Data Analysis Project Report

P. Y. Chou 2/17/2021

Abstract

In this project, we aim to find out the relationship between the acceleration of a car versus several predictors. We use linear regression to fit several combinations of predictors, including additional transformed predictors, and pick the model that yields normal partial residual plots and the lowest DIC. Finally, we discuss the fitted model and draw the conclusion.

Introduction

In this project, we want to identify the relationship of different vehicles' acceleration (ACC) and predictors including weight-to-horsepower ratio (WHP), the speed at which they were traveling (SP), and the grade G (G=0 indicates the road is horizontal.)

Data

We will use the data set which contains all the information mentioned previously. Unfortunately, I don't know how the data were collected. But you can rest assured that this data set is not arbitrary, as it is used in one of the assignments in my linear regression class.

Acceleration is a well-studied physical property, and it follows a simple equation $\mathrm{F} = \mathrm{ma}$. Therefore, the ratio $\frac{\mathrm{F}}{\mathrm{m}}$ is a crucial variable. Indeed, the weight-to-horsepower ratio, which is inversely proportionate to $\frac{\mathrm{F}}{\mathrm{m}}$, is included in the data set. Also, it is intuitive that the faster a car goes, the harder it accelerates. Hence, it is necessary to include traveling speed. Lastly, we need to account for the degree of slope of the road for calculating the effective power for acceleration. This information is included in the variable G.

The data set contains 50 observations without any missing value. According to Figure 1, we can observe that:

1) WHP takes on 4 real values, and G takes on 3 integer values.
2) ACC has a somewhat negative relation with WHP and SP.
3) The variance of ACC seems to decrease along with WHP and SP.
4) SP has little correlation with WHP.
5) The predictor G has little correlation with other factors and observations.

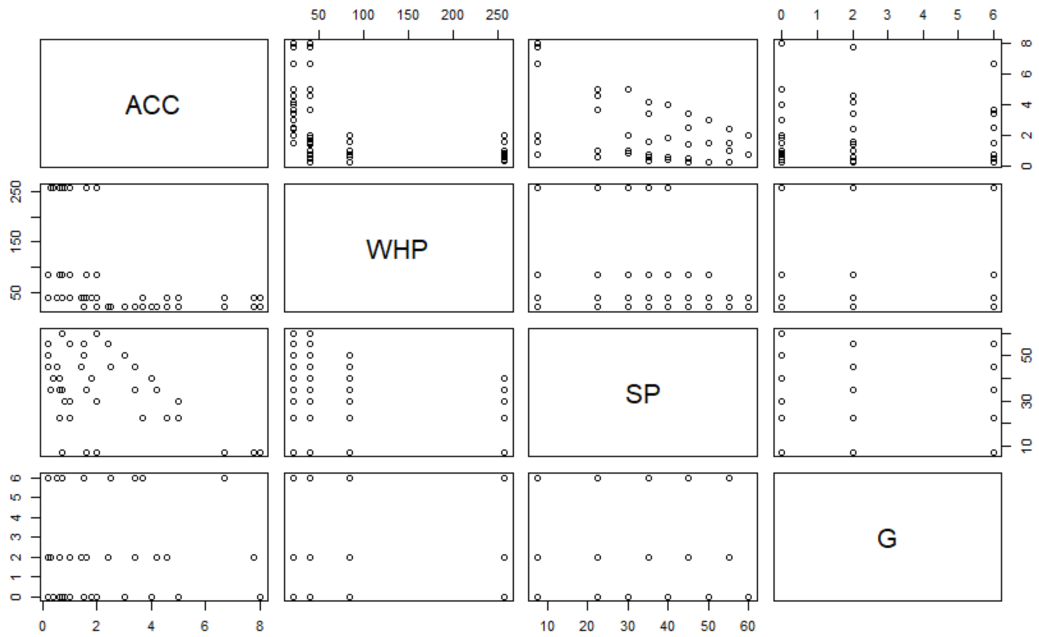

Figure 1

Model

We will fit a linear model for the responses and predictors for they are all continuous values. Since WHP and G takes on a small set of values, one may fit separate models where they are treats as discrete factors.

Firstly, we fit a naïve model with non-informative prior and see how it does:

$$y_i|X_i, \beta, \sigma^2 \sim ind\ N(X_i\beta + \beta_0, \sigma^2), i = 1, \ldots, 50$$
$$\beta_j \sim iid\ N(0, 1e6), j = 0,1,2,3$$
$$\sigma^2 \sim inverse\ Gamma(1/2, 1 * 1500/2)$$

where the column of X includes WHP, SP, and G

For the JAGS setup, we skip manual initialization, use 1000 iterations as burn-in, and run 3 chains for 5000 steps. According to the convergence analysis, the parameter indeed converges. Because the autocorrelation is a little big for β_0, I increase the iterations to 10000 for each chain so that the total number of iterations is greater than that suggested by Raftery's diagnostic. The MCMC yields the posterior mean of the parameters as:

```
       b[1]        b[2]        b[3]          b0
-0.01830005 -0.09379089 -0.15487931  7.20113606
```

The result is expected for b[1] and b[2] which are the coefficients for WHP and SP.

Although b[3] appears to be negative, the posterior distribution (Figure 2) shows that the predictor G is actually not that significant. We also check the partial residual plots using the mean of the posteriors. In Figure 3, the partial residual of WHP seems to exhibit a quadratic trend. It will be a good idea to include additional transformation of the predictors. Before fitting another model, we calculate the DIC to be 281.5.

