Analysis of the Effect of Transmission Type on Vehicle MPG

Amir Manafpour 5/7/2020

Executive Summary

This study evalautes the affect of the type of transmission on the miles per gallon (mpg) value of the vehicles in the mtcars dataset. The results show when running regressional analyssis with MPG as the outcume and only the "am" variable as the predictor: a 7.25 increase in MPG for automatic transmission versus manual transmission is observed with a very low p-value.

However, by incorporating more relevant variables such as wt and hp, the am (transmission type) variable becomes statisically insignficant. This implies that the MPG value of vehicles doesn't depend on transmission type and that other variables can better explain the variability in MPG values.

Exploratory Data Analysis

```
data("mtcars"); df <- mtcars
```

Based on the correlation (see correlation plot) we observe the following:

- Continuous variables: mpg, disp, hp, drat, wt, qsec
- Discreate variables: vs, am, gear, carb, cyl
- High correlation (>= 0.6) between mpg and -> cyl, disp, hp, drat, wt, vs, am
- But a lot of the variables themselves are highly correlated to each other also

Regression Analysis

First a model was fit using all variables and by considering cyl, vs, am, and gear variables as factors.

```
fit <- lm(mpg ~ factor(cyl) + disp + hp + drat+ wt + qsec + factor(vs)
          + factor(am) + factor(gear) + carb, data = df)
summary(fit)$coef
```

```
##                   Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept)   15.09261548 17.13627433  0.8807408 0.38946336
## factor(cyl)6  -1.19939698  2.38736481 -0.5023937 0.62116357
## factor(cyl)8   3.05491692  4.82986776  0.6325053 0.53459525
## disp           0.01256810  0.01774024  0.7084518 0.48726645
## hp            -0.05711722  0.03174603 -1.7991927 0.08789210
## drat           0.73576811  1.98461241  0.3707364 0.71493502
## wt            -3.54511861  1.90895437 -1.8570997 0.07886857
## qsec           0.76801287  0.75221895  1.0209964 0.32008122
## factor(vs)1    2.48849171  2.54014636  0.9796647 0.33956206
## factor(am)1    3.34735713  2.28948094  1.4620594 0.16006890
## factor(gear)4 -0.99921782  2.94657533 -0.3391116 0.73824498
## factor(gear)5  1.06454635  3.02729599  0.3516492 0.72897110
## carb           0.78702815  1.03599487  0.7596834 0.45676696
```

Based on the results above, it appears weight and horsepower seems to have the lowest p-values compared to all the other variables, followed by the transmission variable (am).

To compare models, we'll start out with a model only including am, then add in the wt variable, followed by the hp variable.

wt-hp have high correlation (0.659) and so do wt-am = (-.692) so we'll include the interaction of these variables also for comparison purposes.

```
fit1 <- lm(mpg ~ factor(am), data = df)
fit2 <- lm(mpg ~ factor(am) + wt, data = df)
fit2i <- lm(mpg ~ factor(am)*wt, data = df)
fit3 <- lm(mpg ~ wt + hp + factor(am), data = df)
fit3i <- lm(mpg ~ wt*hp + wt*factor(am), data = df)

summary(fit1)$coef
```

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

```
summary(fit2)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## factor(am)1 -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

```
summary(fit3)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## factor(am)1  2.08371013 1.376420152  1.513862 1.412682e-01
```

```
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ wt + hp + factor(am)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 68.734 5.071e-09 ***
## 3     28 180.29  1     98.03 15.224 0.0005464 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit1, fit2i, fit3i)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) * wt
## Model 3: mpg ~ wt * hp + wt * factor(am)
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     30 720.90
## 2     28 188.01  2    532.89 53.57 6.007e-10 ***
## 3     26 129.32  2     58.69  5.90  0.007714 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Selection

The model only considering am as a variable (fit1) has an estimated coefficient of 7.25 for the am = 1 factor. Which would imply that automatic transmission (am = 1) has an MPG of about 7.25 higher versus manual transmission (am = 0). The p-value for fit1 also implies that the "am" variable is a statistically significant variable. However, the R-squared value is only 0.36 which implies a significan portion of the variation in MPG cannot be explained by the "am" variable alone.

By adding in the wt and hp variables, we see p-values significanly less than 0.05 by running ANOVA on the 3 nested models. The same occurs when we run the ANOVA on models including the interaction of variables too. The R-squared value for models (fit2 + fit3) including wt and hp variables are significantly higher (>0.75) which implies a lot more of the variance in MPG can be explained by including the wt and hp variables.

