UNIVERSITY OF TEXAS AT SAN ANTONIO

The Analysis of Car Emissions Data using Bayesian Modeling Averaging with Comparison to Frequentist Methods

Liang Jing, Bazoumana Kone, and Chris Louden

Dec. 2008

1 INTRODUCTION

One of the side effects of the internal combustion engine is the emission of pollutants such as carbon dioxide, particulates and nitrous oxides. In 2001, the city of Oslo, Norway was concerned with the levels of pollution in the neighborhood of Alnabru.

This data set consists of 500 observations of nitrous oxide levels in the neighborhood of Alnabru from October 2001 to August 2003. Seven other measurements were also taken as shown in the below, and the scatter plots of paired variables are shown in Figure 1.1.

X_1 : Number of cars per hour

X_2 : Temperature 2 meter above ground (degree C)

X_3 : Wind speed (meters/second)

X_4 : Temperature difference between 2 meters and 25 (degree C)

X_5 : Wind direction (degrees)

X_6 : Hour of day

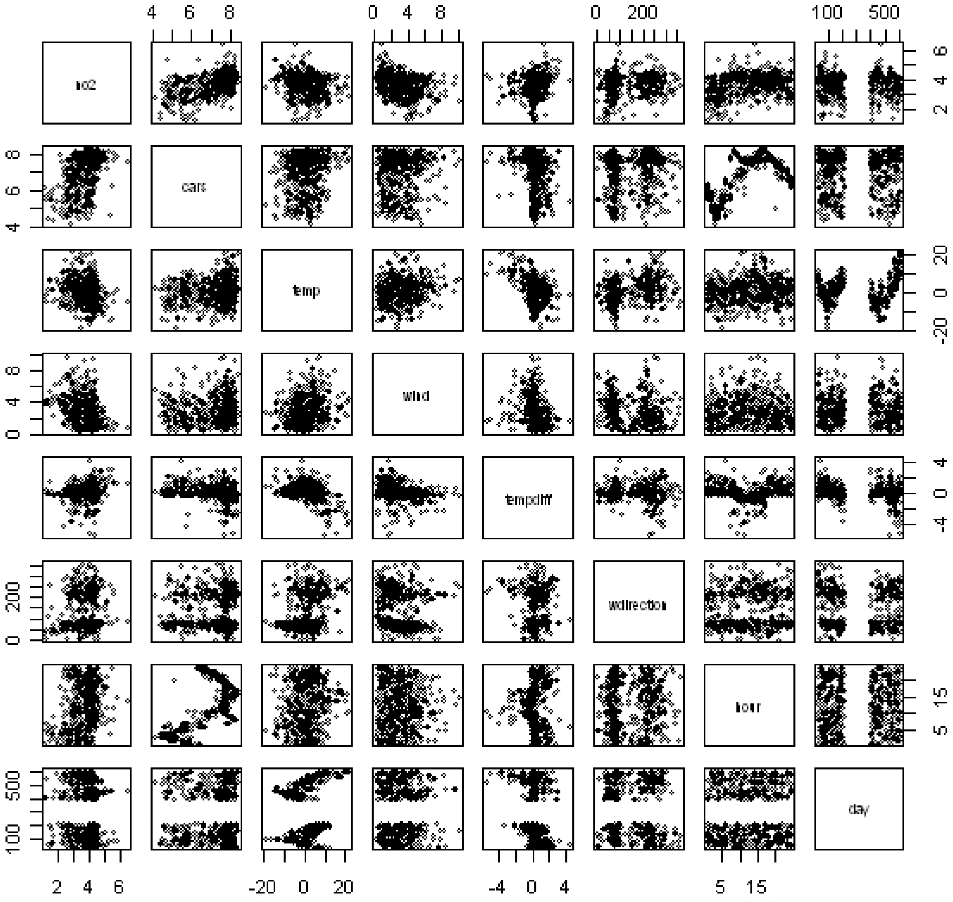
X_7 : Day number

This data was analyzed in many ways using both Bayesian and Frequentist methods. Each analysis was then compared against the others to see which one has the better predictive power. The first goal was to compare Bayesian Model Averaging to more traditional Bayesian approaches to modeling.

