Effect of Tranmission Choice on Fuel Economy in 1974-Era Automobiles

BT, Coursera Course Project

2014-08-17

Executive Summary

This report was commissioned to determine the effect of a manual or automatic transmission on fuel economy in 1974-era automobiles.

To answer this question, this report evaluated models of outcome fuel economy (*miles per gallon* or *mpg*) using linear regression and selected a single model for intepretation.

This report finds evidence that a manual transmission improves fuel economy over a manual transmission, by an estimated 2.9 miles per gallon.

A brief overview of the data

The *mtcars* data set describes 32 1974-model automobiles, and provides complete values for eleven variables. Six are continuous variables (mpg, disp, hp, drat, wt, qsec). Three are discrete numerical variables (cyl, gear, carb). Two are boolean categorical variables (vs, am).

Analysis Goals

The goal of this analysis is to answer two questions:

1. Is an automatic transmission (am=0) or manual transmission (am=1) better for fuel economy (mpg)?
2. What is the quantified mpg difference between automatic and manual transmissions?

We will build a linear model of the outcome, mpg, based on explanatory variables.

Model Selection

Models were evaluated using the Bayesian Information Criterion (BIC), using the R *leaps* package. Our analysis goals favor models that include am, and the model with the best BIC score does.

See Appendix A for more details.

A Model For Fuel Economy

This best BIC model uses wt, $qsec$ and am as explanatory variables:

$model_{mpg} = -3.917\,wt + 1.226\,qsec + 2.936\,am$

Here wt is weight in 1000 pounds, $qsec$ is the time to travel a quarter mile from a start, and am indicates automatic (am=0) or manual transmission (am=1).

```
summary(lm(mtcars$mpg ~ mtcars$wt + mtcars$qsec + mtcars$am))
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$wt + mtcars$qsec + mtcars$am)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## mtcars$wt     -3.917      0.711   -5.51    7e-06 ***
## mtcars$qsec    1.226      0.289    4.25  0.00022 ***
## mtcars$am      2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

95% confidence intervals for coefficients:

```
print(qsec_ci_95 <- 1.2259+c(-1,1)*0.2887**qt(0.975,df=31))
```

```
## [1] 1.147 1.305
```

```
print(wt_ci_95 <- -3.9165+c(-1,1)*0.7112*qt(0.975,df=31))
```

```
## [1] -5.367 -2.466
```

```
print(am_ci_95 <- 2.9358+c(-1,1)*1.4109*qt(0.975,df=31))
```

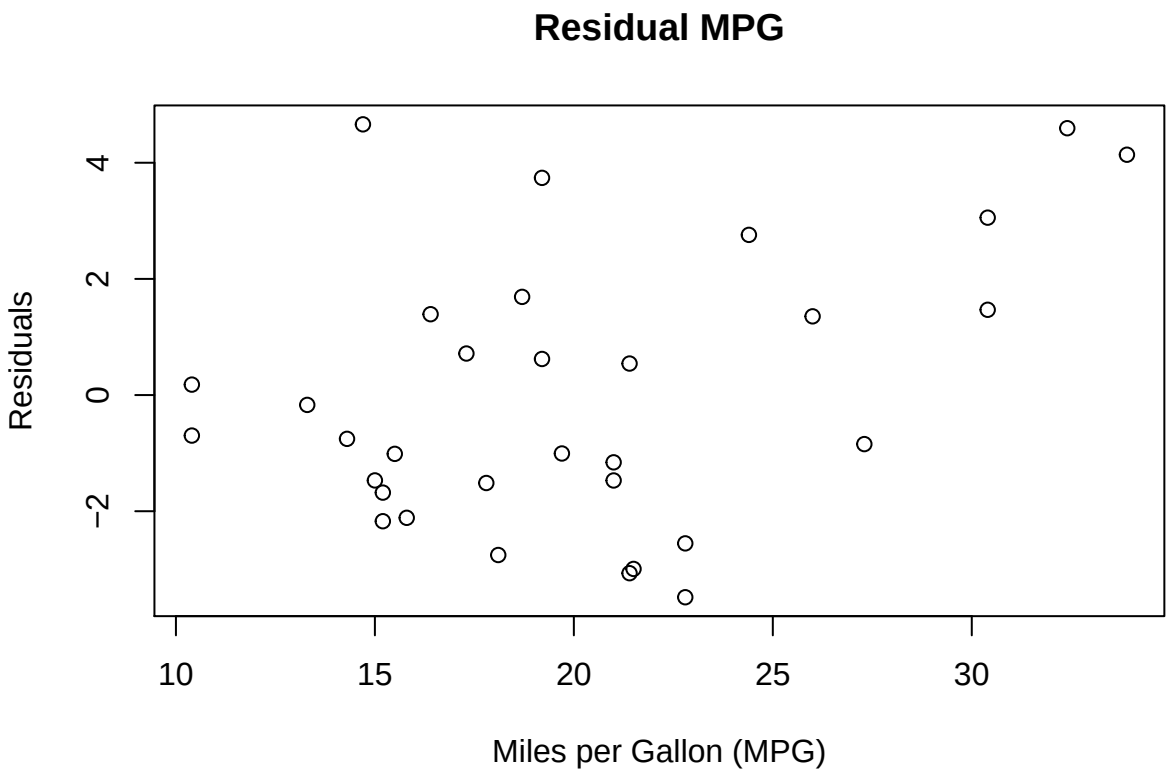
```
## [1] 0.05825 5.81335
```

The estimated regression coefficient for *am* is *2.936*.

The 95% confidence interval for this coefficient is [0.058, 5.813], providing evidence (p=0.047) in favor of rejecting the null hypothesis in favor of the hypothesis that *am* has a positive coefficient. **This would indicate that using a manual transmission improves fuel economy (by an estimated 2.9 mpg), versus an automatic transmission**.

wt and *qsec* also show strong evidence for non-zero effects, with p=.000007 and p=0.0002 respectively. These observations are consistent with mechanistic expectations. For *wt*, moving more mass requires more energy, so it should be negatively correlated with *mpg*. Similarly, a slow quarter-mile requires less energy, so *qsec* should be positively correlated with *mpg*.

