Regression Models Transmission vs. MPG Report

Dakota Carter

January 23, 2016

Executive Summary

Motor Trend's car dataset is explored to evaluate the claim that type transmission strongly influences the gas milage of cars. The multi-stage analysis uses linear regression, and in the end finds a modest relationship between the two variables, with manual cars getting less than two more miles per gallon on average.

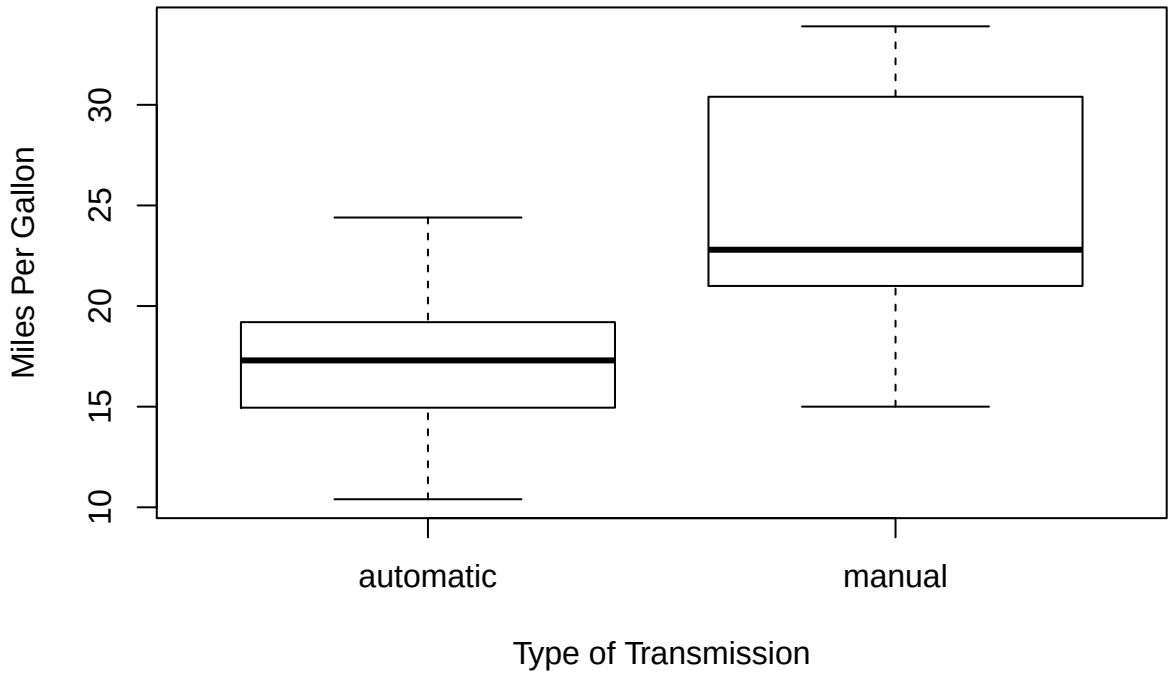
Load & Process Data

```
suppressMessages(suppressWarnings(library(dplyr)))
data(mtcars)
#Factorize vs and am, make am more readable
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- sub(0, "automatic", mtcars$am)
mtcars$am <- sub(1, "manual", mtcars$am)
mtcars$am <- as.factor(mtcars$am)
```

See Figure 1 (in appendix) for a peek at the data

Exploratory Analysis

```
boxplot(mpg ~ am, mtcars, ylab="Miles Per Gallon", xlab="Type of Transmission")
```



From here it looks as if cars with a manual transmission get better gas mileage, but of course we must dig deeper. First we build a simple linear model!

```
single_reg <- lm(mpg ~ factor(am), data=mtcars)
single_reg
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Coefficients:
##      (Intercept)  factor(am)manual
##           17.147             7.245
```

In this model manual transmission cars get on average 7.3 more miles to the gallon than automatic cars, however in Figure 1 (see appendix) we can see that only about 40% of the variance of the regression can be explained by our model, so there is probably more affecting the mpg than just type of transmission. Again we must dig deeper. Now we build a multivariate linear model.

```
summary(lm(mpg ~ . ,data=mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs1          0.31776    2.10451   0.151   0.8814
## ammanual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

From this summary of effects we can conclude that the important predictive variables for a multivariate model are: cyl, hp, wt, carb, drat, disp, and am.

```
multi_reg <- lm(mpg ~ cyl+hp+wt+carb+drat+disp+am, data = mtcars);
multi_reg
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + carb + drat + disp + am, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl           hp           wt         carb
##   33.140529    -0.845406    -0.020055    -2.663390    -0.419967
##        drat         disp     ammanual
##    0.783716     0.005875     1.719368
```

This model explains about 86% of the regression variance, much better than with just transmission. To see the summary of this linear model, and relevant p-values see appendix figure 3. To see the diagnostics plot see appendix Figure 4.

We can see from the Q-Q plot that the data are normally distributed and from the residuals plot that the data have approximately the same variance (or are homescedatic).

Conclusion

Since the last model explains 85.8% of the regression variance, we can conclude that the initial findings of a strong relationship between type of transmission and gas milage conflated the effects of other variables, namely weight and number of cylinders. In the new model on average manual cars get 1.72 more miles to the gallon than automatics, a much more modest association. With a p-value of 0.316 though we fail to reject the hypothesis that transmission type has any effect on gas mileage. We have such a small sample that more data is needed.

Appendix

Figure 1

```
sample_n(mtcars, 10)
```

```
##                     mpg cyl  disp  hp drat    wt  qsec vs        am gear
## Pontiac Firebird   19.2   8 400.0 175 3.08 3.845 17.05  0 automatic    3
## Maserati Bora      15.0   8 301.0 335 3.54 3.570 14.60  0    manual    5
## Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0 automatic    3
## Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1 automatic    4
## Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1    manual    4
## Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1 automatic    4
## Fiat X1-9          27.3   4  79.0  66 4.08 1.935 18.90  1    manual    4
## Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1    manual    4
## Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1 automatic    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0 automatic    3
##                    carb
## Pontiac Firebird      2
## Maserati Bora         8
## Merc 450SE            3
```

```
## Merc 280              4
## Fiat 128              1
## Merc 230              2
## Fiat X1-9             1
## Datsun 710            1
## Toyota Corona         1
## Cadillac Fleetwood    4
```

Figure 2

```
summary(single_reg)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17.147      1.125  15.247 1.13e-15 ***
## factor(am)manual    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Figure 3

```
summary(multi_reg)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + carb + drat + disp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8660 -1.3907 -0.4814  1.5252  5.3075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.140529   8.987870   3.687  0.00116 **
## cyl         -0.845406   0.791853  -1.068  0.29631
## hp          -0.020055   0.020278  -0.989  0.33255
## wt          -2.663390   1.523541  -1.748  0.09322 .
```

```
## carb        -0.419967   0.669436  -0.627  0.53636
## drat         0.783716   1.586146   0.494  0.62573
## disp         0.005875   0.016130   0.364  0.71886
## ammanual     1.719368   1.680204   1.023  0.31637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.58 on 24 degrees of freedom
## Multiple R-squared:  0.8581, Adjusted R-squared:  0.8167
## F-statistic: 20.73 on 7 and 24 DF,  p-value: 9.891e-09
```

Figure 4

```
par(mfrow = c(2, 2))
plot(multi_reg)
```