2 BAYESIAN MODEL AVERAGING

One of the foremost jobs of a statistician is to model data. When one models data, one must strike a balance between creating a model that fits the existing data well but that is not overfitted to the point that the predictive value of the model is nil. Furthermore, it may be possible for the statistician to develop two models which fit the data appropriately, but which vary wildly in their predictions. Thus the question arises, which model should the statistician pick?

Figure 1.1: Histogram of Nitrogen dioxide levels



Historically, this question has been answered by looking at various statistics that describe the model. For example, the Bayesian Information Criterion (BIC), describes how well a model matches the data. What if, however, two different models had a similar BIC but were very different in prediction. Bayesian Model Averaging (BMA) provides a way to avoid such a problem (Hoeting et al., 1999).

The idea behind BMA is to consider the posterior probability of some quantity of interest given the data. That probability is the sum of the probabilities of the quantity given the data and that a certain model is used weighted by the probability of that model being the correct one given the data. More generally, in the case of model selection, one may simply look at the posterior probability that a model is the correct model.

Formally, let Δ be a quantity of interest. The posterior probability of Δ given the data is

$$\Pr(\Delta|D) = \sum_{k=1}^{K} \Pr(\Delta|M_k, D)\Pr(M_k|D) \tag{2.1}$$

where $M_1, M_2, ..., M_k$ are the possible models for this data, and D is the data itself. The posterior probability of the model given the data is

$$\Pr(M_k|D) = \frac{\Pr(D|M_k)\Pr(M_k)}{\sum_{l=1}^{K}\Pr(D|M_l)\Pr(M_l)} \tag{2.2}$$

where

$$\Pr(D|M_k) = \int_{\Theta_k} \Pr(D|\boldsymbol{\theta}_k, M_k)\Pr(\boldsymbol{\theta}_k|M_k)\,\mathrm{d}\boldsymbol{\theta}_k \tag{2.3}$$

and $\boldsymbol{\theta}_k$ is a vector of parameters for the model M_k.

As Hoeting et al. mentioned, there are many problems associated with preforming this analysis. One of the largest ones is the size of the set of all possible models, $\mathscr{M}$. Thus the summations involved can become quite unwieldy. There are work around to avoid this, such as considering only a subset of $\mathscr{M}$ that contains only the most plausible models.

Such considerations are made in the software available to preform BMA. Chris Volensky maintains a collection of software written to preform this type of analysis. The software used in this paper is a package, BMA, that preforms this analysis in R.

3 Analysis

The first step to modeling the data is to try and divine what distribution it follows. A histogram of the data, Figure 3.1, shows that the data is skewed to the right. Therefore it may be necessary to consider more complicated models than simple linear regression such as a generalized linear model (GLM). These models were considered from both a Frequentist and a Bayesian point of view to compare the two methods. In addition, BMA allows one to find the models with the highest posterior probability given the data. A model that was the average of the top five models weighted by their probability was also studied. Figure 3.2 displays the residuals of this model along with those of the best model. It can be seen that, for this data, there is little difference in the residuals.

