Regression Analysis Project

Jonathan Bennett

Table of Contents

1. Executive Summary

This analysis addresses the relationship between a group of variables and the resulting miles per gallon for a subset of cars from the mtcars dataset. Specificically the analysis determines whether an automatic or manual transmission achieves better gas mileage, and attempts to quantify the difference. For details on the dataset, please visit:
https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html)

Findings: Using the AIC approach, an automatic transmission decreased gas mileage by 0.07 MPG. Appropriate input variables were wt, cyl, hp, and, of course, am. However, the standard error for the am is 1.4, meaning that a 95% confidence interval spans -2.84 to 2.69. Therefore the finding that an automatic transmission is less efficient can't be stated with a 95% guarantee. Rather, the transmission type is inconclusive as a predictor of MPG.

This analysis employs a rigorous quantitative and qualatitive approach to model selection. This investigates five model approaches:
1) Use all regression variables (cyl, disp, hp, drat, wt, qsec, am, gear, vs, carb)
2) Akaike Information Criterion (AIC-http://en.wikipedia.org/wiki/Akaike_information_criterion (http://en.wikipedia.org/wiki/Akaike_information_criterion))
3) Bayesian Information Criterion (BIC-http://en.wikipedia.org/wiki/Bayesian_information_criterion (http://en.wikipedia.org/wiki/Bayesian_information_criterion))
4) Only use am as a predictor

The AIC approach made the most sense because it optimizes fit without overfitting. Other approaches minimized rsquared values, but introducing variables uncorrelated to the output will increase rsquared values so minimizing rsquared was not a determining factor. This analysis performed some data maniuplation prior to fitting:
1) Converted appropriate inputs to factor type (am, cyl, gear, car, vs). This made sense because the data were not numeric in nature. Arguably gear, carb and cyl may be, but am and vs are most certainly

not.
2) Removed training outliers that were outside of the 95th percentile. Given the few data points, outliers were likely to skew results so the outliers were deemed as detrimental to a good fit for purposes of this investigation.

For purposes of brevity, much of the code in this Rmd file has been hidden as the assignment mandated a length of less than seven pages total.

2. Coefficient Interpretation

The regression equation to obtain the mpg is: mpg = 35.92996 - 3.70181 * wt - 0.02670 * hp - 1.35179 (if 6 cylinders) - 1.16743 (if 8 cylinders) - 0.07253 (if automatic transmission)
We start at 35.92996 MPG (intercept) before adding other elements. It makes sense the we lose appx 3.70 MPG for each unit of weight, and 1.35 for each unit of hp. A four cylinder car has no MPG adder, but both 6 and 8 cylinders reduce gas mileage. The regression equation indicates 8 cylinder cars are more efficient (less or a reduction) than 6 cylinder cars, although this might not folow intuition. More data or analysis is warranted on this variable.

```
## Loading required package: ggplot2
## Loading required package: reshape
```

```
## Warning: package 'reshape' was built under R version 3.1.3
```

```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 3.1.3
```