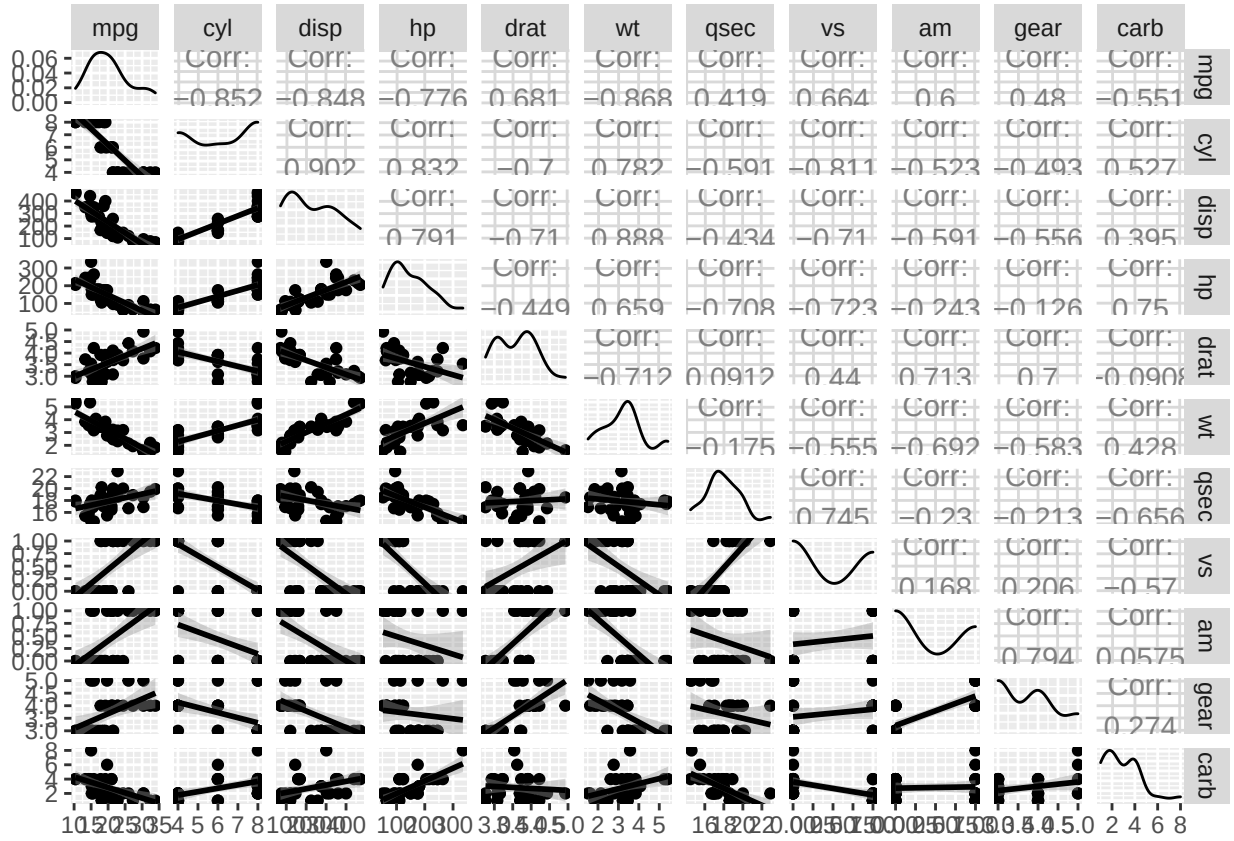
For fit2 and fit3 models the am variable becomes statiscally insignificant with p-values of 0.988 and 0.141 respectively. Ultimately this analysis implies that the type of transmission is statistically insignifcant compared to horsepower and weight.

