STAT 886 Project: Predicting Car MPG using Bayesian Regression and Comparing with Frequentist Results

Oliver Titus

May 2021

1 Introduction

The fuel economy of a car can depend on many factors. According to the U.S. department of energy, factors that may affect fuel economy include how you drive, vehicle maintenance, fuel variations, vehicle variations, as well as engine break-in [3]. In this paper, we use a data set from the UCI Machine Learning Repository called Auto MPG [5]. The data set contains the miles per gallon (our response variable) for many different cars along with several explanatory variables including the number of cylinders, displacement, horsepower, acceleration, and weight of the vehicle.

The purpose of this paper is to use this data set to compare two different regression approaches. We compare Bayesian regression to Ordinary Least Squares (OLS), where OLS is a frequentest approach to linear regression. We propose three different model specifications and choose the best one using Deviance Information Criterion (DIC). We compare the Bayesian results of best model to the OLS results.

In the next section, we will describe the data set as well as provide some summary statistics. In the models section, we will setup the Bayesian framework and describe the models we tested. The results section will compare the Bayesian and Frequentest approaches. We will discuss overall and conclusions in the data section and then propose future work.

2 Data

The data set used is from UCI Machine Learning Repository. This data set is a slightly modified version from the original which was from the StatLib library which is maintained at Carnegie Mellon University. The data set conerns city-cylce fuel economy in miles per gallon (mpg), which is to be predicted by three multi-valued discrete as well as five continuous attributes. This data set was used to test graphical analysis packages at the 1983 American Statistical Association Exposition [4].

The data set includes the following continuous attributes:

- mpg: the miles per gallon for the car
- displacement: the size of the engine
- horsepower
- weight (lbs)
- acceleration (0-60mph times)

The data set also includes the following discrete attributes:

- number of cylinders
- model year
- origin

The data set also includes the car name which is a string. See Table 1 in Appendix A for summary statistics for these attributes. One thing to note is that there were 6 observations that had missing values for horsepower. These rows were removed from the data.

3 Models

We first consider the ordinary least squares regression model, where we assume that the observation errors are independent and have a homogeneous variance. In general, the model for all observations can be written:

$$y \mid \beta, \sigma^2, X \sim N(X\beta, \sigma^2 I)$$

To setup of the Bayesian framework for our model, we use the standard noninformative prior distribution for (β, σ) which in general can be written [1]:

$$p(\beta, \sigma \mid X) \propto \sigma^{-2}$$

The joint posterior distribution can be written:

$$p(\beta, \sigma^2 \mid y) \propto p(\beta \mid \sigma^2, y) p(\sigma^2 \mid y)$$

And the conditional posterior distribution for β can be written:

$$\beta \mid \sigma, y \sim N(\hat{\beta}, V_\beta \sigma^2)$$

where

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$V_\beta = (X^T X)^{-1}$$

Our goal is estimate the marginal posterior distribution for the parameters [2]. We can write the marginal posterior distribution of σ^2 as

$$p(\sigma^2 \mid y) = \frac{p(\beta, \sigma^2 \mid y)}{p(\beta \mid \sigma^2, y)}$$

We also want estimate the marginal posterior distribution for the β vector. To do this, we simulate the marginal posterior distribution for β using Markov Chain Monte Carlo (MCMC) methods. We use R2OpenBugs, which utilizes Gibbs Sampling, to simulate the marginal distributions for the parameters. We use the following non-informative priors for the parameters:

$$\beta_0 \sim N(0, 1000)$$
$$\beta_i \sim Uniform(-1000, 1000)$$
$$\sigma^2 \sim Inv - Gamma(0.5, 0.5)$$

We consider 3 model specifications for predicting MPG:

$$\log(mpg) = \beta_0 + \beta_1 * model.year + \beta_2 * acceleration + \beta_3 * horsepower \tag{1}$$

$$\log(mpg) = \beta_0 + \beta_1 * model.year + \beta_2 * acceleration + \beta_3 * displacement \tag{2}$$

$$\log(mpg) = \beta_0 + \beta_1 * model.year + \beta_2 * acceleration + \beta_3 * weight \tag{3}$$

Something to note is that we do not include all the explanatory variables in these models in order to avoid overfitting and multicollinearity. We did not want to include horsepower, displacement, and weight in the same model as these variables were highly correlated. See Tabel 2 for the correlation matrix. We do not consider the number of cylinders in these models as there is not a lot of variation associated with this variable. We also do not include origin as that variable is categorical and it would not be able to be used in the regression framework.