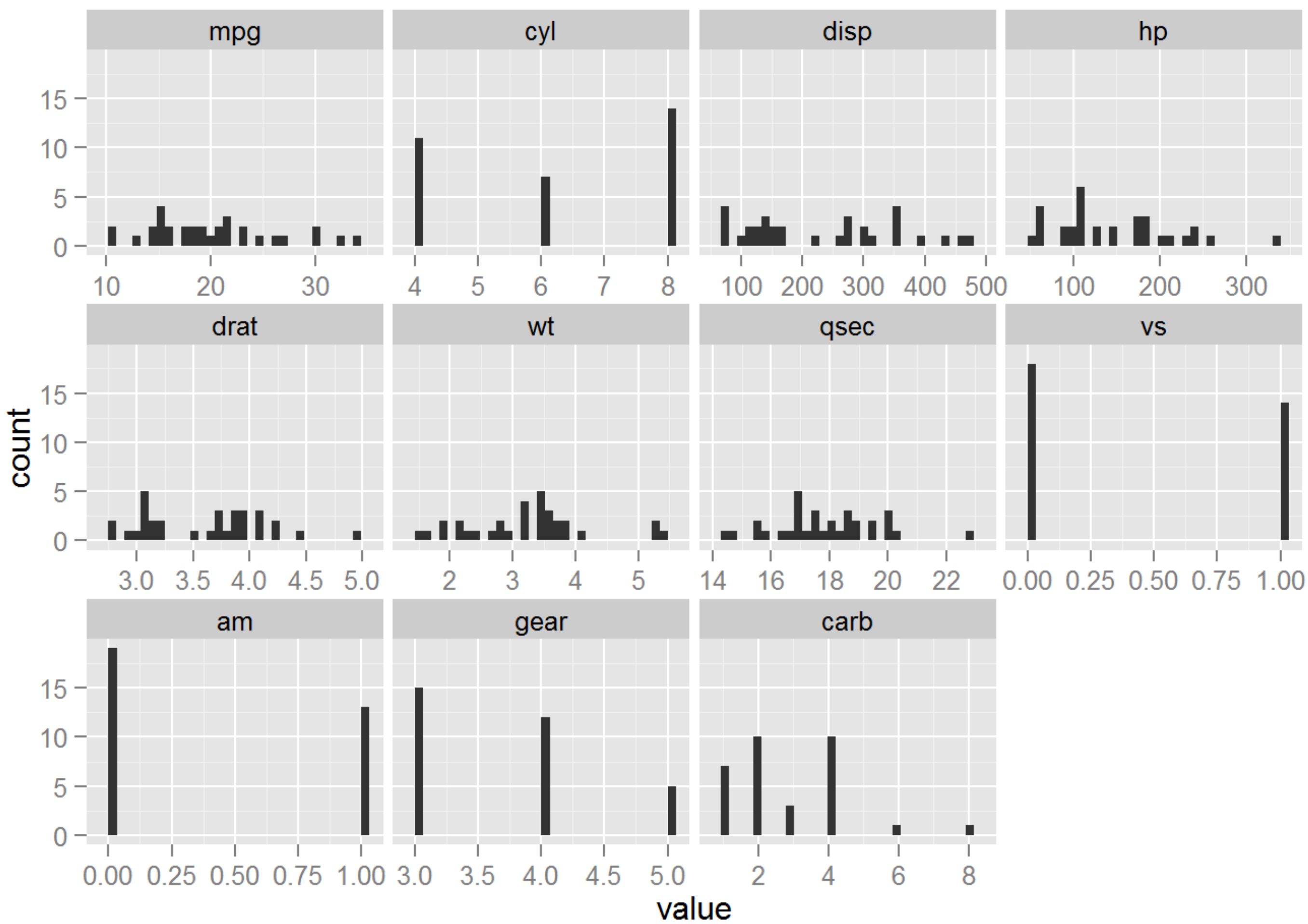
3. Exploratory Analysis

```
d <- melt(mtcars)
```

```
## Using  as id variables
```

