Does a cars MPG depend on its Transmission?

Janamejaya Chowdhary

February 6, 2017

Executive Summary

On behalf of the magazine, Motor Trends, analyze vehicle features for a collection of cars (mtcars dataset) in order to (a) determine whether an Automatic or a Manual Transmission is better for a cars MPG and (b) quantify the MPG difference between Automatic and Manual Transmissions.

Exploratory data analysis

To start, the mtcars dataset is loaded as a dataframe and the contents summarized as follows

```
data(mtcars); colnames(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

The dataset contains values of 11 features for each car. The features of primary interest are **mpg** (miles per gallon) and **am** (**0**: Automatic Transmission (**AT**) and **1** for Manual Transmission (**MT**)). Correlation of **mpg** with all features is calculated below

```
mcor <- cor(mtcars); mcor <- round(mcor,2); mcor["mpg",]
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

The features most and least correlated with **mpg** are **wt** (cor=0.89) and **qsec** (cor=0.42), respectively. The feature of interest, **am**, has intermediate correlation (cor=0.6) with **mpg**. The full correlation matrix and pair feature dependence are presented in Appendix A1 and A2 respectively. Correlation between **cyl** and **disp** (cor=0.9) is very high. Of these, the highest correlation is between **wt** and **disp** (cor=0.89). It may be prudent to remove **disp** from the following Regression analysis.

Univariate Linear Regression analysis

A simple univariate regression between **mpg** and **am** is performed first.

```
fit1 <- lm(mpg ~ as.factor(am), data=mtcars)
summary(fit1)$coeff
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    17.147368   1.124603 15.247492 1.133983e-15
## as.factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

From this, the mean **mpg** value for **AT** is the intercept 17.15. The slope 7.24 corresponds to the difference in mean **mpg** values for **AT** and **MT**. So, the mean **mpg** for **MT** is 24.39. To establish the significance of this difference between mean values, the null hypothesis that there is no difference between mean **mpg** values for **AT** and **MT** is tested next.

```
ttest <- t.test(mpg ~ am, data=mtcars)
```

The 95% confidence interval (-11.28, -3.21) does not contain zero. Hence, the null hypothesis is rejected. Thus, the difference in mean **mpg** between **AT** and **MT** is non-zero.

Multivariate Linear Regression analysis

For the linear model (fit1), percentage of variance explained $R^2 = 0.36$ and its value adjusted for number of predictors is adj.$R^2 = 0.34$. The small values indicate that other variables may be significant. A multivariate regression is performed with all features excluding **disp** as it is highly correlated with other features.

```
fit2 <- lm(mpg ~ cyl + hp + drat + wt + qsec + vs + factor(am) + gear + carb, data=mtcars)
```

The values of $R^2 = 0.87$ and adj.$R^2 = 0.81$ are much better fit1 but p-values for coefficients are too large.

```
## (Intercept)         cyl          hp        drat          wt        qsec
##        0.51        0.92        0.49        0.57        0.04        0.35
##          vs factor(am)1        gear        carb
##        0.94        0.24        0.62        0.32
```

To improve fit quality, the feature with largest p-value is dropped, the linear model refit, and the process continued till all p-values are small. The order of feature deletion is **vs**, **cyl**, **gear**, **hp**, **drat**, and **carb**. The final model (below) has $R^2 = 0.85$, adj.$R^2 = 0.83$ (better than fit2), and the following coefficients

```
fit8 <- lm(mpg ~ wt + qsec + as.factor(am), data=mtcars)
```

```
fit8$coefficients
```

```
##    (Intercept)             wt           qsec as.factor(am)1
##       9.617781      -3.916504       1.225886       2.935837
```

Based on these coefficients, cars with MT have on average 2.94 more **mpg** than those with AT.

To improve the model, **am** is eliminated from fit8 or interaction between **wt** and **qsec** are introduced in fit8. The former improves the fit, the latter degrades it with respect to fit8. Based on Appendix A2, **mpg** depends on **wt** and **qsec** as a function of **am**. So introduce **am** as an interaction term but not as an independent feature. Several combinations of terms in the model were checked and the best fit quality was selected (see Appendix A3). The intercept term was removed to obtain the model

```
fit23 <- lm(mpg ~ wt:as.factor(am) + qsec:as.factor(am)-1, data=mtcars)
```

The final value of $R^2 = 0.99$ and adj.$R^2 = 0.99$ are better than that for fit8. The p-values for each coefficient (below) are also statistically quite significant

