# Assignment 2: Monte Carlo Report

STAT8010 – Into to R

Brian Higgins – R00239570

# 1. INTRODUCTION

This report will be a brief outline of questions 3 and 4 for assignment 2 to use Monte Carlo Simulations to predict the average Co2 Emission value using the dataset. We will first fit Linear Regression models on a reduced model supplied by the assignment and a second model using the Akaike Information Criterion or (AIC) to find the best model. Normally we would not have a Y Variable to compare on but in this project, we can compare the average results from running thousands of Monte Carlo Simulations with the actual results of the model to see how Monte Carlo can be used to get very close results.

**Note:** The results for the simulations in this report are an example. If the simulations are run again in the code, there will be different outputs, the general values will be very similar but there will be slight variations. Any discussion is related to the current outputs and is told as a comparison to highlight how well Monte Carlo works.

# 2. MODEL 1: REDUCED LINEAR REGRESSION MODEL

## 2.1. Linear Regression Model

The first Linear Regression model is given by the assignment:

$$CO2{=}\beta0\ {+}\beta1 Eng.size{+}\beta2 Cylinders{+}\beta3 Fuel.conscomb{+}\beta4 Fuel.consmpg + \epsilon$$

We will run this model and remove the Beta Values and store these for later, we will also remove the standard deviation of the residual value as this will be part of our error rate. The error value is in the form of "rnom(0, Var(e), n=length(Data ) )" so we want our residual data to be normally distributed and this will be used to help give us variations in the Monte Carlo simulations as each simulation is run.

## 2.2. Normalisation

Not in the requirements of this assignment, but normalisation was briefly investigated on the residual data. Fig 2.2.1 and 2.2.2 show scatter plots and histograms where the residual data appears to be normally distributed. I also looked into running a Shapiro Wilks test which I have learned in other modules but this can only be run on 5000 data points as it becomes as with larger data there is a higher chance to reject normality. For the first 5000 data points, the data was normal, I will ask about this more next semester.

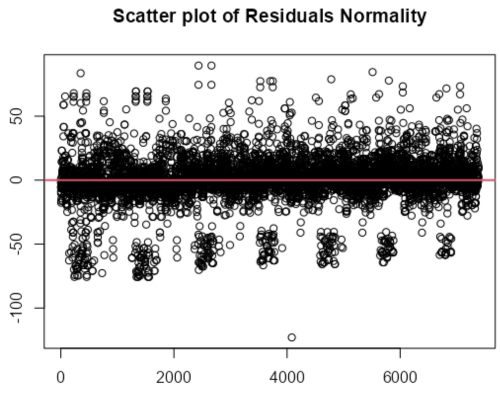


Fig 2.2.1. Scatter Plot

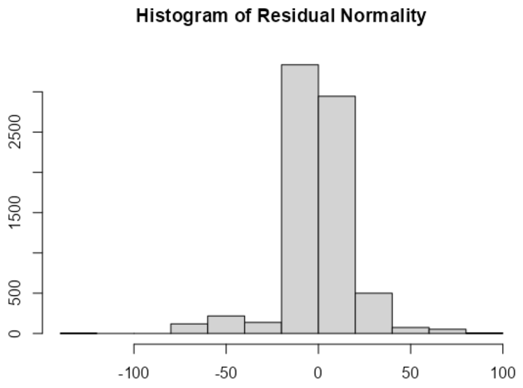


Fig 2.2.2. Histogram

## 2.3. Simulations

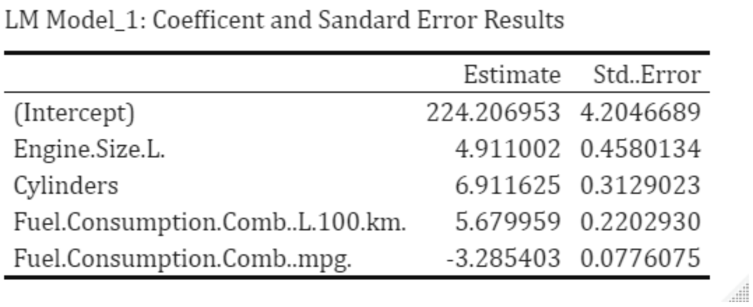
Previously we removed the Beta and Error-values and we will now create a new Estimated Co2 Value by multiplying each of the Model_1 Linear Regression values by the Predicators by the error rate. Once we have this, we will then do a new Linear Regression Model on this and take this for 1,000 or 10,000 simulations. We will then take the mean of the coefficients, standard error and the new Estimated Co2 value.

$$\text{Estimated\_Co2\_Value}=\beta 0+\beta 1*Eng.size+\beta 2*Cylinders+\beta 3*Fuel.conscomb+\beta 4*Fuel.consmpg\ +\ \epsilon$$

## 2.4. Compare results

We now have the mean values of the Coefficients and the Standard Error as mentioned in the assignment report and we can compare these to the original Linear Regression values.

