Analyzing the Price of Mazda Cars by age Using Bayesian Linear Regression

Derek Albosta

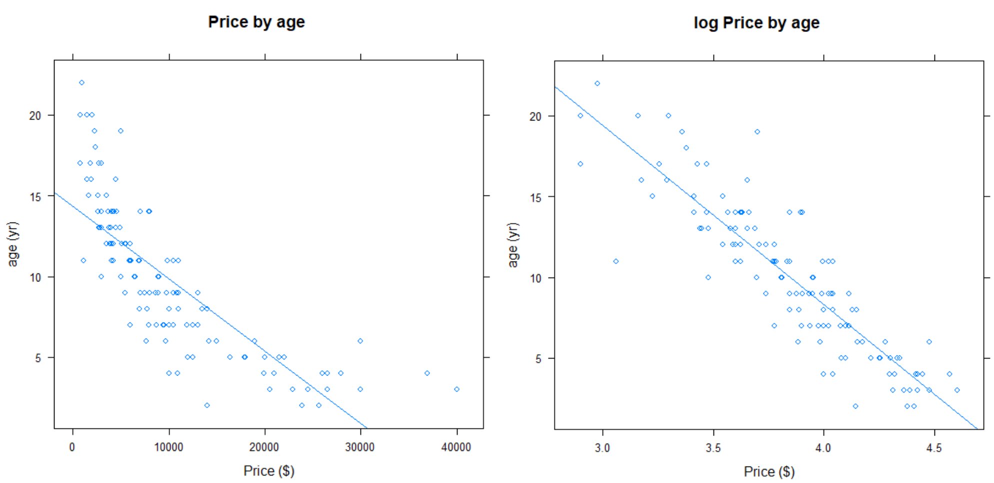
April 15, 2020

1 Introduction

Data obtained in 1993 contains the age of Mazda cars along with their prices[1]. The data set contains 124 Mazda cars and the prices each car was purchased at in 1993. The age range of the cars runs from 1971-1991. When working with the dataset, an age variable was appended to the dataset by subtracting the purchase date (1993) by the year the car was made. Using a Bayesian linear regression, we will analyze the impact that age has on the price of Mazda cars.

2 The Data

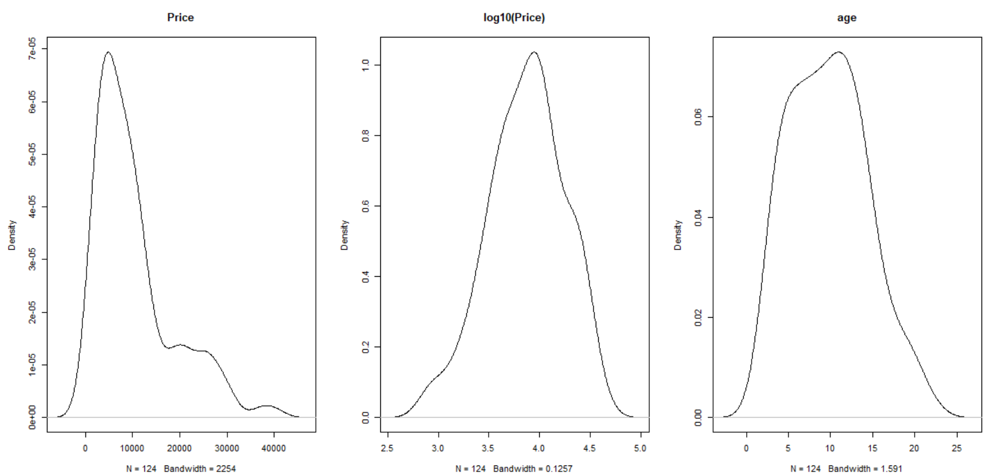
In order to see the relationship of age on the price of the cars, we first built a simple ab-line over our predictor and predicted variable. Below are results of an ab-line fitted over price by age and the log transform of price by age:



AB-line fitted on Price and log Price by Age

As we can see, our relationship of the variables of the non transformed graph show a clear curvature along our predictor variable age. Because of this, we have transformed our price variable. Our transformed figure on the left show a tighter linear relationship of our two variables and suggests that we should transform our price variable. We can further reinforce this notion by plotting our variables in a density plot to see if they are normally distributed. Density plots of price, logPrice, and age are shown below:

[1]Melbourne Age, various issues during 1991.



Density plot of variables of interest

As we can see, the non-transformed price is not normally distributed but our log transformed version is. Therefore, we will be using the log transformed version of price in our linear model.