We choose which of these three models is best by checking the Deviance Information Criterion (DIC) where the deviance is generally defined as

$$D(\theta_m, m) = -2 \log p_m(y \mid \theta_m)$$

We choose the model that minimizes the DIC.

4 Results

In order to ensure the model converges, we use 10,0000 iterations for burn-in and 100,000 total iterations with 1 chain. Results for all three models using OpenBugs are given in Appendix A in Tabels 3 through 5 where the mean and standard deviation for the marginal distributions for the parameters are given as well as the quantiles. Trace Plots as well as marginal densities are given in Appendix B for these models.

Looking at the R2OpenBugs results for model 1 in Tabel 1, we see that the relationship between model year and the log of MPG are positive with a 95% credible interval of (0.02147, 0.03092). The relationship between acceleration and the log of MPG is negative with a 95% credible interval of (-0.0144, -0.0013). The relationship between horsepower and the log of MPG is negative with a 95% credible interval of (-0.00836, -0.00715). In Figure 3, we see that the trace plots for the intercept and model year show a slight wavy pattern which means the model is not mixing very well for these parameters. Evidence of this is also apparent based on the marginal densities for the intercept and model year in Figure 4 as these do not look perfect. The rest of the parameters seem to be converging just fine as the trace plots look decent as well as the marginal densities.

The results for model 2, we see that the relationship between model year and log of MPG as well as acceleration and log of MPG are negative. The 95% credible interval for the coefficient for model year is (0.02440, 0.03317) and for acceleration, the 95% credible interval is (-0.0144, -0.0013). The relationship between displacement and log of MPG is negative with 95% credible interval of (-0.00269, -0.00233). The trace plots and marginal densities in Figure 6 are similar in shape to as shown in model 1 with model year and the intercept showing signs of not mixing perfectly, with the rest of the parameters looking fine.

Examining the results for model 3, the relationship between model year and the log of MPG is positive with 95% credible interval of (0.02726, 0.03455). One thing to note that is different about this model compared the first and second is that the relationship between acceleration and the log of MPG is positive overall with 95% credible interval of (-0.000093, 0.00938). It would make sense for the coefficient on acceleration to be positive as the car with a longer 0 to 60 mph time would have a more fuel efficient engine. The relationship between weight and the log of MPG is negative with 95% credible interval of (-0.00032. -0.00029). This model had the lowest DIC compared to the other two models. Because of this, we consider it to be the best model of fit compared to the other two. We compare the R2OpenBugs results of this model to the R Stan results as well as the Frequentest OLS results.

Using the R Stan 'brms' package, we specified the same model as well as priors as to that of model 3. Only thing different here is that we used the default 1000 iterations for burn-in and 2000 iterations total. We see that the results are identical to the OpenBugs results in Table 6. The marginal densities as well as trace plots for the R Stan version of Model 3 are given in Figure 9. We see that the marginal densities and trace plots look decent. This shows that the model is converging, which is also evident with the R-hat values in Table 6. What is interesting here is that far fewer iterations were needed in R Stan compared to OpenBugs for the model to converge.

The Frequentist results for Model 3 is given in Table 7. In this table, the estimates are given as well as the standard errors in parentheses. We see that the coefficients are identical to the OpenBUGS results and very close to the R Stan results. Only difference here is that we can look at p-values. We see that the intercept and coefficient for weight is highly statistically significant with p-values less than 0.001. Acceleration on the other hand is weakly statistically significant with p-value less than 0.1. The coefficient for model year is highly statistically significant with p-value less than 0.001. We also see that this model results in an R^2 of 0.87 which tells us that 87% of the variation in the log of MPG can be explained by weight, acceleration, and model year.

5 Conclusion

We found that the model with the weight, model year, and acceleration seem to be the most appropriate based on this model having the lowest DIC. Looking at the results as a whole, we found strong evidence of a negative effect on the miles per gallon (MPG) and weight. We found weak evidence of a positive effect on MPG with acceleration. We also found strong evidence of a positive effect on MPG with the model year. The Bayesian and frequentest results for this model were almost identical. We suspect this may be due to the data set being large.

In future work, it may be interesting to see if specifying different prior distributions for the model would change the DIC. There might be a better choice of prior for the parameters. It also may be interesting to try and incorporate the Origin variable into the model and see if it has a significant relationship with MPG. We could incorporate this variable doing an ANCOVA model where we look at how the regression coefficients compare across different origins. Another way to do this would be to include dummy variables into the standard regression model. Comparing other approximation methods such as the EM algorithm, importance sampling, and other computational methods would be other realms of future work. Also, it would be interesting to compare the Bayes factor for the parameters to the p-values in the OLS results to see if those results are consistent.