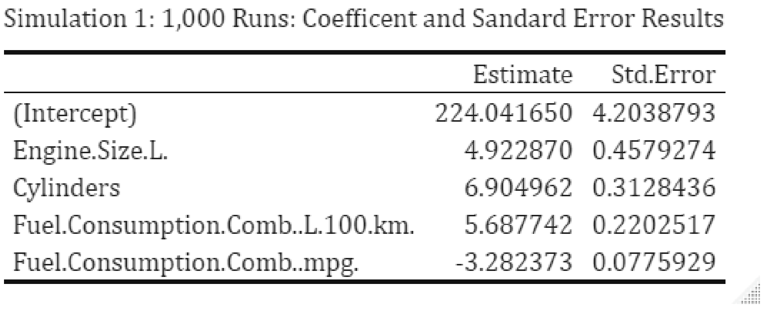
The three tables below in Fig 2.4.1, 2.4.2 and 2.4.3 show the Coefficients of Standard Error for the Linear Regression Model and two simulations that ran 1,000 and 10,000 times. As you can see from both simulations they have gotten very close to the results in the model. In this case, there was not much of a difference between the 1,000 and 10,000 simulations, a couple of the values are a bit closer. We will just look at the 1,000 results for discussion purposes for simplicity.



LM Model_1: Coefficent and Sandard Error Results

| | Estimate | Std..Error |
|---|---|---|
| (Intercept) | 224.206953 | 4.2046689 |
| Engine.Size.L. | 4.911002 | 0.4580134 |
| Cylinders | 6.911625 | 0.3129023 |
| Fuel.Consumption.Comb..L.100.km. | 5.679959 | 0.2202930 |
| Fuel.Consumption.Comb..mpg. | -3.285403 | 0.0776075 |

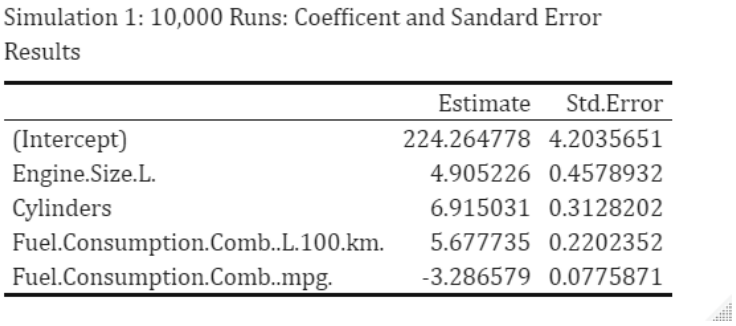
Fig2.4.1 Model 1



Simulation 1: 1,000 Runs: Coefficent and Sandard Error Results

| | Estimate | Std.Error |
|---|---|---|
| (Intercept) | 224.041650 | 4.2038793 |
| Engine.Size.L. | 4.922870 | 0.4579274 |
| Cylinders | 6.904962 | 0.3128436 |
| Fuel.Consumption.Comb..L.100.km. | 5.687742 | 0.2202517 |
| Fuel.Consumption.Comb..mpg. | -3.282373 | 0.0775929 |

Fig 2.4.2. Simulation 10,000



Simulation 1: 10,000 Runs: Coefficent and Sandard Error Results

| | Estimate | Std.Error |
|---|---|---|
| (Intercept) | 224.264778 | 4.2035651 |
| Engine.Size.L. | 4.905226 | 0.4578932 |
| Cylinders | 6.915031 | 0.3128202 |
| Fuel.Consumption.Comb..L.100.km. | 5.677735 | 0.2202352 |
| Fuel.Consumption.Comb..mpg. | -3.286579 | 0.0775871 |

Fig 2.4.3 Simulation 10,000

Next, we will look at the Co2Emissions results for the Linear Regression Model and the 1k Simulation. In Fig 2.4.4 and 2.4.5 you can see the Actual and Simulated Co2 Emission values, the Estimated values look a little more dense around the middle. However, it's in the tables that we see the most interesting result and also shows the power of Monte Carlo Simulations. You can see that the Min value is much lower than the Min value in the actual data, whereas the Max is not as high, however, the Mean value are almost the same. Remember the Estimated Co2 values are the results of 1,000 simulations so it has been averaged out.

