Regression Models: Motor Trend Cars - Impact of Transmission Type on MPG

Mark Culp

April 30, 2017

Executive Summary

This report examines Motor Trend car road tests extracted from the 1974 Motor Trend US magazine. These road tests were conducted on 1973-1974 car models. The tests examined 10 aspects of automobile design and performance for 32 different models. We focused here on the miles per gallon (MPG) performance of automatic versus manual transmissions.

We concluded that cars with manual transmissions had better gas mileage than cars with automatic transmissions. The differences in MPG attained under our Simple Linear Regression (SLR) analysis were dramatically higher than those obtained using Multivariable Linear Regression.

The effect of a car's transmission on MPG can be estimated with a relatively few number of variables contained in the mtcars data set. We identified a suitable model using three variables: transmission type, weight, and cylinders.

Detail of Analysis

We initially identified seven variables that appeared to have an impact on MPG. Some calculations on the correlation between these variables and MPG were made to see if our intuition was correct. A summary of those calculations is presented below:

Table 1: Potential Regressors

cyl	disp	hp	wt	am	gear	carb
-0.852	-0.848	-0.776	-0.868	0.6	0.48	-0.551

The above table summarizes these calculations. The "am" variable is our transmission type regressor. It is a numeric value with 0 designating a car with an automatic transmission, and 1 designating a car with a manual transmission. Our table shows a 0.6 positive correlation with MPG as we compare cars with automatic transmissions to those with manual transmissions.

The number of forward gears and the number of carburetors on an automobile had lower correlations to MPG than the other five variables so we dropped them. Our goal was to keep the model simple, and easily interpretable at the risk of introducing bias from omitting relevant predictors.

We then examined two sets of regressors that appeared to be highly correlated to each other. First was displacement and weight. The second was cylinders and horsepower. The correlation between displacement and weight was almost 89%. The correlation between cylinders and horsepower was about 83%.

While these variables describe very different car characteristics, their high correlations made them redundant. We therefore excluded displacement and horsepower from our final model (fit3). Horsepower was examined in the fourth model (fit4), but as expected, it contributed very little predictive value to our model. The slope of the linear model only decreased by about 0.03 MPG.

Simple linear regression (SLR) identified an approximately 7.25 MPG difference between cars with automatic transmission and cars with manual transmission. This was essentially the difference between the means of the

two groups. The differences were much less when other variables, or regressors, were considered. Our analysis showed the weight of the car and the number of cylinders had a greater impact on MPG than transmission type.

We used the Analysis of Variance (anova) and VIF functions to assess the variances introduced by our multivariable model. The anova function provided an F-statistic for fit4 indicating that the introduction of horsepower into the model was not statistically significant at the 95% confidence level. The VIF function showed significant variance inflation from the introduction of all three of the regressors. The weight variable introduced 3.6 times the variance produced by an ideal regressor.

We used the dfbetas function to identify outliers in our data. The Chrysler Imperial and the Toyota Corona were found to have low dfbetas relative to automatic transmissions. Their low dfbetas here appears to have resulted from the relatively low MPGs each car had given their respective cylinder numbers and weights.

Appendix

```
# Load data set
library(datasets)
data("mtcars")
```

Exploratory Analysis

```
# Examine columns and data types
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
# Range of MPG values
range(mtcars$mpg)
```

```
## [1] 10.4 33.9
```

```
# Car transmissions: 0 = automatic, 1 = manual
table(mtcars$am)
```

```
##
##  0  1
## 19 13
```

Fitting Simple Linear Regression (SLR)

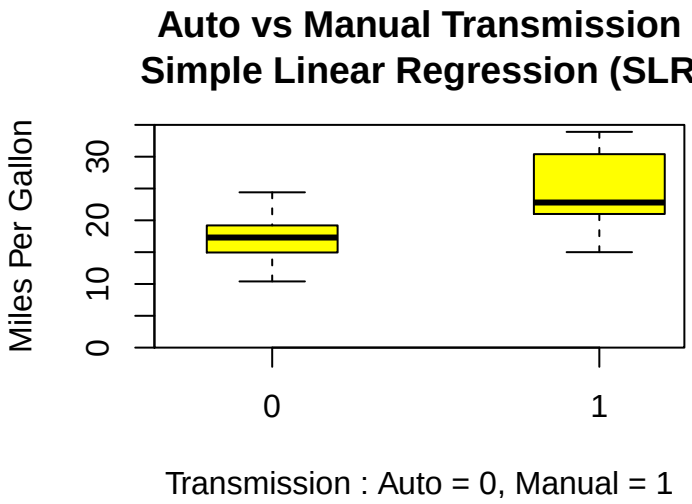
```r
# So, we estimate a 7.25 mpg increase in moving from an
# automatic to a manual transmission.  We expect to get
# 17 mpg with an automatic transmission.
fit1 <- lm(mpg ~ factor(am) - 1, data = mtcars)

# Summarize SLR comparison of auto vs man transmission
summary(fit1)$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## factor(am)0 17.14737   1.124603 15.24749 1.133983e-15
## factor(am)1 24.39231   1.359578 17.94109 1.376283e-17
```

```r
# So we can reject the null hypothesis that the transmission has no impact on MPG.

boxplot(mpg ~ am, data = mtcars,
        boxwex = 0.4, at = 0:1 - 0.2,
        main = "Auto vs Manual Transmission\n Simple Linear Regression (SLR)",
        col = "yellow",
        xlab = "Transmission : Auto = 0, Manual = 1",
        ylab = "Miles Per Gallon",
        xlim = c(-0.5,1), ylim = c(0,35), yaxs = "i")
```



