Bayesian Statistics: From Data to Models

University of California—Santa Cruz(coursera)

A Regression and Bayesian Analysis comparing MPGs of Cars from different countries

**Abstract**

Using a linear regression and Bayesian method I showed that America cars on average consume more fuel than other foreign cars.

**Introduction**

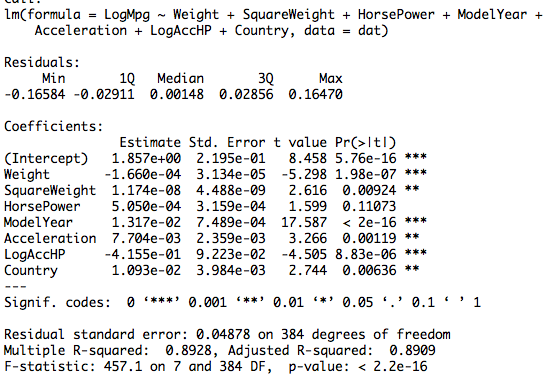
It is generally assumed that American cars have higher gasoline consumption than European or Japanese cars. The goal of this paper is to test this belief and to also create a predictive model for estimating gasoline consumption using mile per gallon (MPG) as a measure. Consumers often purchase European or Japanese cars at a higher price than American cars with similar quality or condition. In light of this hypothesis, this paper focuses on testing if this hypothesis or belief is statistically plausible. Ergo, we hope to answer-- does American cars on average have lower miles per gallon (MPG) than other cars?

**Data and Model**

This data originally came from StatLib Library at Carnegie Mellon University, and was used by the American Statistical Association in 1983 for exposition. The data was retrieved from the UCI Machine Learning Repository. UCI removed 8 instances from the original data due to missing “MPG” values, and I removed 6 instances due to missing “Horsepower” values.

**Linear** **Regression** **Specification**

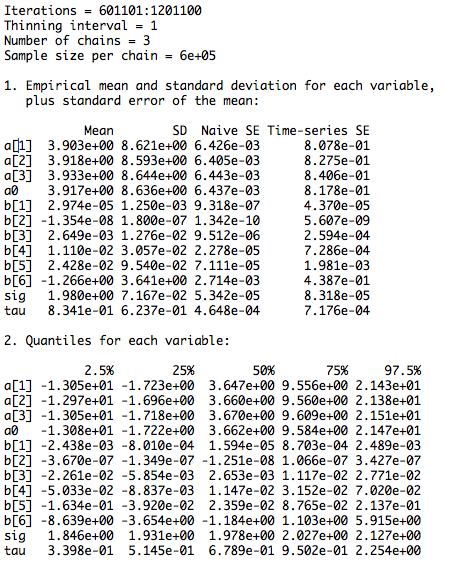
I assumed that the samples are from a normal population and that they satisfied the basic linear regression assumptions. I transformed the MPG variable to the log scale.



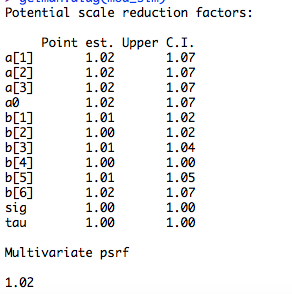
All coefficients are significant at alpha=0.05, but “HorsePower” is insignificant. I retained the “HorsePower” variable because I had an interaction term of HorsePower and Acceleration. The coefficients also have the expected signs. For example, the weight coefficient is negative, which is consistent with the fact that a heavier car has lower MPG.

**Bayesian** **Model** **Specification**

Where j is the index of countries/continent (America, Japan and Europe). To answer the question (does American cars on average have lower miles per gallon (MPG) than other cars?), I used a random intercept model, and corresponds to the respective intercepts. If American cars consume more fuel on average, we expect a[1] to be less than a[2] and a[3]. As the table be shows, a[1] is the least of the intercepts.

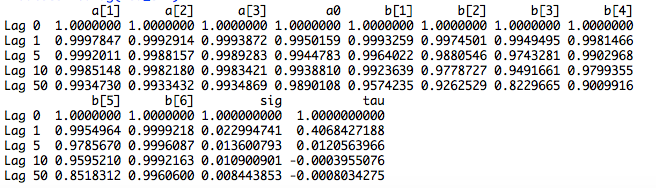


The potential scale reduction factors (using the gelman diagnostics) look good. They are all nearly 1.

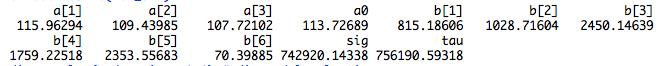


Autocorrelation

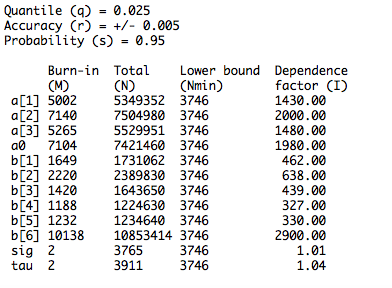
There is strong indication of autocorrelation between the lags, with the exception of the sig and tau parameters. The autocorrelation between the 0th lag and the 50th lag is nearly 1 for most of the parameters.



Effective sample size for most of the parameters, excluding sig and tau is very low considering the large number of sample size (500,000). This is also in agreement with the autocorrelation diagnostic. To increase the independent sample size or to better estimate some of the parameters we need more iterations (perhaps around 1 to 2 million). Unfortunately, I couldn’t perform such high number of iterations. I tried a million iterations and my computer crashed.



The Raferty and Lewis diagnosis shows that the model requires high number of independent samples to estimate 95% of the distribution with in an accuracy of +- 0.005. Interestingly, b[6] (the interaction of horsepower and Acceleration) needs nearly 10.8 million independent sample size. The below table is only for the first chain. However, all of the chains are similar.



**Conclusion**

Considering all of the diagnostic results, the model is on average good. However, more iterations need to be performed to ensure that the simulation converge to the targeted distributions.