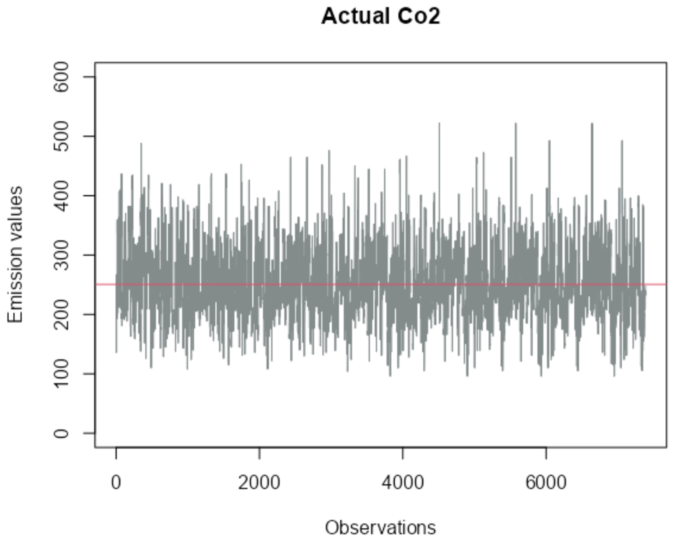


Fig 2.4.4. Actual Co2 Emissions

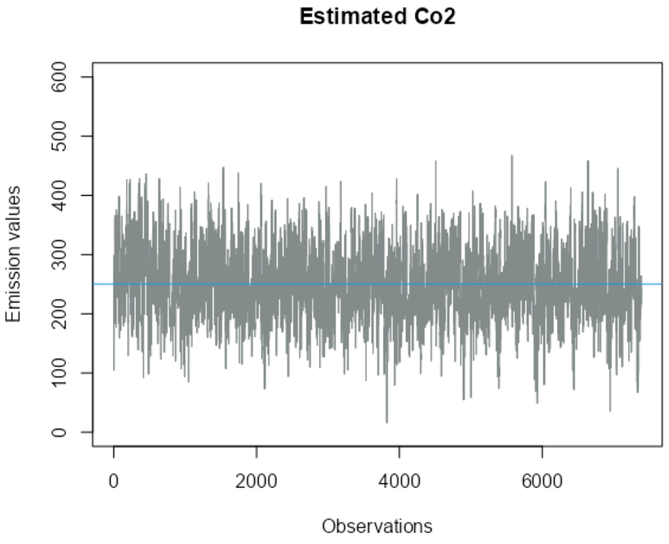
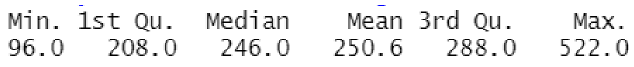
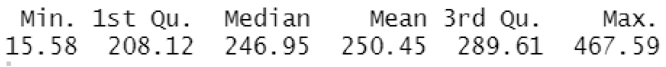


Fig 2.4.5. Simulated Co2 Emissions

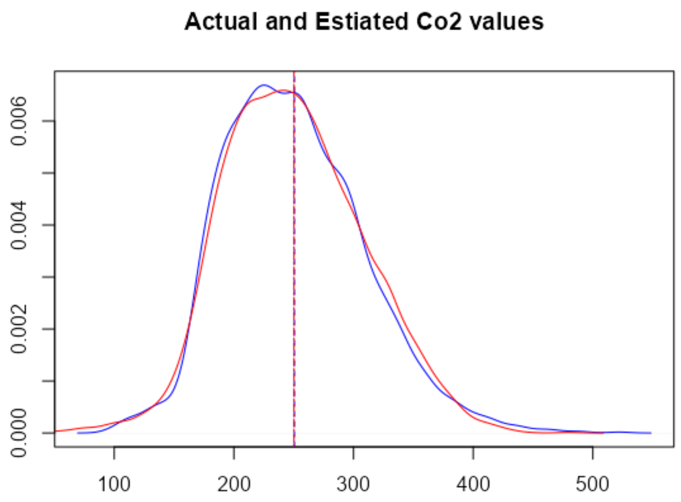


```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   96.0   208.0   246.0   250.6   288.0   522.0
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.58  208.12  246.95  250.45  289.61  467.59
```

