Predicting Automatic v Manual Car Transmission with Bayesian Logistic Regression
By Iman Sellars

Motivation

The main motivation behind this analysis is to gain a better understanding of which factors make a car more likely to have a manual versus an automatic transmission. This is a classification problem that offers a setting to apply Bayesian logistic regression. The mtcars dataset has a rich set of vehicle characteristics, making it ideal for studying transmission type. The Bayesian framework allows us to quantify uncertainty and perform feature selection.

Data and Methods

32 cars and 10 standardized predictor variables:

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

The goal is to model binary response using Bayesian logistic regression with stochastic search variable selection (SSVS). This approach uses spike-and-slab priors on the regression coefficients, so the model can estimate effect sizes and the probability that a predictor should be included in the model. Posterior inference is done using MCMC sampling in JAGS, producing coefficient estimates, 95% credible intervals, inclusion probabilities, and model-based predictions of transmission probabilities for each car. Traceplots, posterior density plots, and Gelman-Rubin statistics were generated as diagnostics for this model.

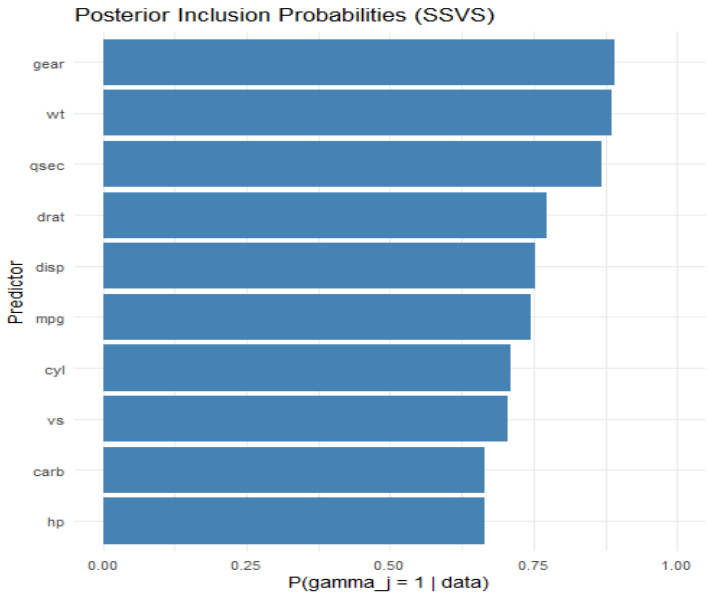
MODEL EQUATIONS ~
This model assumes that each car's transmission type (manual vs automatic) follows a Bernoulli likelihood, where the probability of a manual transmission is determined by a logistic regression.

Likelihood
$\text{logit}(P(\text{manual}))=\beta 0 + \beta 1(hp) + \beta 2(wt) + \beta 3(mpg) + \ldots$
Automatic ~Bernoulli(pi)
Now this is standard Bayesian logistic regression.

Each coefficient receives a weakly informative normal prior and reflects prior uncertainty.
Priors:
$Bj \sim N(0,10^2)$

Major Findings

In this bayesian logistic regression with SSVS, several useful insights were drawn about the characteristics of cars associated with manual transmission. Posterior Inclusion probabilities show that features gear, weight, and quartermile time These 3 predictors had inclusion probabilities greater than 0.86, and therefore should definitely be used for the model. There was a second class of predictors drat, disp, and mpg. These predictors output inclusion probabilities greater than 0.74, and do not contribute to the posterior as much as the previously mentioned predictions, but they could still be useful towards the model. Finally there was a third class of predictors that reported inclusion probabilities less than 0.72. Cyl, vs, carb, and hp. Furthermore, coefficient estimates also reveal how important a predictor is. Our estimates reveal that the majority of the predicting power is found in the same three predictors that had the highest inclusion probabilities, giving us further assurance that these predictors are important for capturing effects in the model.