Figure 3.1: Histogram of Nitrogen dioxide levels

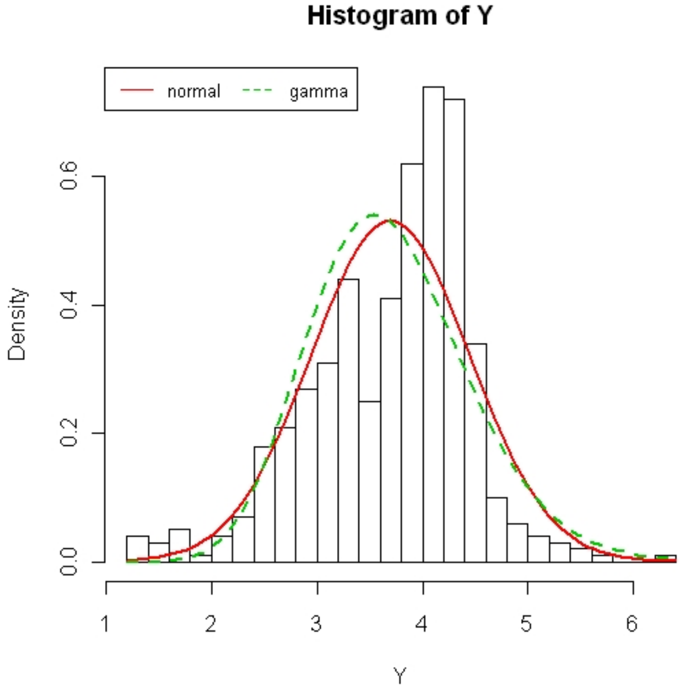
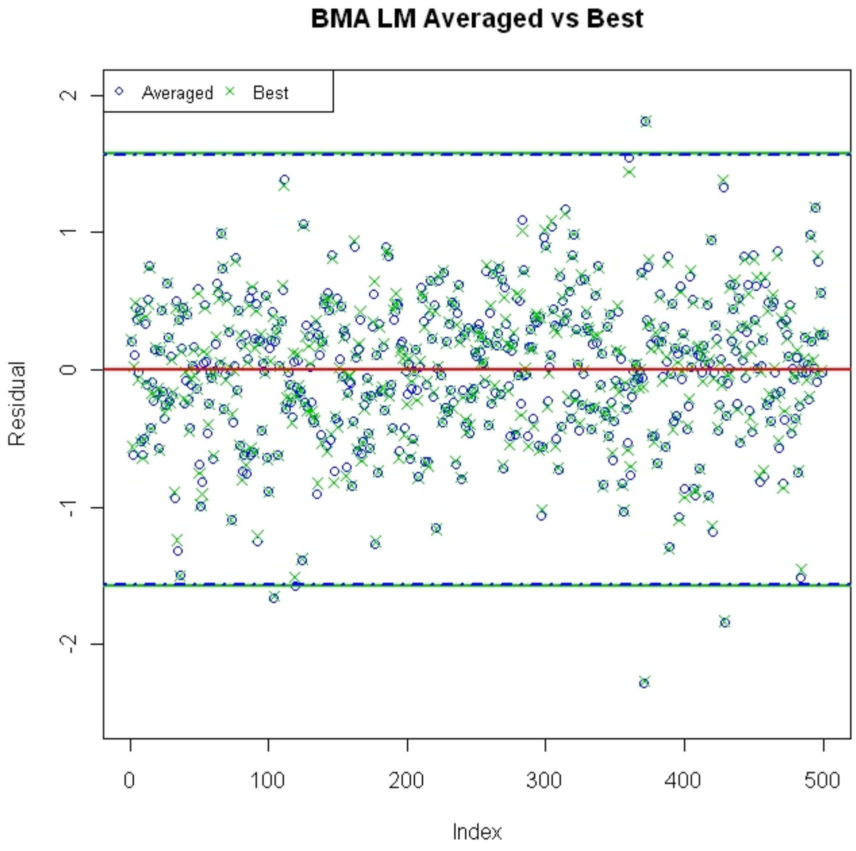


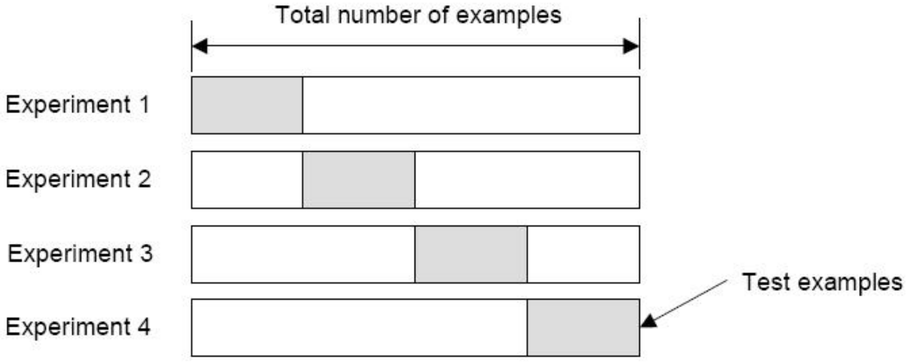
Figure 3.2: Residuals for the Highest Probability Model (Best) vs the Average of the top 5



