicles.

We should also collect a larger sample size to ensure we have the information necessary to be confident that our data is telling us what we need to know about our test vehicles.

Based on the type of data we have, it appears that several types of statistical tests can be used to perform a meaningful analysis. However, this would also be reliant on several other pieces of information that we should factor in to provide a useful analysis.

Some of the data points we should include:

* Vehicle build location: this information will allow us to do additional analysis about the types of materials used and the import/export considerations that may affect sell price and functionality.
* Selling Lot: Knowing more information about the lots that have vehicle inventory may provide information about price, supply and consumer types.
* Vehicle price: This may drive the cost up or down.
* Resale value: People may decide what they are will to spend based on whether on not the vehicle has a good resale value.
* Maintenance costs: Determine how the cost of maintenance may affect decisions on purchasing and other budgetary considerations.
* Color: Look at the color of vehicles and how they may affect buyer decision, needs or outcomes (tickets, visibility, cleaning).
* Safety: How do vehicles handle on the road? What data will be available about the safety, durability and capability of the vehicle?
* Speed: Look at factors such as how well a vehicle handles on the road, city and highway.
* Cargo: What type of space does the vehicle have. How much cargo will it be able to carry?
* Passengers: The amount of passengers, comfort of seating and leg room are important.

Based on these observations, I think it would be beneficial to do a multiple linear regression analysis with the new data or a chi-squared test. It’s also possible that an ANOVA test would yield useful information, once we obtained the extended data. In conclusion, we have 150 vehicles for our suspension coil test and 50 vehicles for our mecha cars. Where possible, it would be advantageous to have the same sample size. In this analysis it is not likely to be possible, since we only have 50 test type of cars.