A line plot in Fig 2.4.5 of both shows a slight variation in the two Co2 values. They are still very similar. The Mean for both distributions are shown with a blue and red dashed line. They are practically on top of each other as the lines have blended so the Monte Carlo simulation has done a very good job of estimating the Co2 Emission value from the 1,000 simulations and taking the average.



Lastly a Fig 2.4.5. Co2 Distribution

In Fig 2.4.6 we have created 7 red lines with different simulation results to show that while there is some variation the dashed red mean line does not change, you can notice that you can no longer see the blue actual mean line.

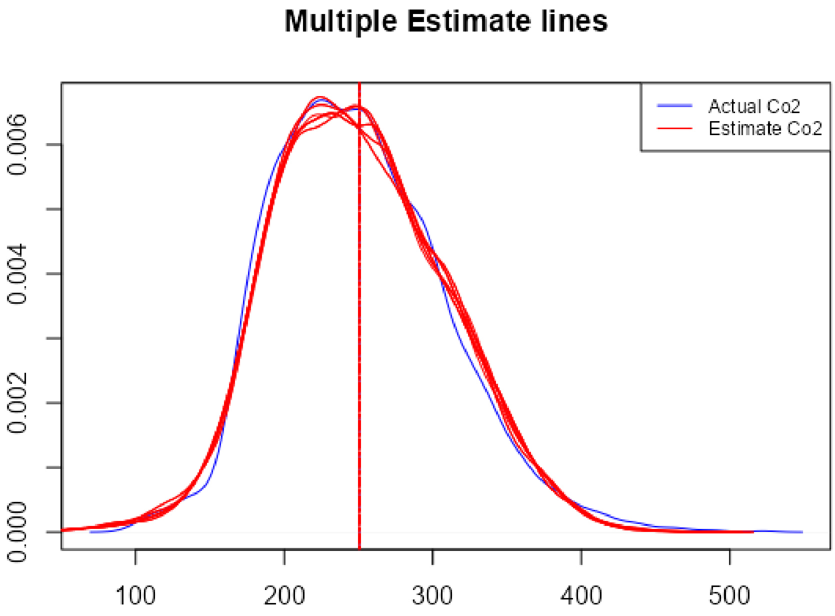


Fig 2.4.6. Seven Simulated Lines

Lastly in Fig 2.4.7 and 2.4.8 we can also see a histogram of both distributions, they look almost identical but they are slightly right and left of each other.