Residuals Diagnostics

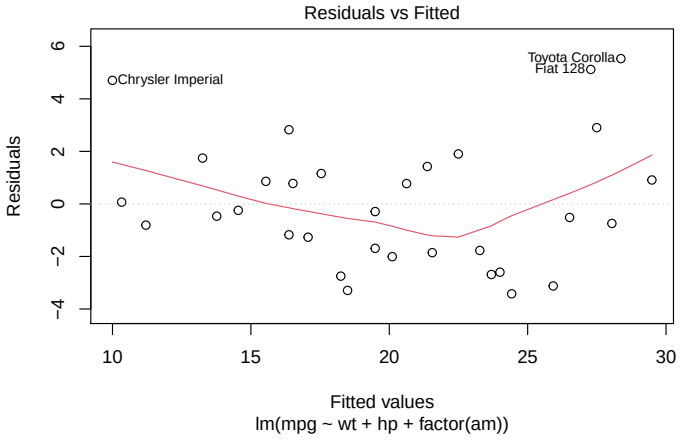
Based on the residual plots (see appendix) above we can observe three data points (Toyota Corolla, Fiat 128, and Chrysler Imperial) that could be potential outliers. These data points may have more leverage than other data points which will need further analysis before omitting them.

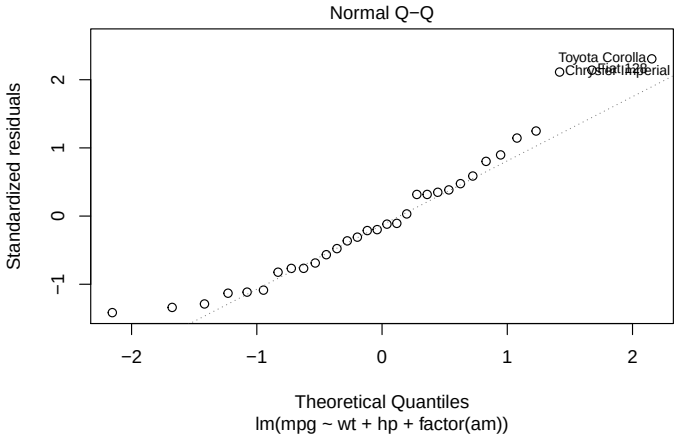
Appendix - Figures

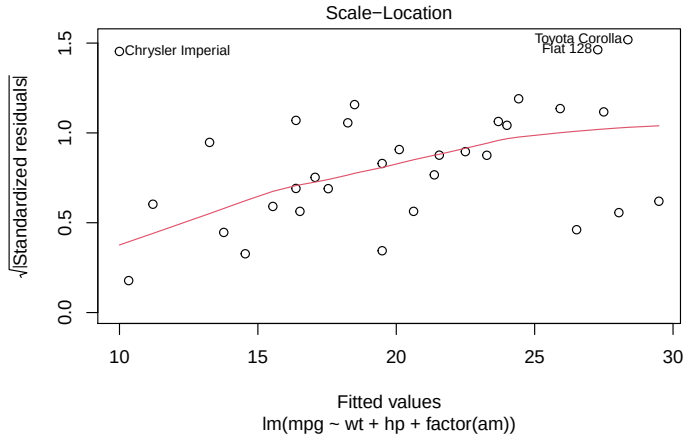
```
ggpairs(df, lower = list(continuous = "smooth"))
```

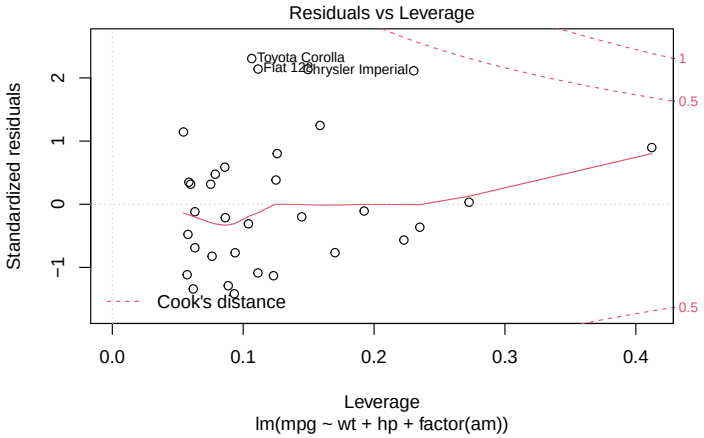



```
plot(fit3)
```






Scale−Location
√|Standardized residuals|
Chrysler Imperial
Toyota Corolla
Fiat 128
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(mpg ~ wt + hp + factor(am))


Residuals vs Leverage
Standardized residuals
Toyota Corolla
Fiat 128
Chrysler Imperial
Cook's distance
1
0.5
0.5
−1
0
1
2
0.0
0.1
0.2
0.3
0.4
Leverage
lm(mpg ~ wt + hp + factor(am))