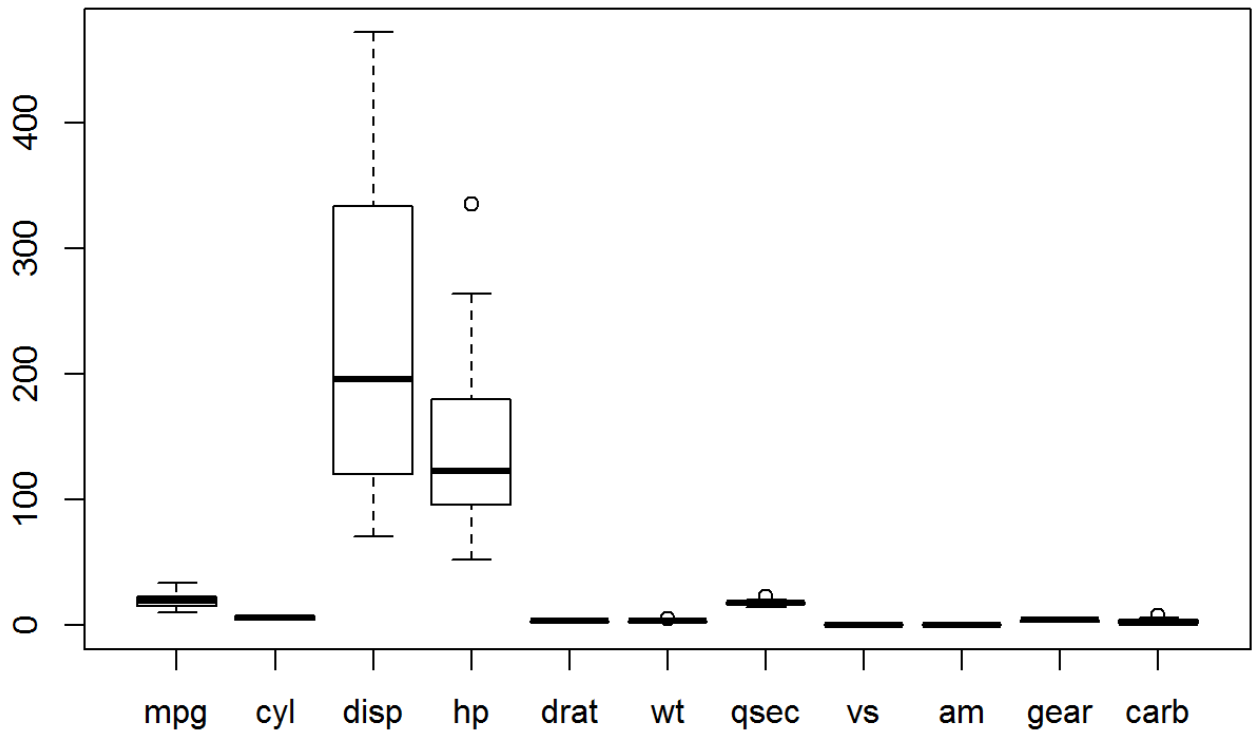
```
ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
# d <- plot(mtcars$mpg) #not generated for purposes of brevity

#look at outliers and remove all datapoint outside of +- two standard deviations
boxplot(mtcars)
```




```
outliers <- !((filteroutlier(mtcars$mpg,confidence)) & (filteroutlier(mtcars$disp,confide
nce)) &
                (filteroutlier(mtcars$hp,confidence)) & (filteroutlier(mtcars$drat,conf
idence)) &
                (filteroutlier(mtcars$wt,confidence)) & (filteroutlier(mtcars$qsec,conf
idence)))
mtcars2 <- mtcars[!outliers,]
```

4. Model Selection

```
## Start:  AIC=70.91
## mpg ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + wt    1    317.86 105.97 39.642
## + cyl   2    314.45 109.38 42.403
## + disp  1    255.42 168.41 50.760
## + hp    1    240.37 183.45 52.814
## + am    1    140.00 283.83 63.288
## + gear  2    156.82 267.01 63.821
## + vs    1    127.37 296.46 64.332
## + drat  1    106.73 317.10 65.948
## + qsec  1     47.75 376.08 70.042
## + carb  4    122.53 301.30 70.721
## <none>              423.83 70.911
##
## Step:  AIC=39.64
## mpg ~ wt
##
##        Df Sum of Sq    RSS    AIC
## + hp    1     38.69  67.28 30.739
## + qsec  1     26.49  79.48 34.739
## + cyl   2     27.97  78.00 36.287
## + vs    1     16.25  89.72 37.648
## + disp  1      9.74  96.23 39.328
## + am    1      8.86  97.11 39.546
## <none>              105.97 39.642
## + drat  1      1.53 104.44 41.293
## + carb  4     23.89  82.08 41.512
## + gear  2      4.94 101.03 42.496
## - wt    1    317.86 423.83 70.911
##
## Step:  AIC=30.74
## mpg ~ wt + hp
##
##        Df Sum of Sq     RSS    AIC
## <none>               67.280 30.739
## + disp  1     2.218  65.062 31.935
## + gear  2     7.041  60.238 32.086
## + vs    1     0.331  66.949 32.621
## + am    1     0.314  66.966 32.627
## + qsec  1     0.071  67.209 32.714
## + drat  1     0.020  67.260 32.732
## + cyl   2     3.778  63.501 33.352
## + carb  4    11.514  55.766 34.235
## - hp    1    38.689 105.969 39.642
## - wt    1   116.175 183.455 52.814
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars3)
##
## Coefficients:
## (Intercept)           wt           hp
##    36.79726     -4.16634     -0.02902
```