The package BMA allows the user to create both linear models and generalized linear models. For this data, the generalized linear models assumed that the data followed a gamma distribution. For the Frequentist analysis, the values of α and β were computed from the data using the sample mean and variance. In both the linear regression and the GLM, the software package R was used. For the Bayesian method, in both the linear modal and the GLM cases, a non-informative prior was used.

To test how well each model fits the data, each model was used to predict new observations. To do this, the data was split into a training and a test set. The training set was used to find the coefficients in the model and that model was used to predict what the values in the test set would be. For the model that was the average of the top five models, 50-fold cross-validation was used. The data was partitioned into 50 sets. At each step one was held out as the test set and the remainder was used as a training set. The model derived was then used to predict the values in the withheld set. This was repeated for each of the remaining 49 sets.

Figure 3.3: The procedure of k fold cross-validation



To test the accuracy of the prediction, two values were considered: the mean absolute error and the mean squared error. These were computed by the following two formulas:

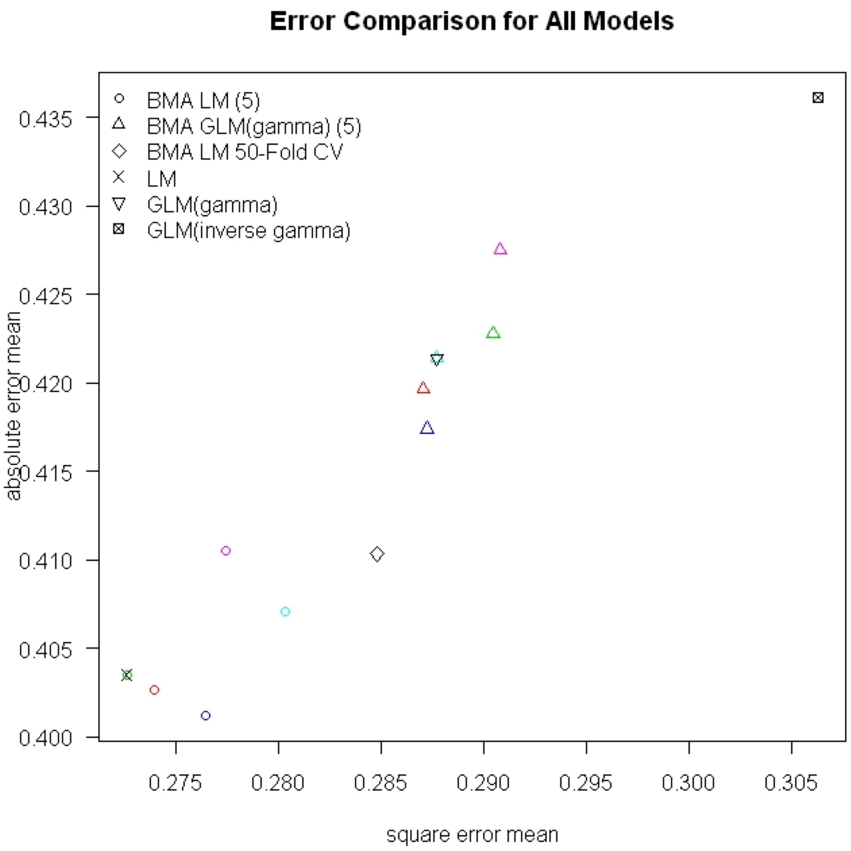
$$e_j = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_{ij}| \tag{3.1}$$

$$s_j = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_{ij})^2 \tag{3.2}$$

where e_j is the mean absolute error and s_j is the mean squared error for model j. In models with a better predictive power, these numbers will be low.

