Car Mileage Prediction Model

Karthik Arumugham

29 April 2016

Executive Summary

The `mtcars` dataset, extracted from 1974 Motor Trend magazine, comprise mileage in miles per gallon (MPG), and ten aspects of automobile design and performance for 32 automobiles (1973-74 models). The regression model attempts to predict gasoline mileage from the other variables. In particular, the analysis attempts to determine whether an automatic or manual transmission is better for MPG, and quantifies the MPG difference. On the basis of the analyses, regression of MPG on TRANSMISSION TYPE, WEIGHT, and QUARTER MILE TIME may be best for prediction. Manual transmission delivers higher mileage by a difference of 2.93 compared to automatic transmission, provided other specs are equal.

Data processing and transformation

Load the `mtcars` dataset and necessary libraries. The description of variables can be found in the mtcars help file. Convert variables `am` and `vs` to factor variables as they are categorical. vs: V/S (0 = V-engine, 1 = straight engine). am : Transmission (0 = automatic, 1 = manual)

Exploratory Data Analysis

1. **Appendix: Plot 1** - The `mpg` outcome variable follows a normal distribution.
2. **Appendix: Plot 2** - The `mpg` outcome variable has a non-linear relationship with important variables such as `disp`, `wt`, `hp` and `qsec`.
3. **Appendix: Plot 3** - `wt` has strong correlation with all variables except `qsec`. So possibly a larger variation in `mpg` can be explained by `wt` and `qsec`.
4. **Appendix: Plot 4**

- 4A: There is a clear separation on the mileage delivered by automatic and manual transmission.
- 4B: There is a clear separation on the mileage delivered by `V-Engine` and `Straight Engine`, with straight engines delivering higher mileage. However this is counter intuitive, as v-engines are much more compact, lighter and has better fuel efficiency. Possibly this may be due to higher influence by displacement and weight.
- 4C: There is a clear separation on the mileage delivered by cylinder types.
- 4D: There is no clear clear separation on the mileage delivered by different gear types. However this needs to be investigated further for any confounding variable.

5. **Appendix: Plot 5**

- 5A: Automatic transmission vehicles are heavier than manual and deliver lower mileage.
- 5B: 8 cylinder engines are heavier than 6, which in turn is heavier than 4 and deliver lowest mileage.
- 5C: V-Engines are faster than straight engines (i.e., lower qsec time), but deliver lower mileage. Since v-engines are mostly 8 cylinders and weight heavier, unlike straight engines which are mostly 4 cylinders.
- 5D: There is no clear seperation for cylinders in terms of qsec time. This may possibly be attributed due to differences in displacement volume attributing to variances in performance. However there is a clear seperation in mileage based on cylinders.

Regression Analysis

A quick stepwise regression with backward elimination returns a model with predictors `am`, `wt` and `qsec`. Linear regression for the suggested model is run with stepwise inclusion of the variables.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## am1          2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

The final model has a high adjusted R^2 of 0.834. All the coefficients except the intercept are significant ($p<0.05$) and the null hypothesis that the beta coefficients are 0 is rejected as there is no 0 in the 95% confidence interval. The intercept is 0. This implies manual tranmission cars deliver a higher mileage by a difference of 2.93 than automatic transmission cars, provided other variables remain constant.

From the anova test, going from first to second, and second to third models result in significant reductions in RSS and hence result in a better model fit.

Regression Diagnostics

1.**Residuals** (Refer **Appendix: Plot 6**)

- **Residuals vs. Fitted Plot** shows no consistent pattern, supporting the accuracy of independence assumption.
- **Normal Q-Q Plot** indicates that the residuals are normally distributed as points lie close to the line.
- **Scale-Location Plot** confirms the constant variance assumption, as the points are randomly distributed.
- **Residuals vs. Leverage Plot** implies that no outliers are present, as all values are within 0.5 bands.