```
##   wt:as.factor(am)0   wt:as.factor(am)1 as.factor(am)0:qsec
##        3.341696e-04        4.422947e-06        1.179543e-12
## as.factor(am)1:qsec
##        1.098203e-16
```

This model has zero intercept. Consequently, a cars transmission type has no direct effect on its **mpg**.

Analysis of residuals

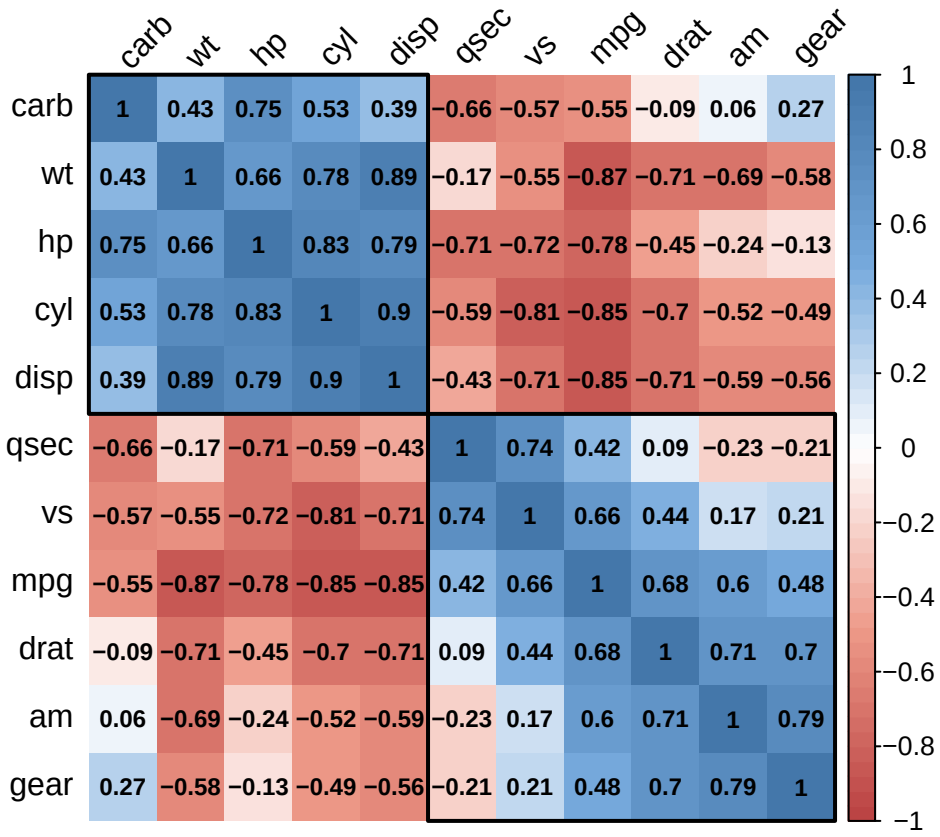
From Appendix A4, Panel B, all residuals are normally distribution. Panel A and C show the absence of any systematic patterns indicative of underlying errors in the model. Panel D shows the lack of outliers or influential points. Thus, the residuals satisfy the underlying assumptions of linear regression and the model fits the data well.

Conclusion

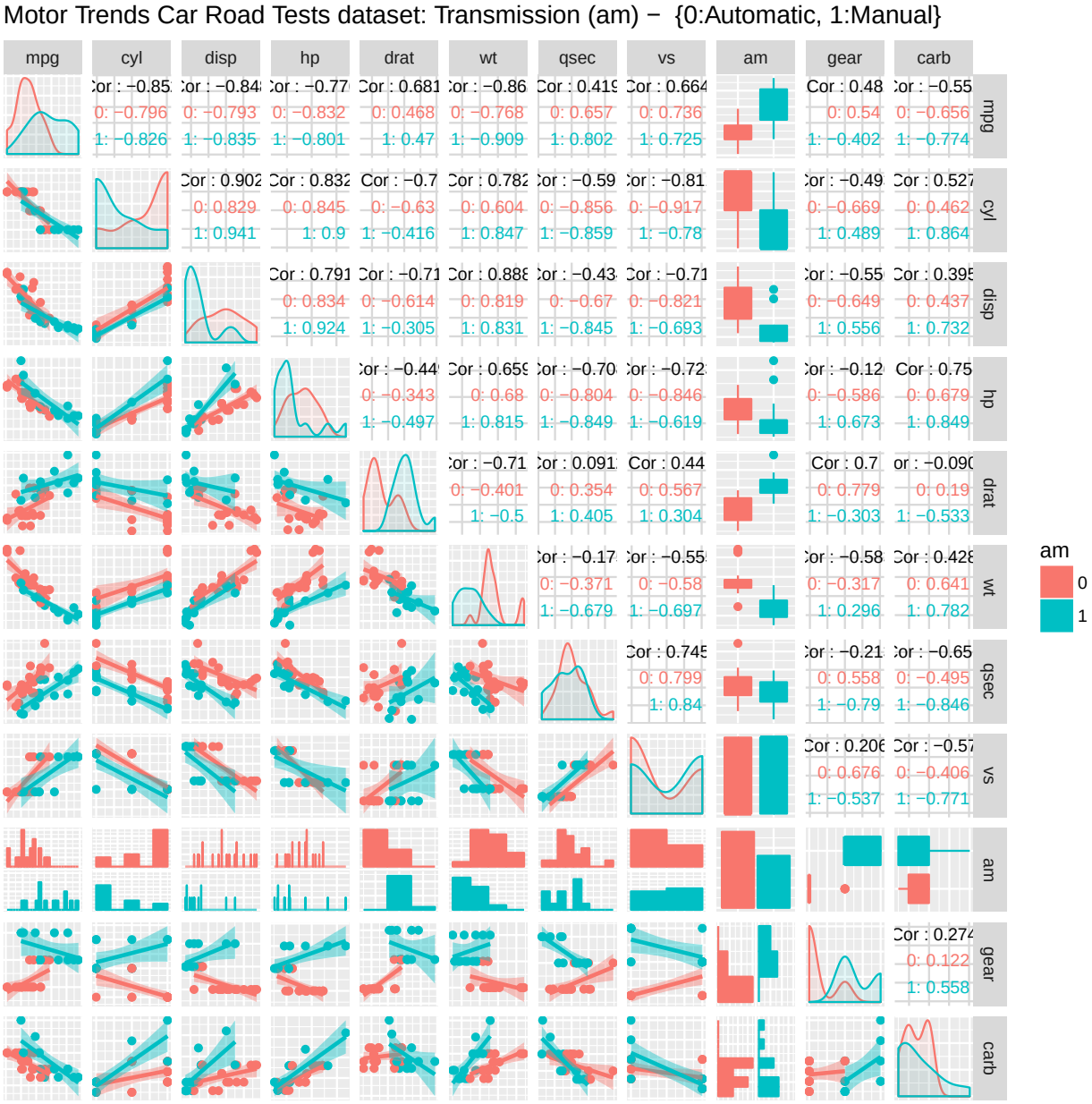
Although model dependent, the best model suggests that there is no difference in mean **MPG** between automatic and manual transmissions. Rather, the dependence of **MPG** on a vehicles weight(**wt**) and its $\frac{1}{4}$-mile time (**qsec**) is linear, although different for automatic and manual transmissions.

Appendix

A1: Correlation of pairs of variables.



A2: Variation of pairs of variables as a function of Transmission type



A3: ANOVA for regression models with interaction terms