6 Appendix A: Tables

Table 1: Summary Statistics

	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
MPG	9	17	22.75	23.45	29	46.6
Cylinders	3	4	4	5.472	8	8
Displacement	68	105	151	194.4	275.8	455
Horsepower	46	75	93.5	104.5	126	230
Weight	1613	2225	2804	2978	3615	5140
Acceleration	8	13.78	15.5	15.54	17.02	24.80
Origin	1	1	1	1.577	2	3
Model year	70	73	76	75.98	79	82

Table 2: Correlation Matrix

	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleraion	Model Year	Origin
MPG	1	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
Cylinders	-0.78	1	0.95	0.84	0.90	-0.50	-0.35	-0.57
Displacement	-0.81	0.95	1	0.90	0.93	-0.54	-0.37	-0.61
Horsepower	-0.78	0.84	0.90	1	0.86	-0.69	-0.42	-0.46
Weight	-0.83	0.90	0.93	0.86	1	-0.42	-0.31	-0.59
Acceleration	0.42	-0.50	-0.54	-0.69	-0.42	1	0.29	0.21
Model Year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1	0.18
Origin	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1

Table 3: R2OpenBUGS Results for Model 1

	Mean	SD	2.5%	25%	50%	75%	97.5%
alpha	2.37694	0.20897	1.987	2.23	2.375	2.509	2.805
beta1	0.02617	0.00239	0.02147	0.02458	0.02611	0.02783	0.03092
beta2	-0.02945	0.00411	-0.03754	-0.03225	-0.02946	-0.02659	-0.02162
beta3	-0.00774	0.00031	-0.00836	-0.00795	-0.00774	-0.00753	-0.00715
tau	36.17638	2.59036	31.24975	34.41	36.12	37.9	41.41
deviance	-328.17152	4.10370	-333.9	-331.2	-328.9	-325.9	-318.2

DIC = -323.3

Table 4: R2OpenBUGS Results for Model 2

	Mean	SD	2.5%	25%	50%	75%	97.5%
alpha	1.52227	0.17798	1.181	1.399	1.523	1.636	1.883
beta1	0.02876	0.00222	0.02440	0.02728	0.0287	0.0303	0.03317
beta2	-0.00778	0.00337	-0.0144	-0.01006	-0.00777	-0.00546	-0.0013
beta3	-0.00251	0.00009	-0.00269	-0.00257	-0.00251	-0.00245	-0.00233
tau	40.7429	2.91787	35.2	38.75	40.68	42.69	46.63
deviance	-379.3896	4.34861	-385.5	-382.6	-380.2	-377	-368.9

DIC = -374.5

Table 5: R2OpenBUGS Results for Model 3

	Mean	SD	2.5%	25%	50%	75%	97.5%
alpha	1.58938	0.14816	1.306	1.487	1.5898	1.684	1.888
beta1	0.03087	0.00185	0.02726	0.02964	0.03082	0.03216	0.03455
beta2	0.00426	0.00266	-0.00093	0.00246	0.00427	0.00609	0.00938
beta3	-0.00030	0.00001	-0.00032	-0.00031	-0.00030	-0.00030	-0.00029
tau	57.24394	4.09945	49.45	54.45	57.15	59.9725	65.52
deviance	-529.13418	5.25421	-536.9	-533	-529.9	-526.1	-516.7

DIC = -524.3

Table 6: R Stan 'brms' Results

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	1.60558	0.14479	1.32419	1.88952	1.00290	3940	2522
model.year	0.03068	0.00184	0.027-7	0.03419	1.00131	3607	2257
weight	-0.00030	0.00001	-0.00032	-0.00029	1.00050	3819	3207
acceleration	0.00427	0.00254	-0.00077	0.00914	0.99987	3065	2452
sigma	0.12276	0.00426	0.11465	0.13165	1.00189	1664	1910

Table 7: OLS Results for Model 3

	Model 3
(Intercept)	1.602^{***}
	(0.14)
weight	-0.0003^{***}
	(0.000008)
acceleration	$0.004^{\cdot}$
	(0.003)
model.year	0.03^{***}
	(0.002)
R^2	0.87
Adj. R^2	0.87
Num. obs.	392