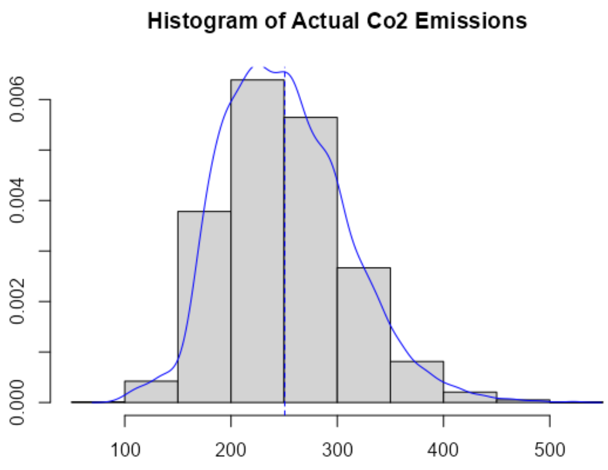


Fig 2.4.7. Actual Co2 Emissions

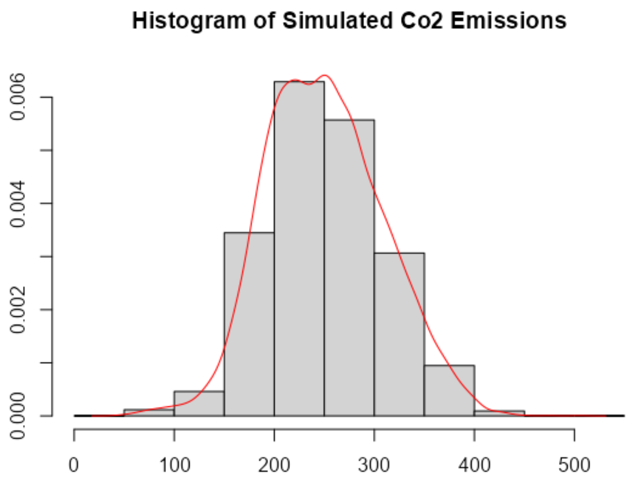


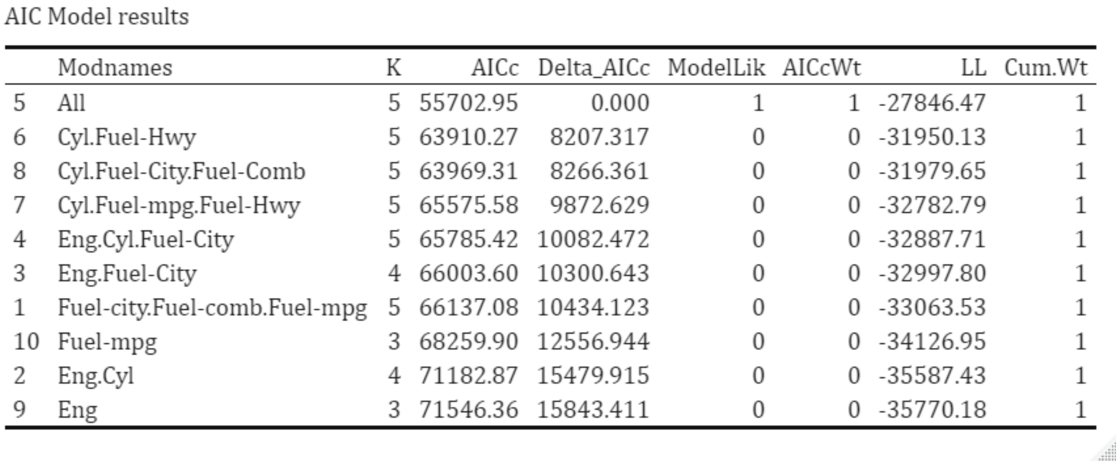
Fig 2.4.8. Simulated Co2 Emissions

# 3. MODEL 2: BEST AIC MODEL

Next, we will look at a second Linear Regression Model and once again compare the results. This time we will use AIC to decide what predictors to add to the model.

## 3.1. Linear Regression Model Using AIC

Previously I have used p-values to look at which predictors are the best to be included in a model. For this project, I wanted to explore a different method and so used the Akaike Information Criterion (AIC) to compare 10 different models and see which was the best. Fig 3.1 shows the result of the ten models, an AIC table automatically sorts the best model to the top and in this case, the best model by AIC was to have All the numeric predictors. I was a little surprised by this given that when I look at the model itself after it ran, some of the p-values were above a normal level of significance and some did not have the usual 3 stars (guideline only) that help to indicate what the Linear Model its self was showing as important of the variables. However, I decided to follow the AIC results for this part to compare with the Reduced Model given in the assignment.



AIC Model results

| | Modnames | K | AICc | Delta_AICc | ModelLik | AICcWt | LL | Cum.Wt |
|---|---|---|---|---|---|---|---|---|
| 5 | All | 5 | 55702.95 | 0.000 | 1 | 1 | -27846.47 | 1 |
| 6 | Cyl.Fuel-Hwy | 5 | 63910.27 | 8207.317 | 0 | 0 | -31950.13 | 1 |
| 8 | Cyl.Fuel-City.Fuel-Comb | 5 | 63969.31 | 8266.361 | 0 | 0 | -31979.65 | 1 |
| 7 | Cyl.Fuel-mpg.Fuel-Hwy | 5 | 65575.58 | 9872.629 | 0 | 0 | -32782.79 | 1 |
| 4 | Eng.Cyl.Fuel-City | 5 | 65785.42 | 10082.472 | 0 | 0 | -32887.71 | 1 |
| 3 | Eng.Fuel-City | 4 | 66003.60 | 10300.643 | 0 | 0 | -32997.80 | 1 |
| 1 | Fuel-city.Fuel-comb.Fuel-mpg | 5 | 66137.08 | 10434.123 | 0 | 0 | -33063.53 | 1 |
| 10 | Fuel-mpg | 3 | 68259.90 | 12556.944 | 0 | 0 | -34126.95 | 1 |
| 2 | Eng.Cyl | 4 | 71182.87 | 15479.915 | 0 | 0 | -35587.43 | 1 |
| 9 | Eng | 3 | 71546.36 | 15843.411 | 0 | 0 | -35770.18 | 1 |

