Predicting Car MPG using Bayesian Regression and Comparing with OLS

By Oliver Titus

UNL

April 27, 2021

Introduction

- The fuel economy of a car can depend on many factors according to the U.S. department of energy:
  - how you drive
  - vehicle maintenance
  - fuel
  - vehicle variations
  - engine break-in
- Used a data set from the UCI Machine Learning Repository to fit a Bayesian linear model to predict the response variable "mpg", based on several explanatory variables.
- Used R2OpenBugs to estimate the posterior distributions for the regression parameters.
- Verified the result using Stan with a package call "brms".
- Also compared the results to the standard OLS estimates.

The data set

Table: 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

The data set

Figure: Historgram of MPG

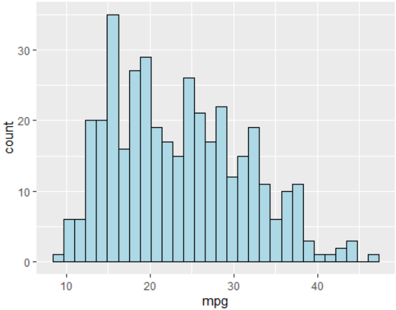
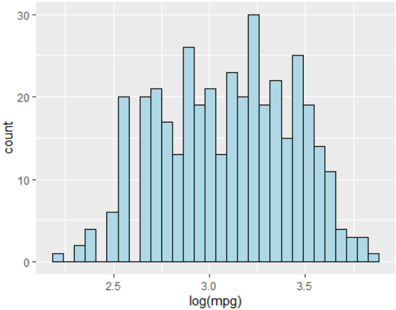
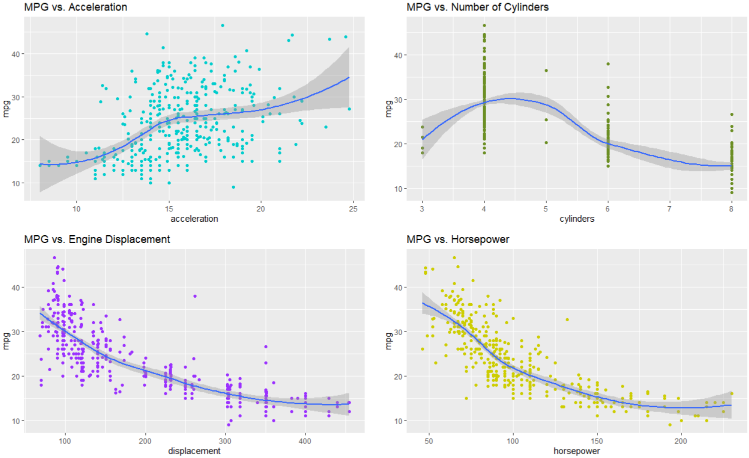


Figure: Histogram of Log MPG



The data set


MPG vs. Acceleration
mpg
acceleration
MPG vs. Number of Cylinders
mpg
cylinders
MPG vs. Engine Displacement
mpg
displacement
MPG vs. Horsepower
mpg
horsepower

The data set


MPG vs. Weight
mpg
40
30
20
10
2000
3000
4000
5000
weight


MPG vs. Model Year
mpg
40
30
20
10
70.0
72.5
75.0
77.5
80.0
82.5
model.year

Correlation Matrix

Table: 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

The Bayesian Framework

In general, the model for all observations can be written:

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

We put the following non-informative priors for the parameters:

$$\begin{aligned} \beta_0 &\sim N(0, 1000) \\ \beta_i &\sim Uniform(-1000, 1000) \\ \sigma^2 &\sim Inv - Gamma(0.5, 0.5) \end{aligned}$$

The conditional posterior distribution:

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

where,

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

Our goal is to estimate the marginal posterior distribution for β.

Three Possible Regression Models

- We consider 3 model specifications for predicting MPG:

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

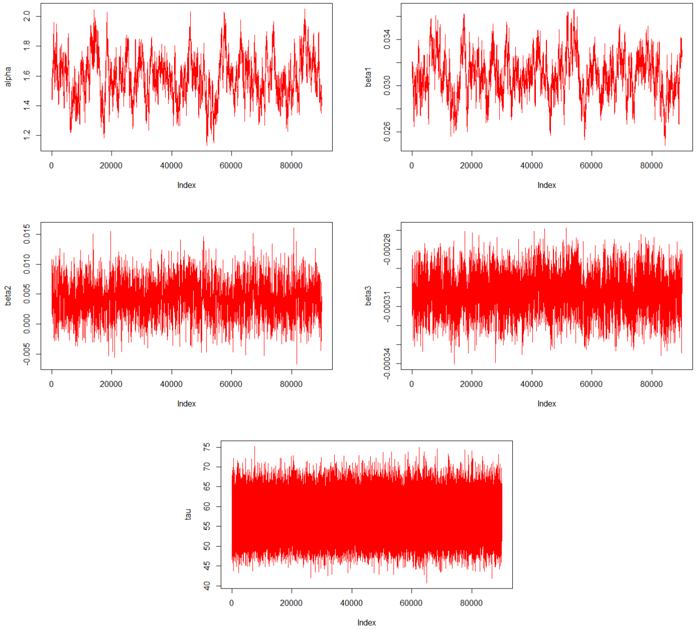
$$\log(mpg) = \beta_0 + \beta_1 * model.year + \beta_2 * acceleration + \beta_3 * displacement \quad (2)$$