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

7 Appendix B: Figures

Figure 1: Historgrams of MPG and Log of MPG

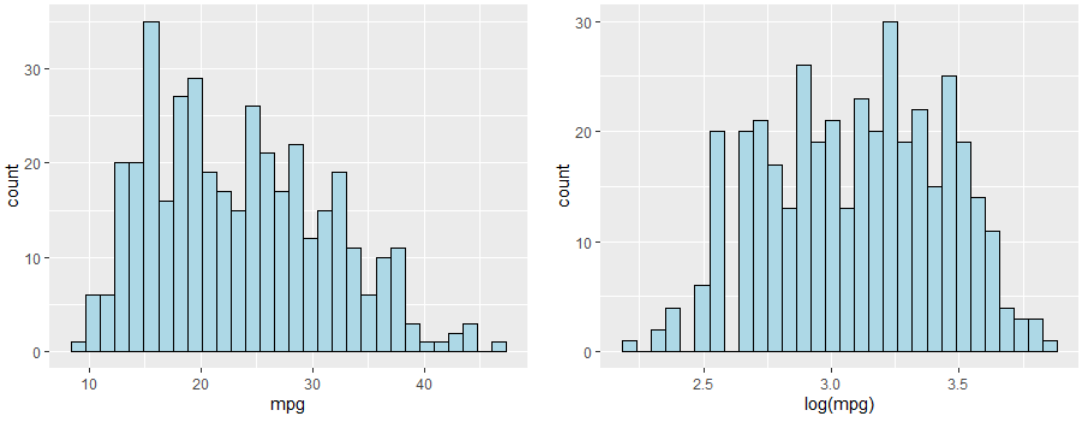


Figure 2: Scatter plots with exponential curve of best fit

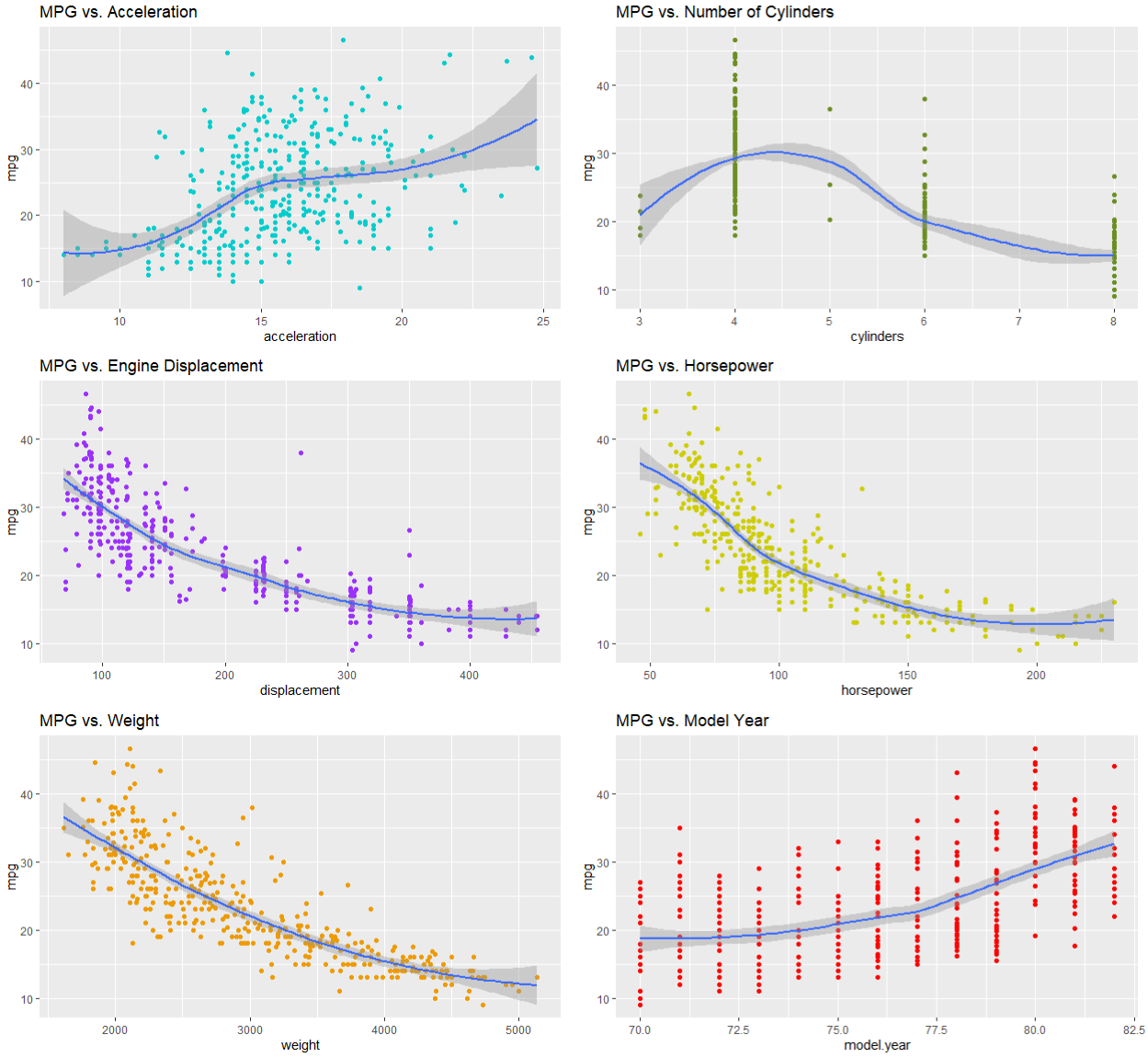