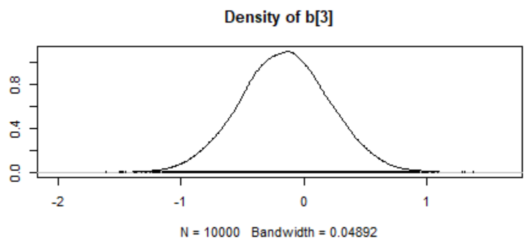

Figure 2

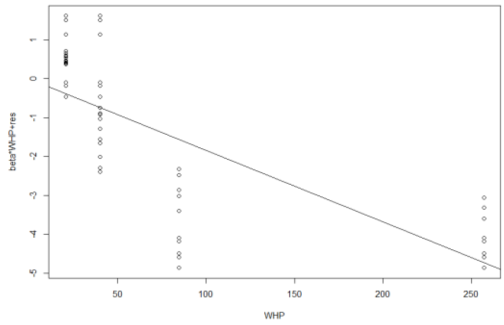

Figure 3

In the second model, we include the WHP^2 as an additional predictor. We fit the second model using the same prior and JAGS setup as the first model. The resulting autocorrelation is quite high. Thus, we increase the iteration of each chain to 100000. The posterior mean of the predictors is:

```
         b[1]          b[2]          b[3]          b[4]           b0
-0.0969762749 -0.1006402149 -0.2248987218  0.0002753742 10.1647473566
```

where b[4] is the coefficient for WHP^2. It turns out all of the predictors are significant in this model, and the new DIC is 280.2. Since the DIC for the second model is lower, we prefer the second model. Most importantly, we can check the partial residual plot for WHP (Figure 4); the quadratic trend disappears.

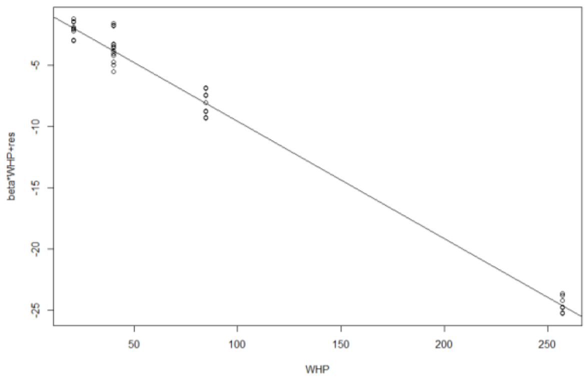

Figure 4

Predictors	**DIC**
WHP SP G WHP^2	280.2
WHP SP G	281.5
WHP SP WHP^2	**277.9**
SP G WHP^2	282.8
WHP SP	279.2
WHP WHP^2	281.9
SP WHP^2	280.1

Table 1

We also try several combinations of predictors with non-informative priors. The result is listed in Table 1. Surprisingly, the model using predictors of WHP, SP, and WHP^2 yields the smallest DIC. Therefore, we are justified to drop the predictor G, and we refer to the new model as -G.

The mean of posterior of the coefficients for model -G is shown below.

```
        b[1]          b[2]          b[3]           b0
-0.0926107429 -0.0962870695  0.0002640494  9.3641944312
```

The sign of the coefficient is the same for the first model and the second model. We can further confirm our belief by assuming a skeptical prior, i.e., assigning positive means to the priors, and see the result of the posteriors.

$$y_i|X_i, \beta, \sigma^2 \sim ind\ N(X_i\beta + \beta_0, \sigma^2), i = 1, \dots, 50$$
$$\beta_j \sim iid\ N(10, 1e6), j = 0,1,2,3$$
$$\sigma^2 \sim inverse\ Gamma(1/2, 1 * 1500/2)$$

where the column of X includes WHP, SP, and WHP^2

The sign of the posterior mean remains the same as shown below.

```
        b[1]          b[2]          b[3]           b0
-0.0903690998 -0.0958593963  0.0002563872  9.2744615549
```

Results

According to our analysis, the best model to describe the relationship between the acceleration of a car and the effect of weight-to-horsepower ratio and traveling speed is: $\text{ACC} = 9.2744616 - 0.0903691^*\text{WHP} + 0.0002563^*\text{WHP}^2 - 0.0958593^*\text{SP}$ The negative relation between WHP and ACC is because the larger the WHP is (more weight, less horsepower), the harder it accelerates. Also, the higher a car travels, the harder it accelerates. This may be caused by the air drag and the limitation of the engine. The positive effect of WHP^2 is a bit difficult to explain. It serves as a correction for the WHP, as we can see the partial residual plot for WHP is no longer showing specific patterns. Also, according to DIC, we prefer to not include the predictor G in the model, which is not a surprise to us for we already identify that G is not so related to the response. In theory, as claimed in Data section, it is necessary to account for the degree of the slope. The problem may lie in the way we interpret the grade G. If G=0 means 0 degree, then G only varies around 0 to 6 degree, and it is reasonable to drop G since the difference of the effective horsepower is negligible. But if G is meant differently, then dropping the predictor needs further justification, because it might be against some basic law of physics.

Conclusions

According to our extensive analysis, we can confidently conclude that there is a negative relationship between the acceleration of a car and the weight-to-horsepower ratio, and between the acceleration of a car and the speed at which it travels.