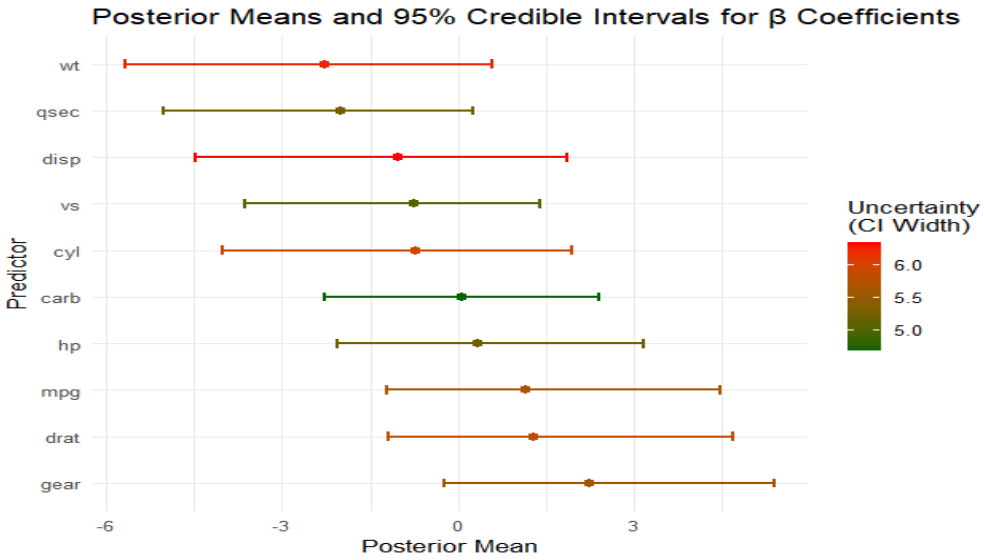
beta_summary

Predictor	mean	sd	lower95	upper95
beta[1]	1.1037510	1.466970	-1.222525	4.346829
beta[2]	0.3284025	1.219475	-2.059414	4.373143
beta[3]	0.3284055	1.758550	-5.745621	0.579473
beta[4]	-2.3394255	1.584980	-4.435959	1.804255
beta[5]	-0.9651414	1.164797	-1.1045996	-4.502729
beta[6]	-0.8144694	1.538990	-4.2517562	1.871302
beta[7]	-0.81414694	-0.81414	-4.2517562	-0.209733
beta[8]	2.35475900	2.347590	-0.209732	-2.341634
beta[9]	-0.71605975	-0.7160575	-3.6814891	3.681481
beta[10]	-0.71605975	-0.1605975	-3.6814891	4.494532

Based on the table below, we see manual transmissions consistently show high predicted probabilities, while automatic transmissions show low probabilities, showing that the model is successfully distinguishing between automatic and manual transmissions.

```
Car Observed PredictedProb
1          Mazda RX4        1   0.933533333
2      Mazda RX4 Wag        1   0.856866667
3         Datsun 710        1   0.826666667
4     Hornet 4 Drive        0   0.006533333
5  Hornet Sportabout        0   0.030600000
6            Valiant        0   0.001200000
```

95% credible intervals show us the general uncertainty surrounding each predictor. This measurement of uncertainty is not usually considered in regular logistic regression, which gives us more assurance about the quality of our model.



Overall, the Bayesian analysis identified which features carry the strongest evidence of influence, quantified uncertainty through credible intervals, and provided a framework for variable selection and prediction. The Bayesian model produced excellent separation between manual and automatic vehicles and can be considered for model deployment predicting the transmission of new unseen cars.

Implications

By quantifying variable importance through posterior inclusion probabilities and assessing effect uncertainty with credible intervals, the model offers a probabilistic understanding of each predictor's contribution to the estimates. Therefore we can provide meaningful insight into the factors that influence whether a car is equipped with a manual transmission. From these findings we can infer certain vehicle characteristics, most notably number of forward gears, weight, and quartermile time.

Further Questions Raised

The dataset is small (n = 32), which makes uncertainty higher than usual. How stable would these findings be with more vehicles or with modern car attributes? Also, some predictors such as weight and horsepower are related to others such as cylinders and displacement. This means we possibly have multicollinearity. Future models can use variable grouping to avoid this and perform more accurate predictions.