2.**Detecting Collinearity** - vif is below the recommended maximum value of 4 (Pan & Jackson, 2008). Hence multicollinearity is ruled out. `am` and `wt` both have relatively higher inflation factors due to correlation.

```
##       am       wt     qsec
## 2.541437 2.482952 1.364339
```

3.**Influential Observations** - From the hat values of the final model, Chrysler Imperial seems to exert a high leverage on the model.

```
##   Chrysler Imperial Lincoln Continental           Merc 230
##               0.230               0.264              0.297
```

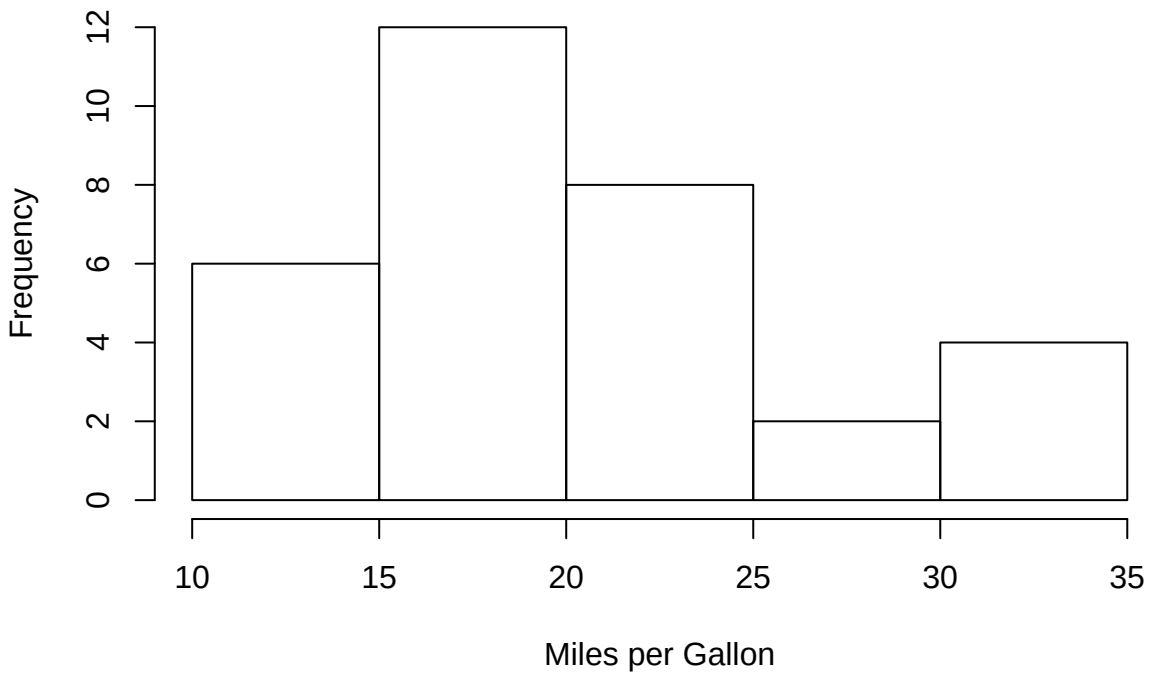
Hence the final model holds good.