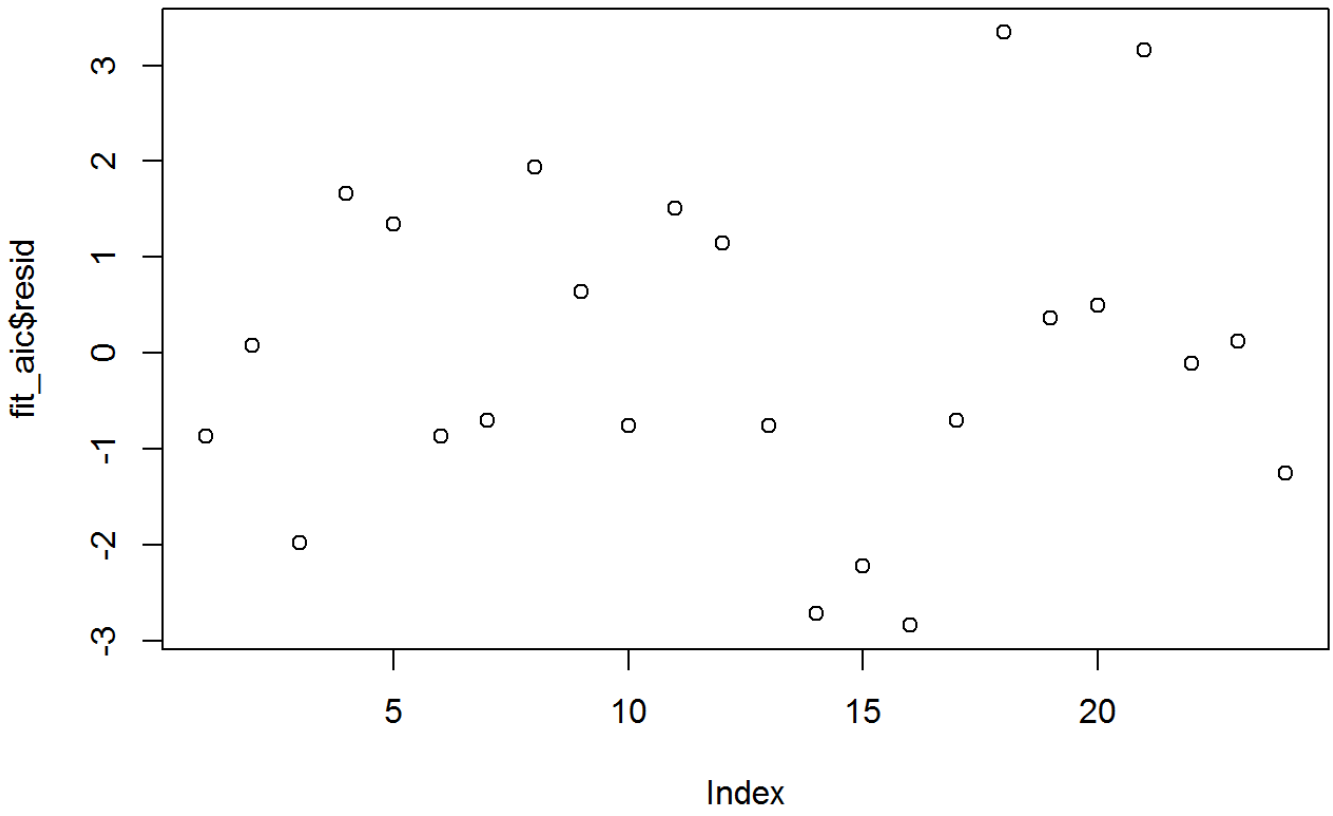
```
#Best AIC with transmission variable approach
fit_aic <- lm(mpg ~ wt + cyl + hp + am, data = mtcars3)
#summary(fit_aic)

#BIC approach
fit_bic <- lm(mpg ~ wt + qsec + am, data = mtcars3)

#all variables
fit_all <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + am + gear + vs + carb, data = mt
cars3)

#only am
fit_onlyam <- lm(mpg ~ am, data = mtcars3)
```

5. Residual Plot



6. Model Comparisons

```
##        approach automatic_mpgadder  rsquared degreesf
## 1 All_variables         1.75684759 0.9726420       16
## 2           AIC        -0.07253162 0.8501950        6
## 3           BIC         0.10277457 0.8125057        4
## 4       only_am         4.98888889 0.3303231        2
```