Figure 3: Trace Plots for the parameters in Model 1

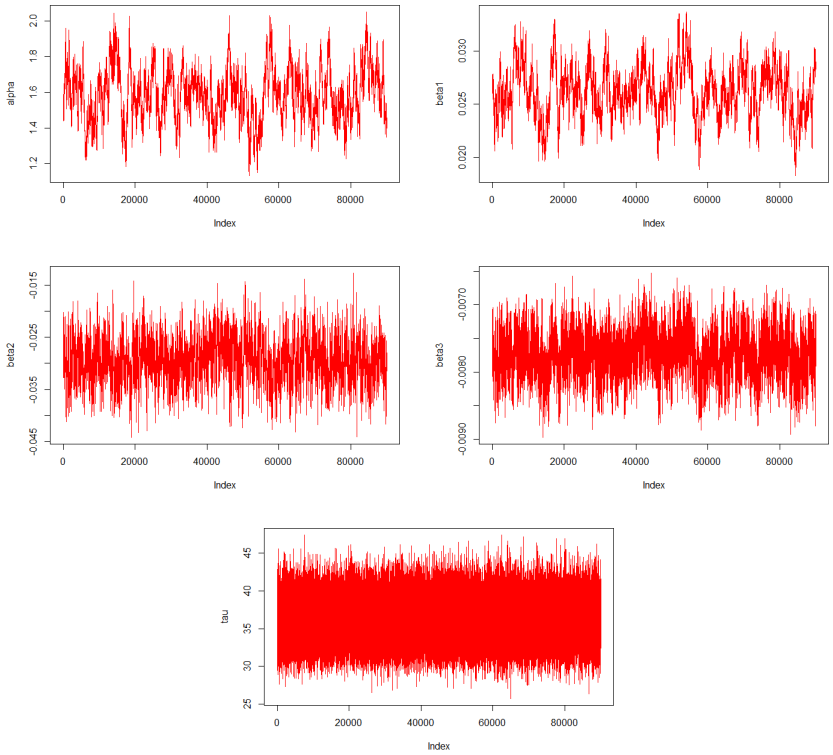


Figure 4: Marginal densities for the parameters in Model 1

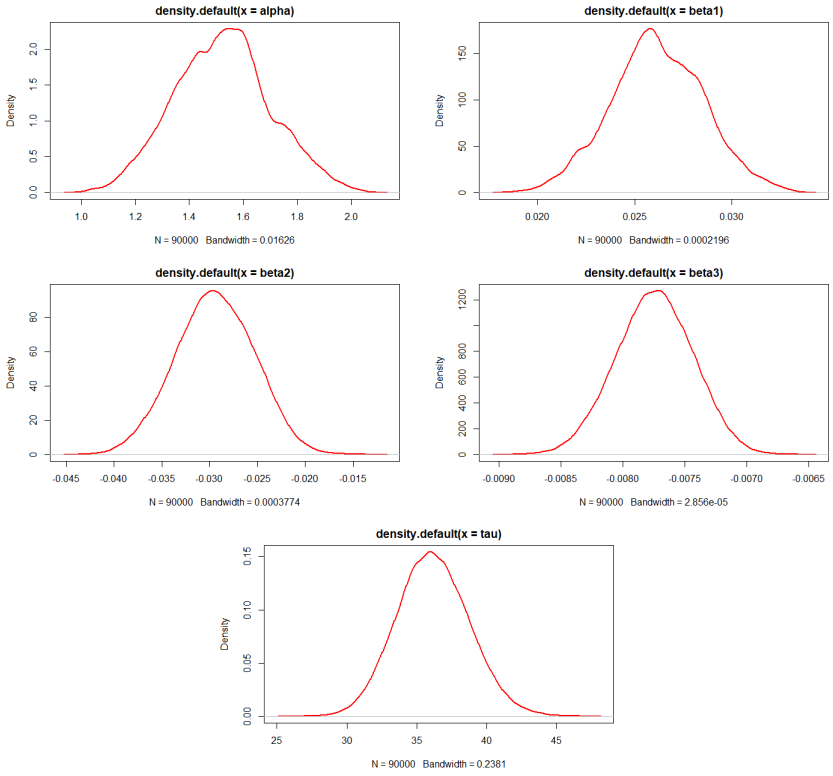


Figure 5: Trace plots for the parameters in Model 2

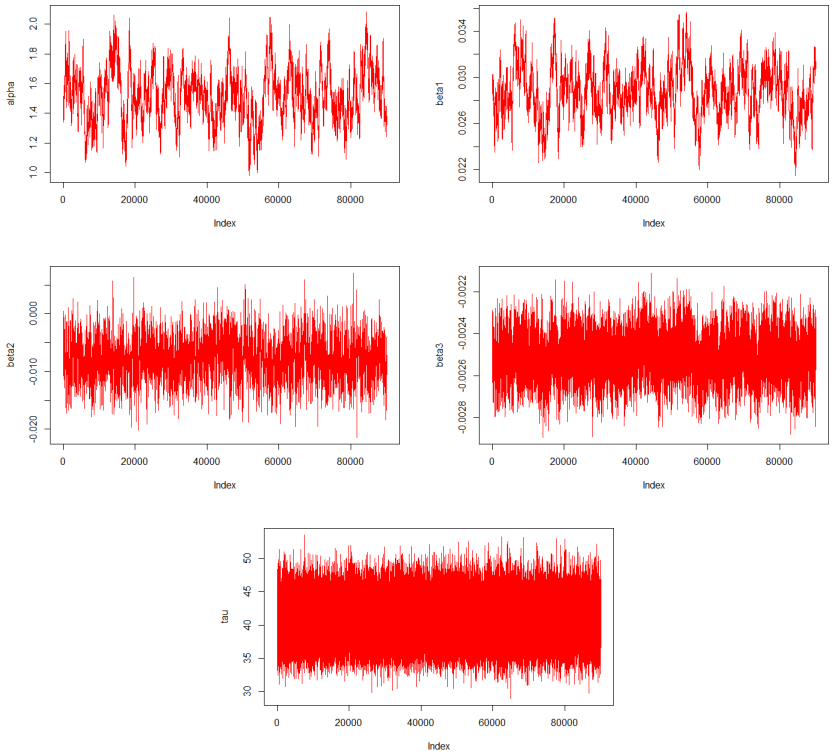