3 The Model

Our model is concerned with finding the relation ship between our log-transformed price variable by using age. We also want to see the variance along our linear model using a t-distribution to keep our model robust to outliers. Below is a model diagram depicting the structure of the model we will use along with its various parameters:

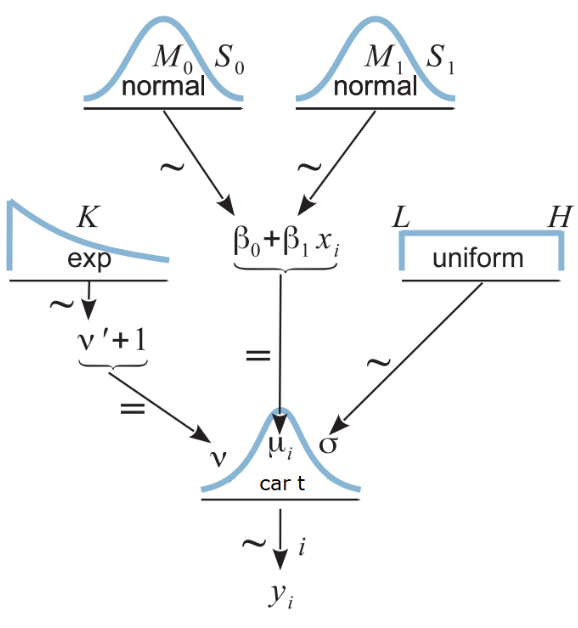


Diagram depicting model hierarchy

At the bottom, y_i represents a value sampled from a t-distribution around the central tendency μ_i. This μ_i is defined as $\beta_0 + \beta_1 x_i$ where β_0 is our y-intercept and β_1 is our slope parameter multiplied by age. Each of the other parameters are shown to be sampled from their corresponding prior distributions.

3.1 Prior distribution parameters

Because we do not have any prior information about the relationship of these two variables, we are using a set of vague priors to define our initial distributions. Our intercept and slope parameters are sampled for a normally distributed prior with a vague tendency on the groups. Our v parameter is sampled from a broad exponential prior and our σ parameter is given a noncommittal uniform prior. It is important to note that we also standardize the data before passing it to our model in order to improve performance. The priors for our model are defined as below:

```
# Priors vague on standardized scale:
    zbeta0 ~ dnorm( 0 , 1/(10)^2 )
    zbeta1 ~ dnorm( 0 , 1/(10)^2 )
    zsigma ~ dunif( 1.0E-3 , 1.0E+3 )
    nu ~ dexp(1/30.0)
```

4 The Posterior

4.1 MCMC Diagnostics

We can ensure the accuracy of our posterior distribution due to our MCMC diagnostics. Each of our chains' metrics satisfy the assumptions which provide evidence showing that the chains converged correctly. The chains' all showed random behavior during the run period. All of our chains' auto-correlation and shrink factors converge towards 1 and 0 respectively and our density plots show tightly clustered distributions. If you would like see the diagnostic maps, please refer to the code.

4.2 Establishing ROPE parameters

There is not that much utility that we can gather by using ROPE parameters for this model. Given that it is a linear regression with vague priors, there is not any region of interest we are looking to test against for our parameters.

4.3 Simulation results

Below are the accompanying graphs generated by our MCMC simulation:

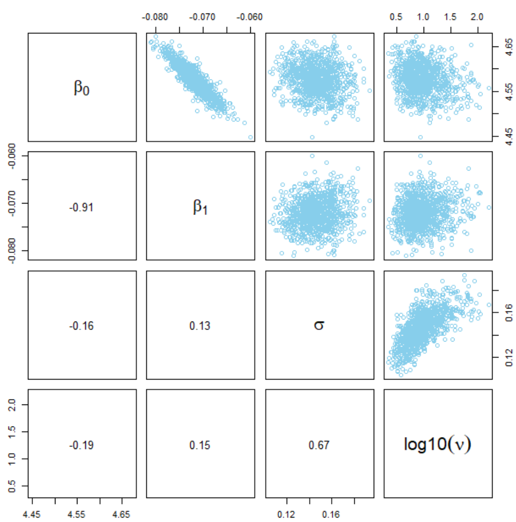


Figure 1: Correlation matrix of parameters

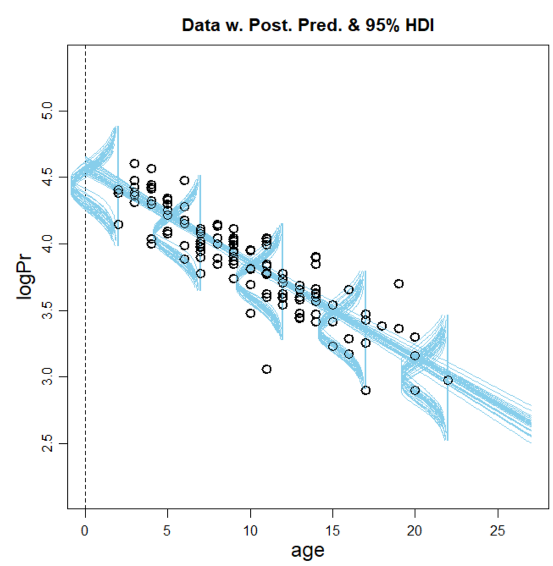


Figure 2: Linear Reg w/ t-noise distributions

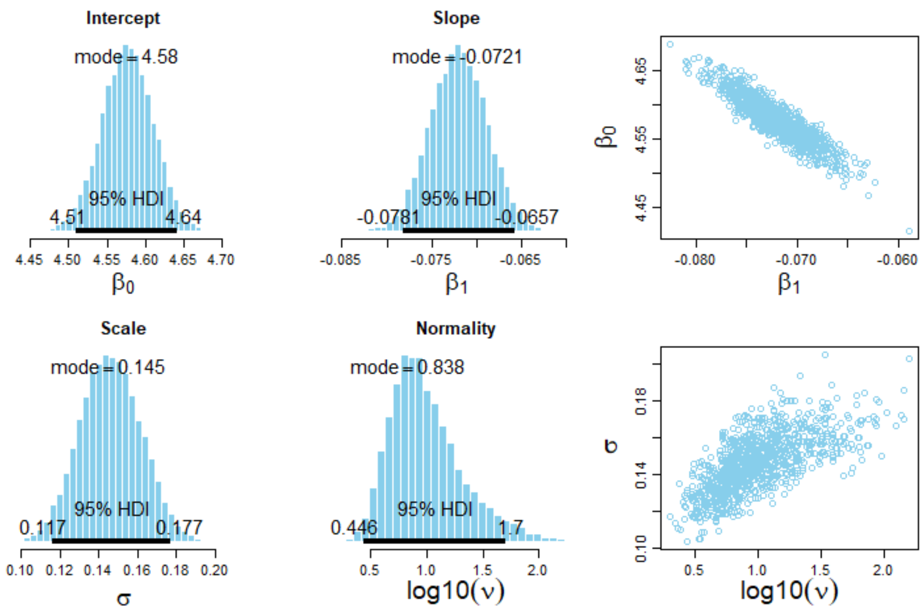


Figure 3: Posterior distributions of parameters

From our resulting posterior distributions, we have a credible range of values for our y-intercept from 4.51 to 4.65 with the mode being 4.58. As for our slope parameter, it is estimated to be in 95% HDI range between -0.0781 to -0.0657 with the mode value being -0.071. As for our variance along our linear model, our t-distribution parameters tells use that the scale is credible among the values between 0.115 to 0.177 with the mode being 0.145 and our normality v being credible within the ranges of 0.446 to 1.7 with a mode of 0.838. These t-distribution parameters suggest that along our linear model, the data is in general close to our ab-line but it is accounting for a lot of outliers since our normality parameter is small. Since this model was done on a log transform of price, these parameters do not have any meaningful interpretation. Using the mode values of our parameters, we can transform our price variable back to its normal scale and find the modal price for the age of a given car. Below bar plot illustrating the median price of cars aged 0(new) to 23 years:

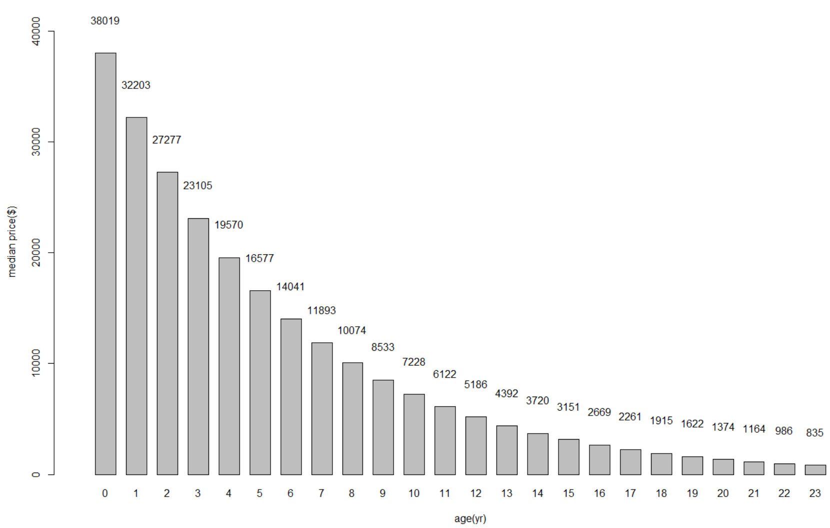


Figure 4: Median price model by age

As we can see, the price of a car depreciates over time on a logarithmic scale. A brand new Mazda will run a price of around $38,000 while it depreciates at a rate of $(10^{\beta_1})^A = (10^{-0.0721})^A$ where A is the age of the car in years.

5 Conclusion

Based on our results, it is evident that age depreciates the value of a Mazda car on a logarithmic scale. We now know from our analysis that the depreciation rate of a car is a function of age in the equation $(10^{-0.0721})^{age}$. In general, a brand new Mazda is worth around the $38,000 price range. To conclude, this shows that Mazda cars are a bad investment.

References

[1] Melbourne Age, various issues during 1991.