Fitting Multivariable Models

```r
library(knitr)

# Examine the correlation between MPG and some of the
# regressors likely to impact it.  Code for creating
# the Table 1, Potential Regessors.
regressor <- c("cyl","disp","hp","wt","am", "gear","carb")
correlate <- c(round(cor(mtcars$mpg,mtcars$cyl),3),
               round(cor(mtcars$mpg,mtcars$disp),3),
               round(cor(mtcars$mpg,mtcars$hp),3),
               round(cor(mtcars$mpg,mtcars$wt),3),
```

```
               round(cor(mtcars$mpg,mtcars$am),3),
               round(cor(mtcars$mpg,mtcars$gear),3),
               round(cor(mtcars$mpg,mtcars$carb),3)
               )

# Create data frame
mtCor <- data.frame(regressor, correlate)

# Table of correlations, Table 1, Potential Regressors
regressorTable <- kable(t(mtCor[,2]),caption = "Potential Regressors", col.names = t(mtCor[,1]))

# Examine correlation between displacement
# and weight.
cor(mtcars$disp,mtcars$wt)
```

```
## [1] 0.8879799
```

```
# Examine correlation between cylinders
# and horse power.
cor(mtcars$cyl,mtcars$hp)
```

```
## [1] 0.8324475
```

```
# Update our model to include weight, cylinders,
# and horse power.
fit2 <- update(fit1, mpg ~ factor(am) + wt - 1)
fit3 <- update(fit1, mpg ~ factor(am) + wt + cyl - 1)
fit4 <- update(fit1, mpg ~ factor(am) + wt + cyl + hp - 1)

# Examine differences between fit3 and fit4 to
# assess value of including horsepower as a variable.
summary(fit3)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## factor(am)0 39.417933  2.6414573 14.922798 7.424998e-15
## factor(am)1 39.594427  1.8721428 21.149255 9.322776e-19
## wt          -3.125142  0.9108827 -3.430894 1.885894e-03
## cyl         -1.510246  0.4222792 -3.576415 1.291605e-03
```

```
summary(fit4)$coef
```

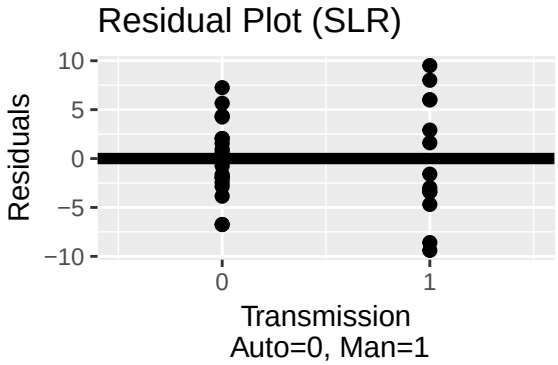
```
##                Estimate Std. Error   t value     Pr(>|t|)
## factor(am)0 36.14653575 3.10478079 11.642218 4.944804e-12
## factor(am)1 37.62458346 2.09640689 17.947176 1.556106e-16
## wt          -2.60648071 0.91983749 -2.833632 8.603218e-03
## cyl         -0.74515702 0.58278741 -1.278609 2.119166e-01
## hp          -0.02495106 0.01364614 -1.828433 7.855337e-02
```

Residual Plot and Diagnostics

```
library(ggplot2)
library(car)
```

```
## Warning: package 'car' was built under R version 3.3.3
```

```
g <- ggplot(data.frame(x = mtcars$am, y = resid(fit1)), aes(x=x, y=y))
g <- g + ggtitle("Residual Plot (SLR)")
g <- g + xlab("Transmission\n Auto=0, Man=1")
g <- g + ylab("Residuals")
g <- g + geom_point(size = 2, colour = "black")
g <- g + geom_hline(yintercept = 0, size = 2)
g <- g + scale_x_continuous(breaks = 0:1, limits = c(-0.5,1.5))
g
```




```
# Compute analysis of variance table
anova(fit1,fit2,fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) - 1
## Model 2: mpg ~ factor(am) + wt - 1
## Model 3: mpg ~ factor(am) + wt + cyl - 1
## Model 4: mpg ~ factor(am) + wt + cyl + hp - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.2925  5.39e-09 ***
## 3     28 191.05  1     87.27 13.8611 0.0009165 ***
## 4     27 170.00  1     21.05  3.3432 0.0785534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Examine variance inflation with 3 factors
vif(lm(mpg ~ factor(am) + wt + cyl, mtcars))
```

```
## factor(am)         wt        cyl
##   1.924955   3.609011   2.584066
```

```
# Examine dfbetas to identify outliers
dfb <- round(dfbetas(fit3),3)
dfb[dfb[,1] > 0.5 | dfb[,1] < -0.5,]
```

```
##                   factor(am)0 factor(am)1    wt   cyl
## Chrysler Imperial      -0.618      -0.596 0.947 -0.45
## Toyota Corona          -0.830      -0.701 0.265  0.34
```