Figure 6: Marginal densities for the parameters in Model 2

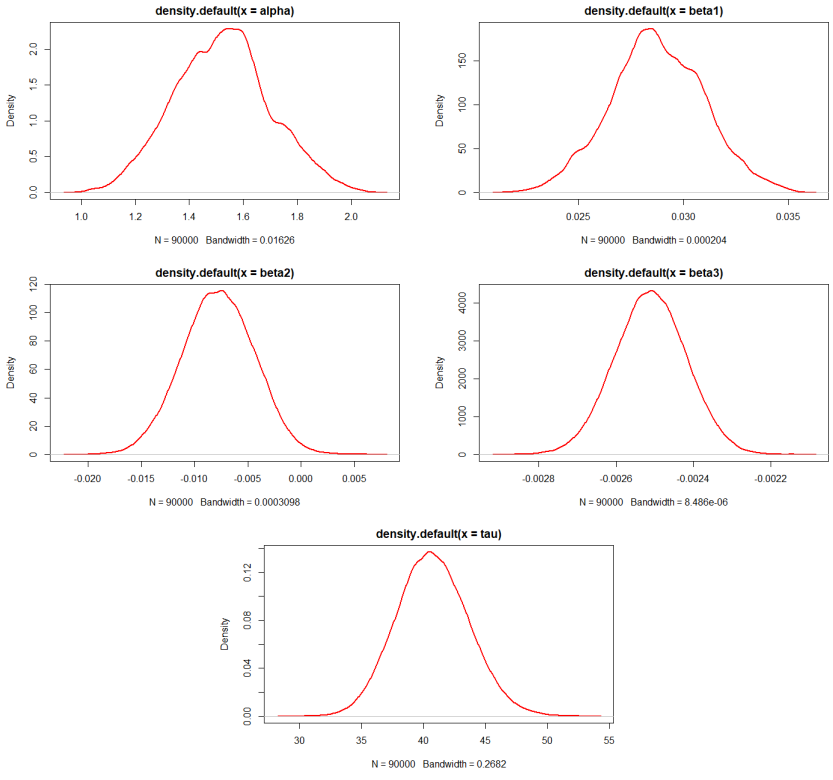


Figure 7: Trace plots for the parameters in Model 3

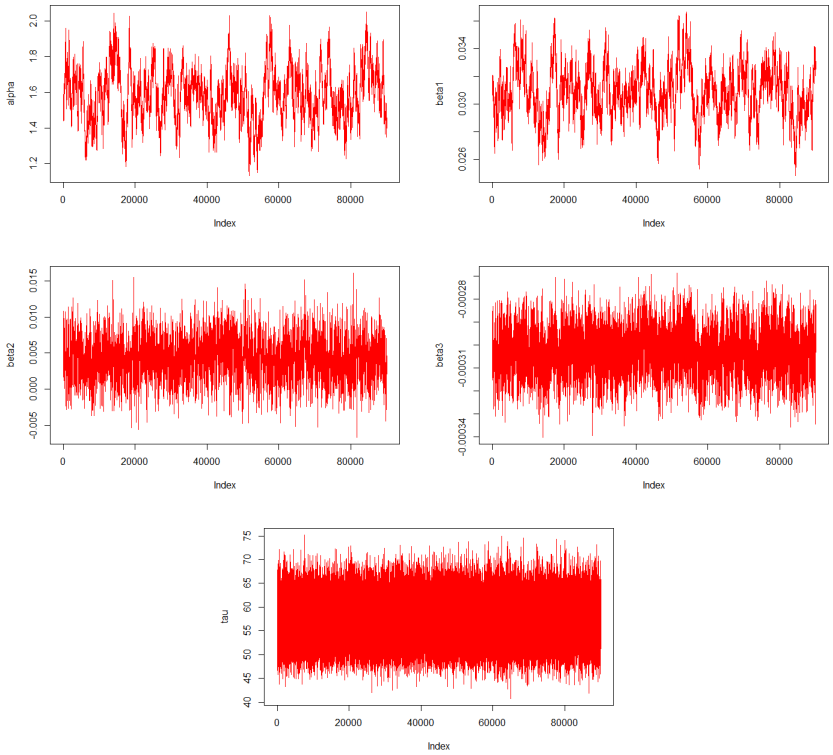


Figure 8: Marginal densities for the parameters in Model 3

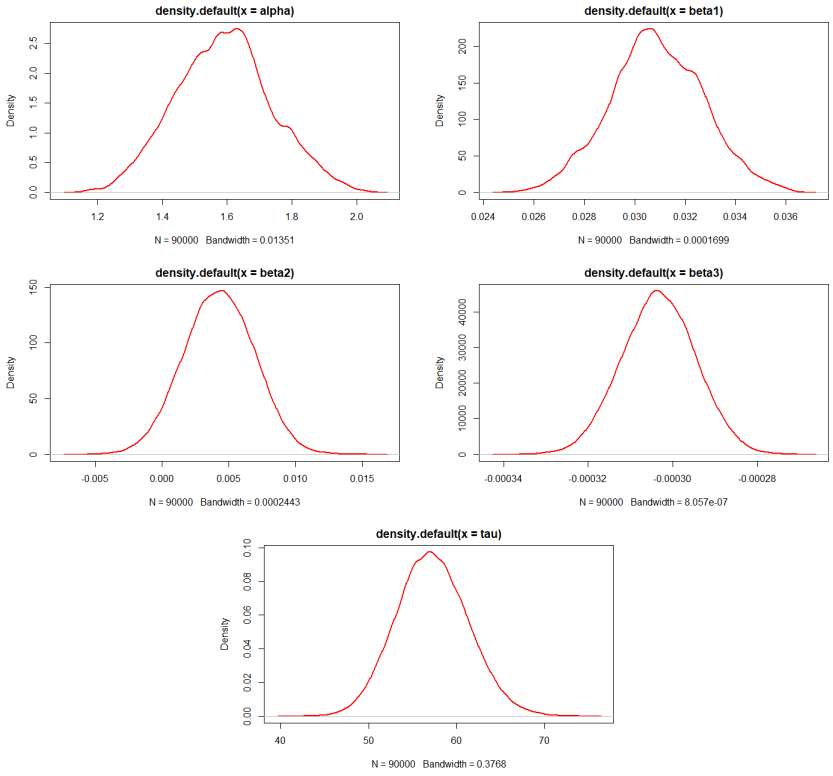