```
## Analysis of Variance Table
##
## Model  1: mpg ~ wt + qsec + as.factor(am)
## Model  2: mpg ~ wt + qsec
## Model  3: mpg ~ wt + qsec + wt:qsec
## Model  4: mpg ~ wt + qsec + wt:factor(am) + qsec:factor(am)
## Model  5: mpg ~ wt + qsec + wt:factor(am)
## Model  6: mpg ~ wt + wt:factor(am)
## Model  7: mpg ~ qsec + wt:factor(am)
```

```
## Model  8: mpg ~ wt + qsec + qsec:factor(am)
## Model  9: mpg ~ wt + qsec:factor(am)
## Model 10: mpg ~ qsec:factor(am)
## Model 11: mpg ~ qsec + wt:factor(am) + qsec:am
## Model 12: mpg ~ qsec + wt:factor(am)
## Model 13: mpg ~ qsec + qsec:factor(am)
## Model 14: mpg ~ wt:factor(am) + qsec:factor(am)
## Model 15: mpg ~ wt + wt:factor(am) + qsec:factor(am)
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      28 169.29
## 2      29 195.46 -1   -26.178  5.9545 0.0215243 *
## 3      28 186.84  1     8.623  1.9615 0.1727384
## 4      27 118.70  1    68.141 15.4999 0.0005227 ***
## 5      28 190.24 -1   -71.540 16.2729 0.0004045 ***
## 6      29 269.76 -1   -79.519 18.0879 0.0002260 ***
## 7      28 190.24  1    79.519 18.0879 0.0002260 ***
## 8      28 161.50  0    28.743
## 9      28 161.50  0     0.000
## 10     29 328.11 -1  -166.610 37.8981 1.403e-06 ***
## 11     27 118.70  2   209.407 23.8165 1.094e-06 ***
## 12     28 190.24 -1   -71.540 16.2729 0.0004045 ***
## 13     29 328.11 -1  -137.867 31.3601 6.107e-06 ***
## 14     27 118.70  2   209.407 23.8165 1.094e-06 ***
## 15     27 118.70  0     0.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A4: Diagnostic plots for non-linear regression