```
model <- lm(mtcars$mpg ~ mtcars$wt + mtcars$qsec + mtcars$am)
plot(mtcars$mpg, resid(model), main="Residual MPG", xlab="Miles per Gallon (MPG)",
    ylab="Residuals")
```

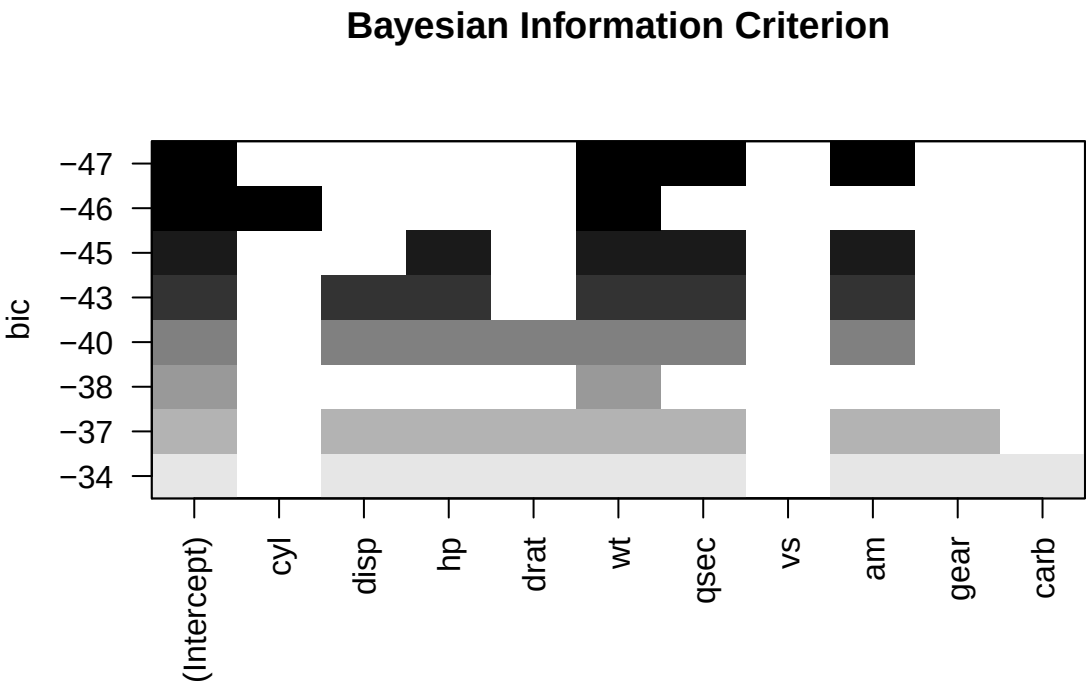


Appendix A: Model Selection

The Bayesian Information Criterion (BIC) is a criterion that weighs the quality of a model's fit against the number of variables included in the model.

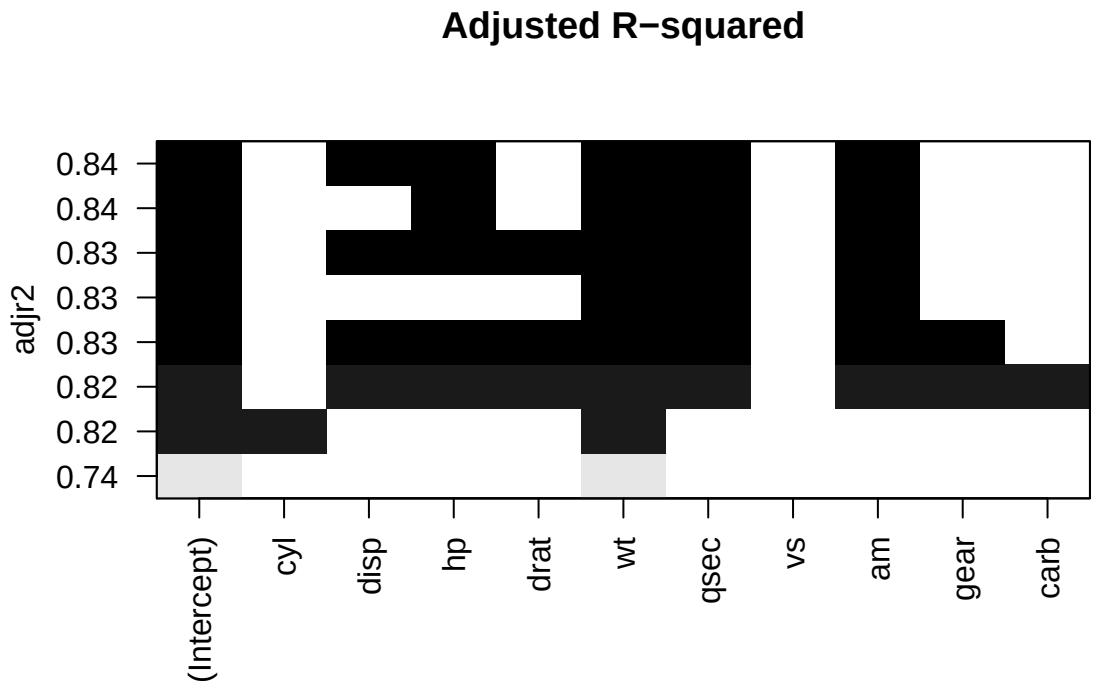
This technique can evaluate all 2^n models to suggest good candidates. Here, with 10 candidate variables, 1024 models are evaluated.