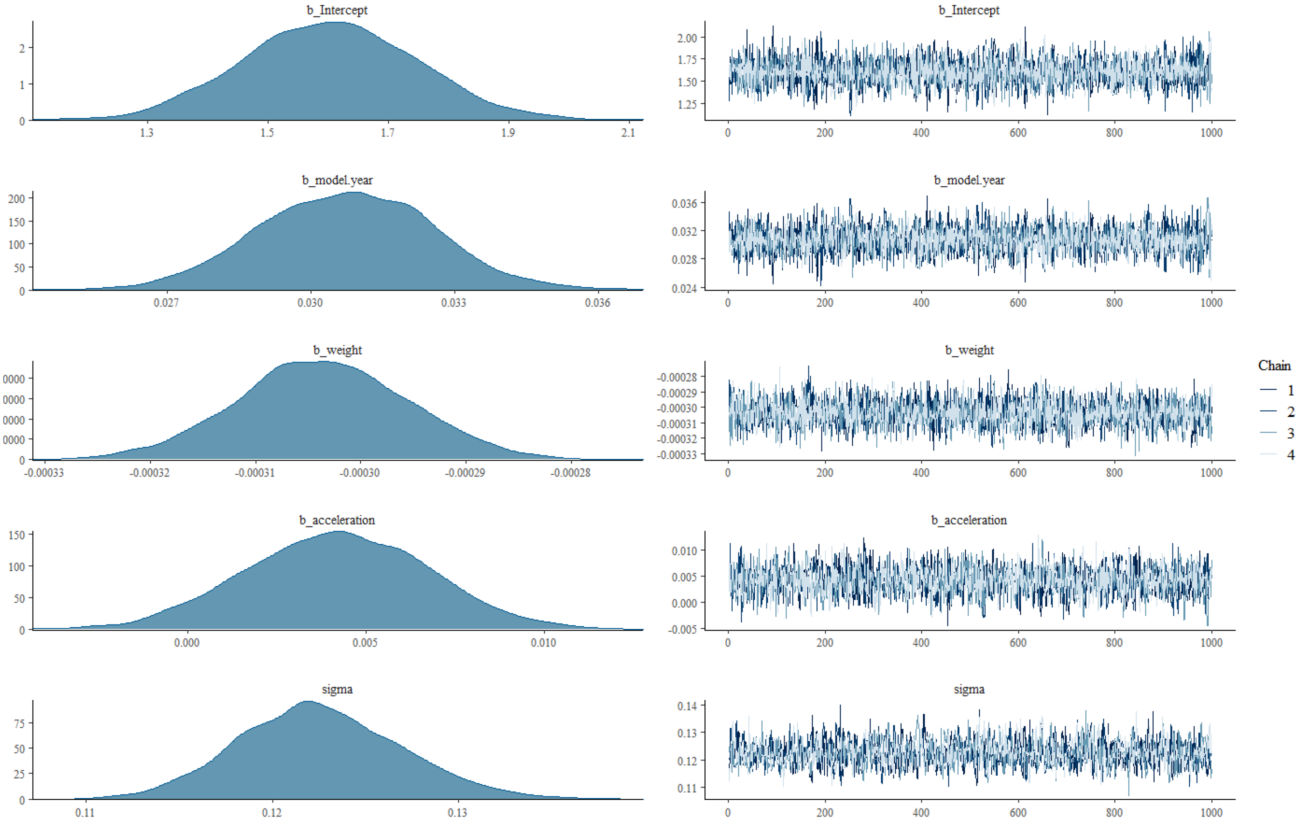


Figure 9: R Stan 'brms' Results for Model 3



References

[1] Jim Albert. *Bayesian computation with R*. Springer, 2009.

[2] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[3] US Department of Energy. Many factors affect fuel economy. https://www.fueleconomy.gov/feg/factors.shtml. Accessed: 5-01-2021.

[4] J Ross Quinlan. Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 236–243, 1993.

[5] UCI Machine Learning Repository. Auto mpg data set. https://archive.ics.uci.edu/ml/datasets/auto+mpg. Accessed: 5-01-2021.