Conclusion

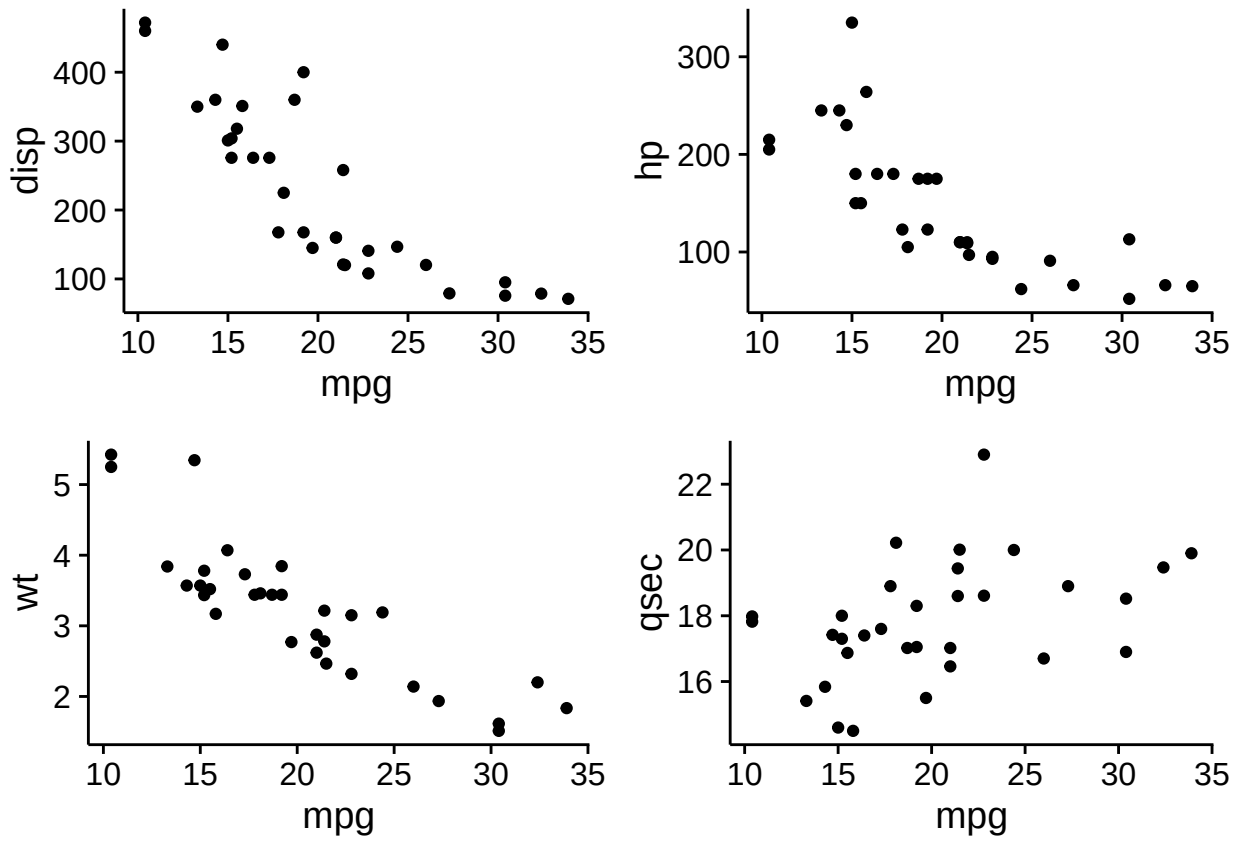
This model is difficult to interpret, and the absence of DISPLACEMENT or HORSEPOWER, which intuition suggests should be important in the prediction of MPG, is surprising. This is primarily due to a bias to non-US automobiles (7 Mercedes, 1 Porsche, 1 Ferrari and 1 Maserati) which have different specs, especially engine specs. Therefore, a universal prediction model isn't possible.