Fig 3.1 AIC table

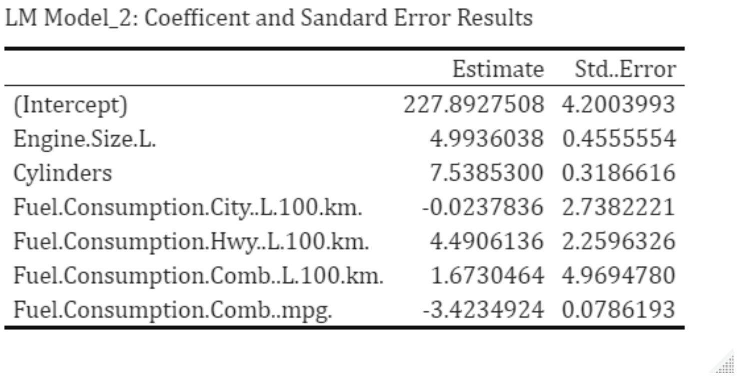
## 3.2. Simulations

Once again we follow the same procedure as above, this time the model was:

$$CO2=\beta 0 +\beta 1 Eng.size+\beta 2 Cylinders+\beta 3 Fuel.conscomb+\beta 4 Fuel.consmpg+\beta 4 Fuel.cons\text{city} +\beta 4 Fuel.cons\text{hwy} + \epsilon$$

## 3.3. Compare results

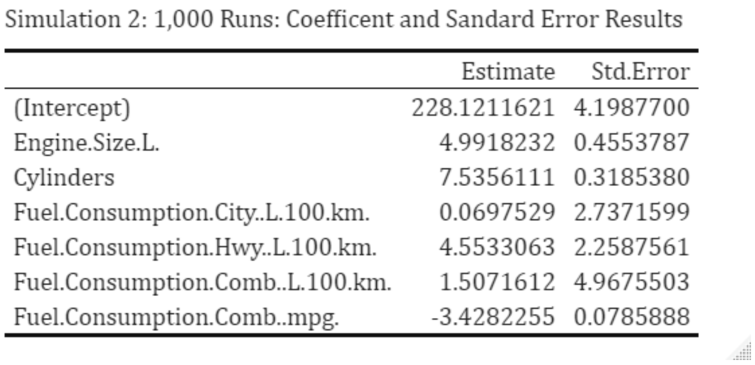
We start by looking at the coefficients and Standard Error for Model 2 and the Simulated Model in Fig 3.3.1 and 3.3.2. Again, after 1,000 simulations we can see that most results have got very close. A few, however, are off by quite a bit, the most noticeable is Fuel.Consumption.L.100.Km which has a -0.023 value in Model 2 and a positive value of 0.069, whether this would be important depends on our Project Outline.



LM Model_2: Coefficent and Sandard Error Results

| | Estimate | Std..Error |
|---|---|---|
| (Intercept) | 227.8927508 | 4.2003993 |
| Engine.Size.L. | 4.9936038 | 0.4555554 |
| Cylinders | 7.5385300 | 0.3186616 |
| Fuel.Consumption.City..L.100.km. | -0.0237836 | 2.7382221 |
| Fuel.Consumption.Hwy..L.100.km. | 4.4906136 | 2.2596326 |
| Fuel.Consumption.Comb..L.100.km. | 1.6730464 | 4.9694780 |
| Fuel.Consumption.Comb..mpg. | -3.4234924 | 0.0786193 |