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

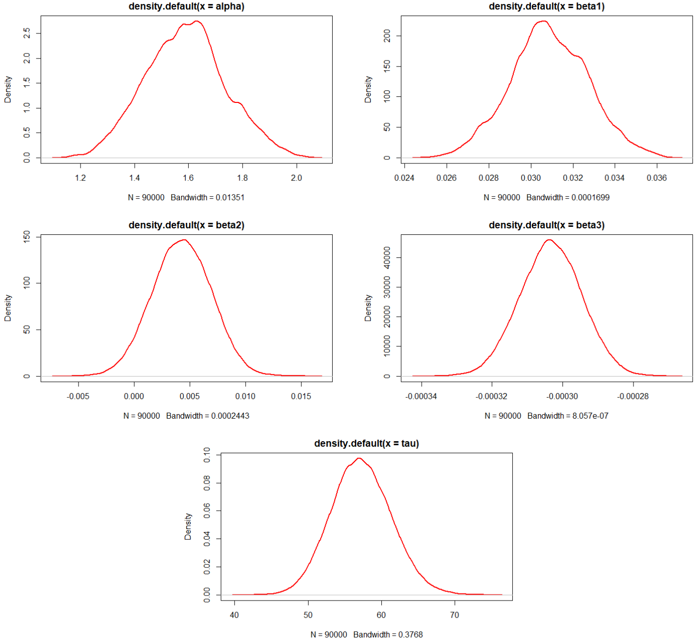
- Not all explanatory variables are included in the model to avoid over-fitting and multicollinearity.

- We pick which of these is the best fit based on the DIC. We find that model 3 had the lowest DIC and therefore has the best fit.

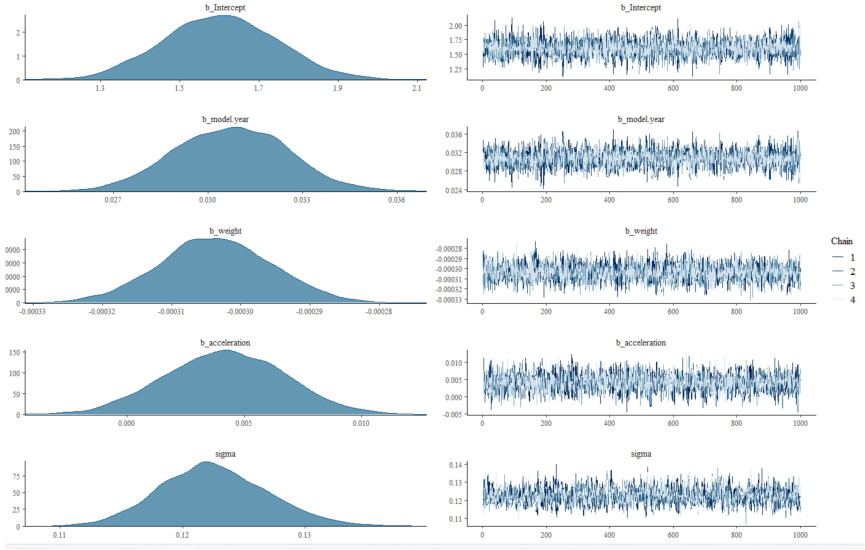
Results: OpenBugs Trace Plots



Results: OpenBugs Marginal Densities



Results: R Stan 'brms' Results



Results: OLS

	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

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

Table: OLS Results

Conclusions

- The model with the weight, model year, and acceleration seems to be the most appropriate.
- There is strong evidence of a negative effect on mpg with the weight.
- There is weak evidence of a positive effect on mpg with acceleration.
- There is strong evidence of a positive effect on mpg with model year.
- Bayesian and OLS results were consistent with these findings.

Future Work

- Trying different prior distributions to see if there's a change with the DIC.
- Trying to incorporate the Origin variable into the model (i.e. dummy variables, ANOVA).
- Comparing other approximation methods such as the EM algorithm, importance sampling, etc.