```
data(mtcars)
library('leaps')
models <- regsubsets(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
                     data=mtcars)
plot(models, scale="bic", main="Bayesian Information Criterion")
```



Here a lower score is better, suggesting a model explaining mpg using (*wt*, *qsec*, *am*). Since our investigatory questions suggest inclusion of the *am* variable, it is worth considering the model which includes *hp*, (*wt*, *qsec*, *am*, *hp*).

```
plot(models, scale="adjr2", main="Adjusted R-squared")
```



Note that the adjusted R-squared model including *hp* actually has a slightly higher adjusted R-squared, and both scores indicate a reasonable fit.

```
summary(lm(mtcars$mpg ~ mtcars$wt + mtcars$qsec + mtcars$am + mtcars$hp))
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$wt + mtcars$qsec + mtcars$am +
##     mtcars$hp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.497 -1.590 -0.112  1.180  4.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.4402     9.3189    1.87   0.0721 .
## mtcars$wt    -3.2381     0.8899   -3.64   0.0011 **
## mtcars$qsec   0.8106     0.4389    1.85   0.0757 .
## mtcars$am     2.9255     1.3971    2.09   0.0458 *
## mtcars$hp    -0.0176     0.0142   -1.25   0.2231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.43 on 27 degrees of freedom
## Multiple R-squared:  0.858,  Adjusted R-squared:  0.837
```

```
## F-statistic: 40.7 on 4 and 27 DF,  p-value: 4.59e-11
```

The model including *hp* provides a similar result for the *am* coefficient to the one omitting *hp*.

Including *hp*, the 95% confidence interval (CI) for the *am* coefficient is [0.07599,5.775] (p=0.04579), while excluding *hp* the 95% CI for the *am* coefficient is [0.05828, 5.8133] (p=0.046716).

Given the relative agreement of these two models, we opt for the simpler one using BIC.