Appendix

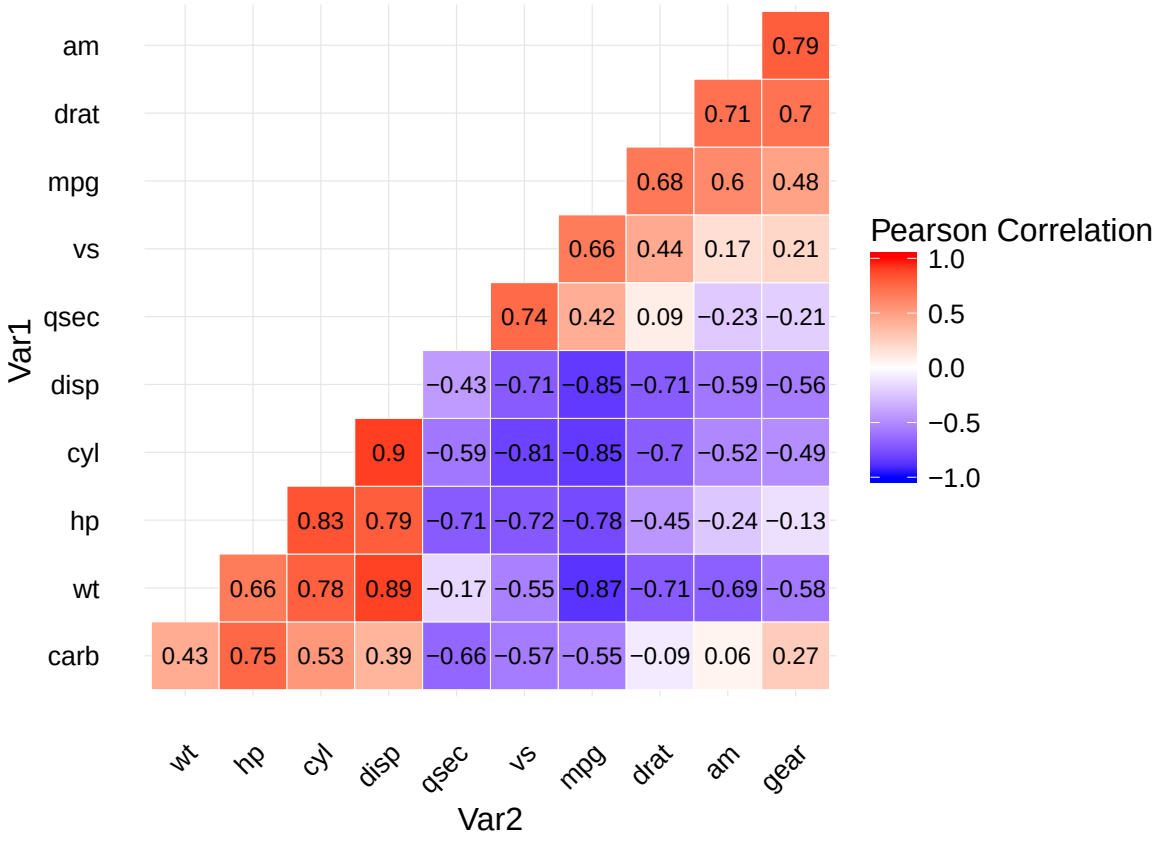
Plot 1: Distribution of mpg variable



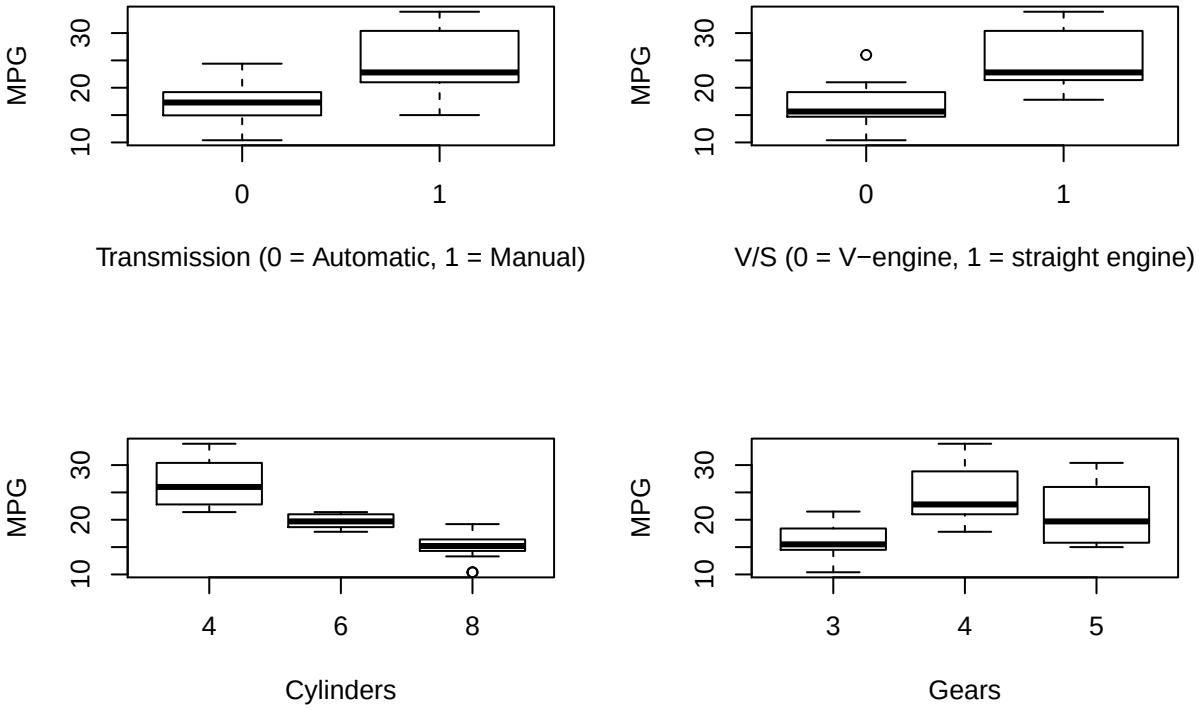
Plot 2: MPG vs other variables



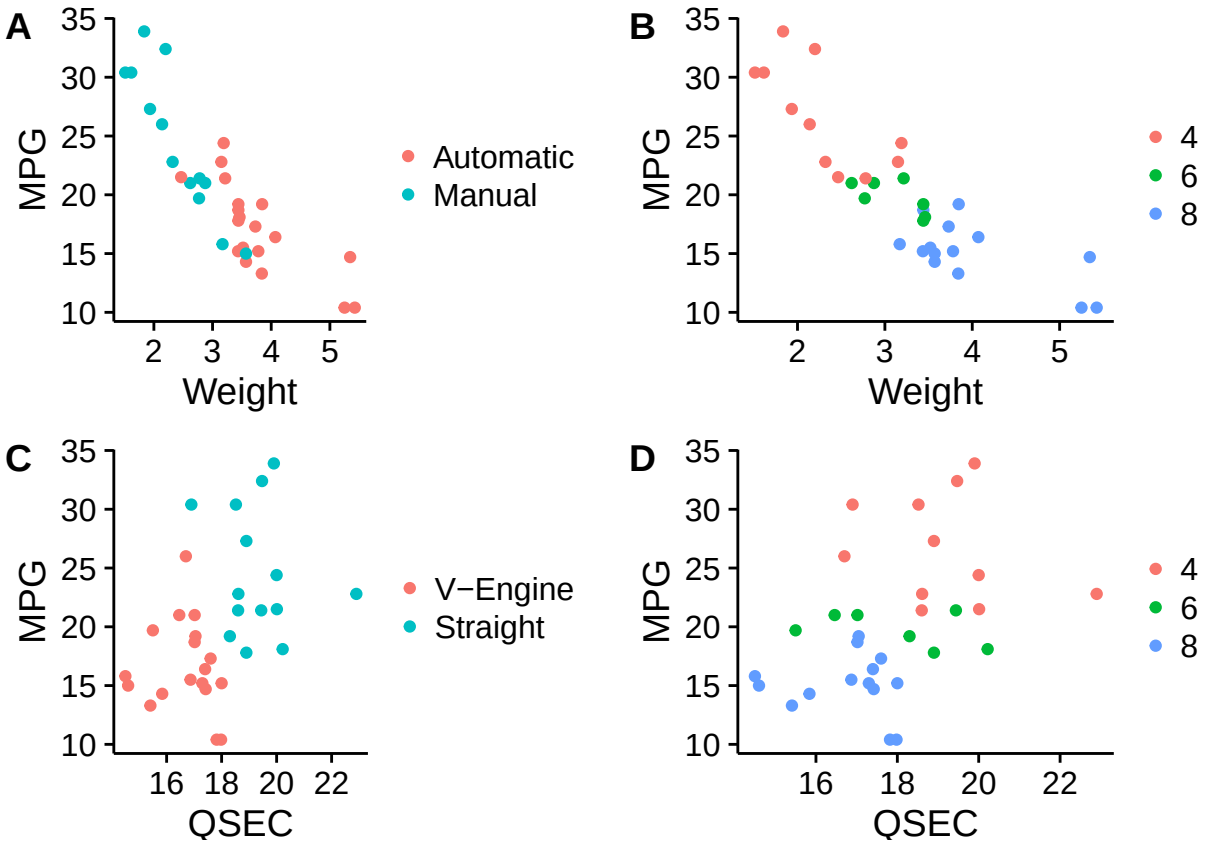
Plot 3: Correlation between all variables



Plot 4: Boxplot of MPG vs AM, VS, Gear Variables



Plot 5: MPG vs WT and QSEC by Transmission, Engine and Cylinder



Plot 6: Residuals