Fig 3.3.1 Model 2 Coefficients

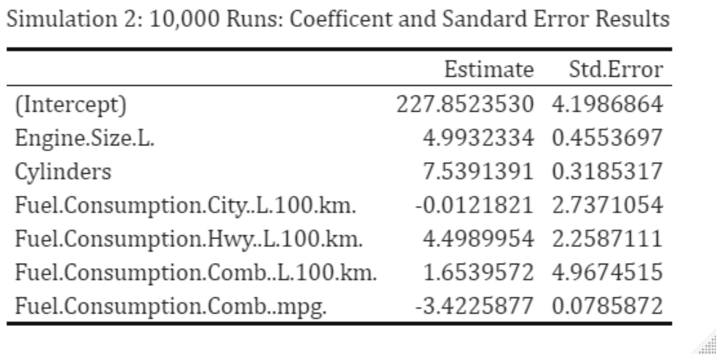


Simulation 2: 1,000 Runs: Coefficent and Sandard Error Results

| | Estimate | Std.Error |
|---|---|---|
| (Intercept) | 228.1211621 | 4.1987700 |
| Engine.Size.L. | 4.9918232 | 0.4553787 |
| Cylinders | 7.5356111 | 0.3185380 |
| Fuel.Consumption.City..L.100.km. | 0.0697529 | 2.7371599 |
| Fuel.Consumption.Hwy..L.100.km. | 4.5533063 | 2.2587561 |
| Fuel.Consumption.Comb..L.100.km. | 1.5071612 | 4.9675503 |
| Fuel.Consumption.Comb..mpg. | -3.4282255 | 0.0785888 |

Fig 3.3.2 Simulation: 1000

To investigate this difference a little more I ran a 10,000 simulation and we can see we get better results for the Fuel.Consumption.L.100.Km variable. This time we have a value of -0.012, closer to the Model value of -0.023. As mentioned this is what makes Monte Carlo Simulations work so well, the more times the simulation is running the more variables that we averaged and the better the result. However, there is a point where the extra time is not worth the extra small improvements. Also, it takes more computer power and my own laptop didn’t seem to like it when I tried 100,000.



Simulation 2: 10,000 Runs: Coefficent and Sandard Error Results

|  | Estimate | Std.Error |
|---|---|---|
| (Intercept) | 227.8523530 | 4.1986864 |
| Engine.Size.L. | 4.9932334 | 0.4553697 |
| Cylinders | 7.5391391 | 0.3185317 |
| Fuel.Consumption.City..L.100.km. | -0.0121821 | 2.7371054 |
| Fuel.Consumption.Hwy..L.100.km. | 4.4989954 | 2.2587111 |
| Fuel.Consumption.Comb..L.100.km. | 1.6539572 | 4.9674515 |
| Fuel.Consumption.Comb..mpg. | -3.4225877 | 0.0785872 |

Fig 3.3.3 Simulation: 10,000

In Fig 3.3.4 and 3.3.5 we see quite similar results to the previous Model. The range of values for the estimate is a little closer but this is more down to the variation in the Simulated data than because of the more complex model.