These two values, shown for each model in Figure 3, show two things about the modeling procedures used. First, the linear models had lower values for both e_j and s_j. This implies that they are a better fit for the

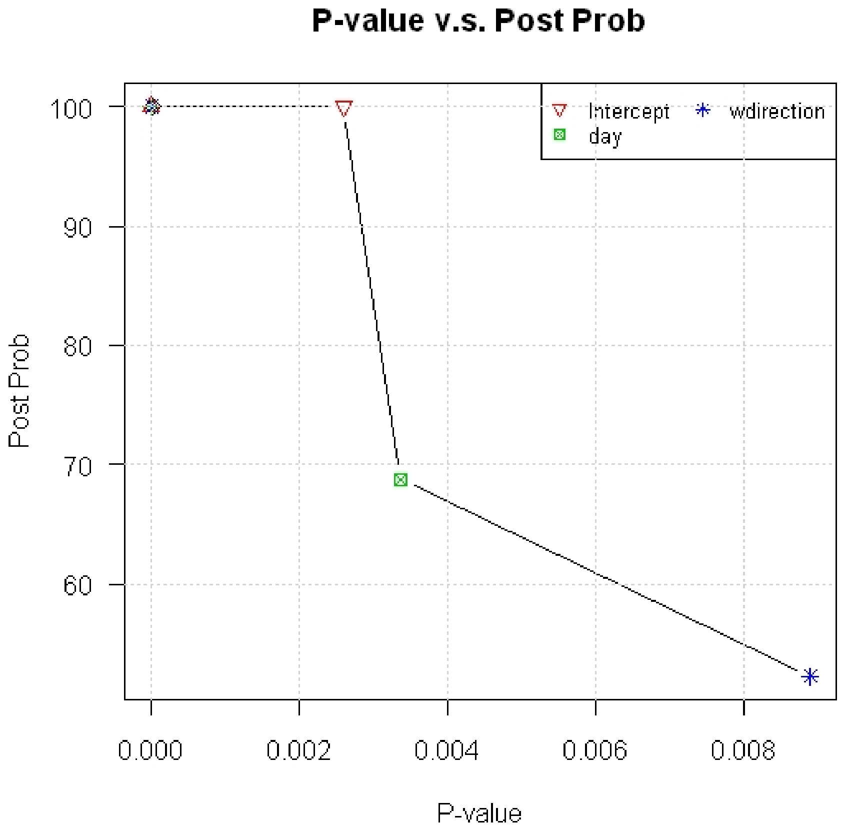
Figure 3.4: Absolute Error Rate vs. Square Error Rate for all Models



data. This is likely due to the choice of distribution for the GLM. Secondly, Bayesian methods and Frequentist methods have nearly the same results. This is likely due to the nature of the data as in general, this may not be so, especially with the use of informative priors.

There was one difference between the the Bayesian and Frequentist analysis. In the Frequentist analysis, one uses the p-value that a coefficient is zero to tell if that coefficient should be in the model. BMA calculates the probability that that coefficient is in the model. This is done by summing the posterior probabilities of the models that contain that coefficient. Comparing the two values, Figure 3, shows that while the Frequentist may be quite certain that both Day and Wind Direction should be in the model since they both have p-values less than 0.01, the Bayesian may be much less sure. The fact that the posterior probabilities are less than 0.70 puts the seed of doubt into the analysis. This is especially so with Wind Direction at just better than half.

Figure 3.5: Posterior Probability versus P-Value



4 Conclusion

Bayesian Model Averaging provides a way to help with model selection. For this data set, however, BMA did not provide improvement in comparison with Frequentist methods. Furthermore, the use of generalized linear models preformed worse in prediction than linear models, despite the fact that the data is skewed to the right. Combining the best models from the BMA routine weighted by their probabilities also did not improve prediction. Overall, for this data set, BMA did not offer much in the way of improvement versus other methods that are known. It does, however, provide an important tool for the statistician to have as well as granting insight into a different way of looking at modeling.

Reference

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999).

Bayesian model averaging: A tutorial. Statistical Science, 14(4).