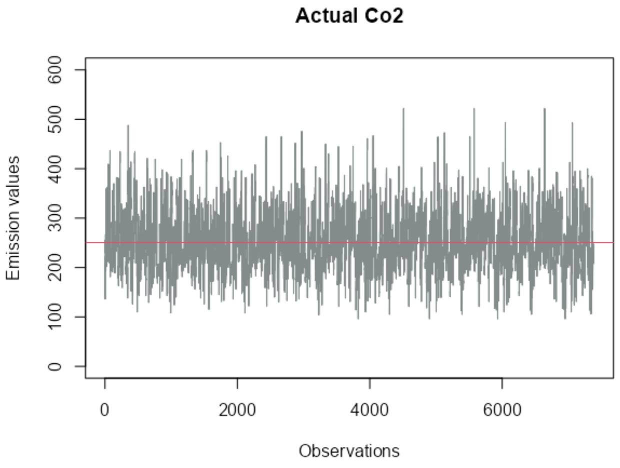


Fig 3.3.4. Actual Co2 Emissions

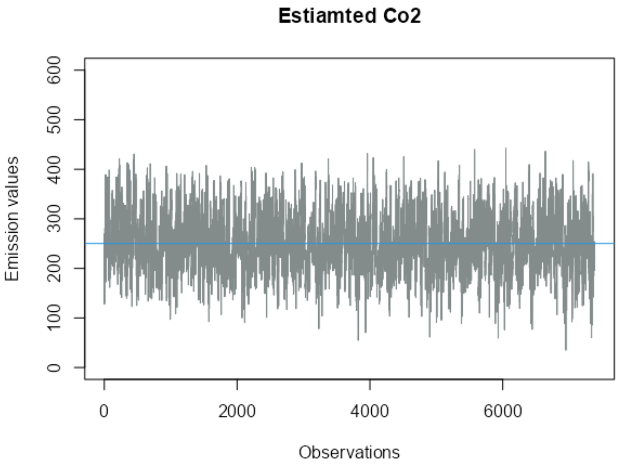
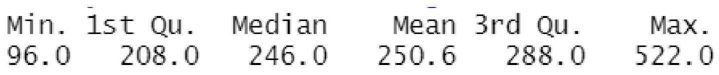
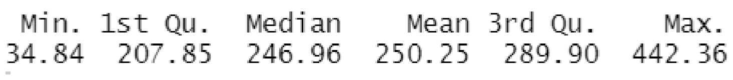


Fig 3.3.5. Simulated Co2 Emissions



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
96.0   208.0   246.0   250.6   288.0   522.0
```



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
34.84  207.85  246.96  250.25  289.90  442.36
```

Lastly, in Fig 3.3.6 we see a distribution comparison again, but it is similar to what we have seen above. There will always be variations in simulation models.

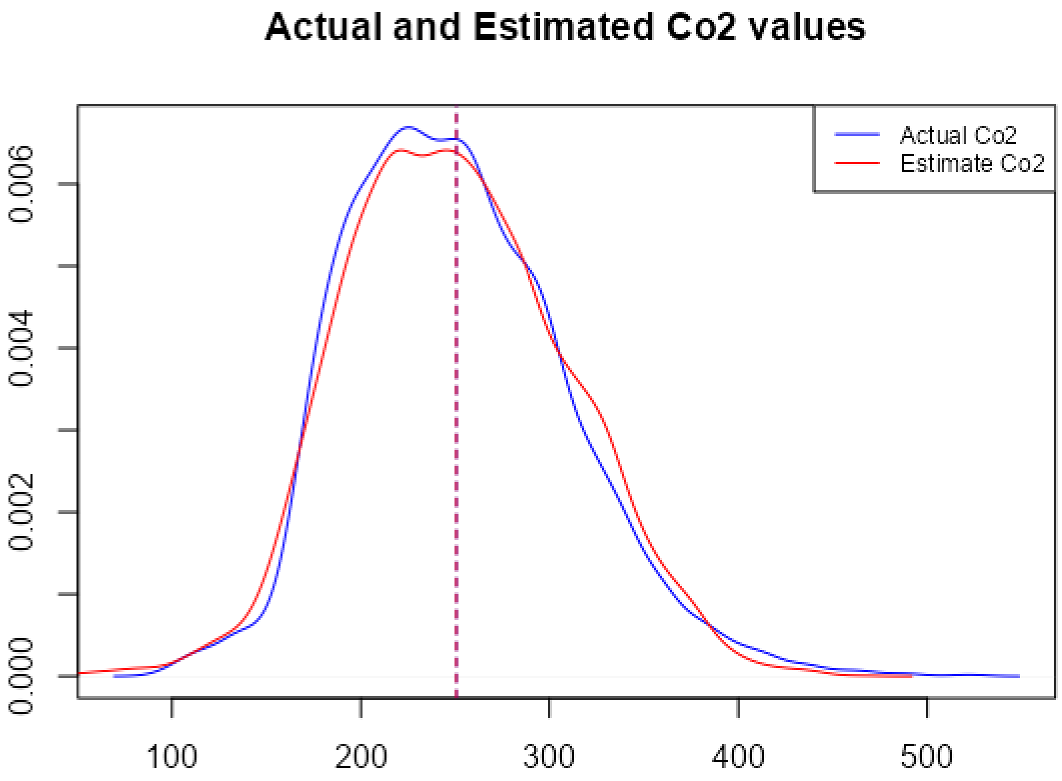


Fig 3.3.6: Distribution of mean Co2 values.

## 4. CONCLUSIONS

In this project, we have shown how to predict coefficient and Standard Error values to a high similarity using Monte Carlo Simulations. We already had the Co2 results so we were able to see just how good the Monte Carlo Simulations are at predicting values.

We also saw in certain situations that the more simulations we ran the better the averaged results and the close they are to the real values. We ran 1,000 and 10,000 simulations and the more simulations the closer the results are to the real results. There is a trade-off in that more simulations take more time and more computer power, weather this small increase in accuracy is important depends on the project.