Motor Trend - Transmission Type vs. Fuel Economy

Mark Huang

March 24, 2017

Executive Summary

The purpose of this report is to investigate the effects of transmission type on fuel economy. We are interested to uncover if an automatic or manual transmission yields better MPG readings quantity this difference. We use the `mtcars` dataset in our analysis.

Exploring the data

We load the `mtcars` dataset and analyze its structure (see *Appendix A*).

```
data(mtcars)
str(mtcars)
```

We see that `cyl, vs, am, gear, carb` must be converted to factors as their numbers don't represent magnitude and will skew the model.

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs, labels=c('V engine','Straight engine'))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

We perform a boxplot of **MPG vs. Transmission Type** (see *Appendix B*) and find that manual transmissions achieve a higher MPG.

Next, we perform a pairs plot of MPG against other variables (see *Appendix C*). Visually, we observe that `cyl, disp, hp, wt, carb` are inversely correlated with `mpg`. We also see that `vs, am, gear` are positively correlated with `mpg`.

Model Fitting and Selection

We attempt to fit an initial linear regression model `fit1` with mpg as the outcome, and the other variables as the regressors. To detect multicollinearity, we use the variance inflation factors function in R `vif(fit1)`. We use the adjusted R squared as it's more accurate.

```
fit1 <- lm(mpg ~ ., data = mtcars)
vif(fit1)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## cyl   128.120962  2        3.364380
## disp   60.365687  1        7.769536
## hp     28.219577  1        5.312210
## drat    6.809663  1        2.609533
## wt     23.830830  1        4.881683
```

```
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

```r
summary(fit1)$adj.r.squared
```

```
## [1] 0.7790215
```

We observe many highly collinear regressors in the model (VIF>4). Including all regressors yielded an adjusted R-squared value of **0.779**. We want to maximize the adjusted R-squared value while removing collinear regressors to improve the model.

We use the `step()` function working backward to find the best model.

```r
step(fit1, direction = "backward")
```

We fit a few models suggested by the `step()` function. `fit1` model has already been fitted above. We use `anova()` to compare the fitted models since they are naturally nested.

```r
fit2 <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear, data = mtcars)
fit3 <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am, data = mtcars)
fit4 <- lm(mpg ~ cyl + disp + hp + wt + qsec + vs + am, data = mtcars)
fit5 <- lm(mpg ~ cyl + hp + wt + qsec + vs + am, data = mtcars)
fit6 <- lm(mpg ~ cyl + hp + wt + vs + am, data = mtcars)
fit7 <- lm(mpg ~ cyl + hp + wt + am, data = mtcars)
a1 <- anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7)
print(a1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
## Model 4: mpg ~ cyl + disp + hp + wt + qsec + vs + am
## Model 5: mpg ~ cyl + hp + wt + qsec + vs + am
## Model 6: mpg ~ cyl + hp + wt + vs + am
## Model 7: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     15 120.40
## 2     20 134.00 -5  -13.5989 0.3388 0.8814
## 3     22 139.02 -2   -5.0215 0.3128 0.7361
## 4     23 139.99 -1   -0.9672 0.1205 0.7333
## 5     24 141.24 -1   -1.2474 0.1554 0.6990
## 6     25 143.68 -1   -2.4420 0.3042 0.5894
## 7     26 151.03 -1   -7.3459 0.9152 0.3539
```

We see from above that model `fit7` has the lowest P-value, and hence the best model. We calculate the variance inflation factors for `fit7` with **25** DF.

Model Evaluation

```r
vif(fit7)
```

```
##         GVIF Df GVIF^(1/(2*Df))
## cyl 5.824545  2        1.553515
## hp  4.703625  1        2.168784
## wt  4.007113  1        2.001778
## am  2.590777  1        1.609589
```

```r
summary(fit7)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

`fit7` has fewer regressors but a higher adjusted R-squared value of **0.84**, or **84%** of the variance explained by the model. The model estimates a **1.809** MPG increase in fuel economy when a manual transmission is used over automatic one, while holding all other regressors constant. The P-value for transmission (**0.206**) is not significant, while the `cyl, hp, wt` regressors significantly affect our prediction.

Model Visualization

We plot model `fit7` (see *Appendix D*) and observe that the **Residual vs. Fitted** values do not reveal any patterns that would suggest a poor model fit. The residuals fall roughly along the normal **Q-Q plot** suggesting a good model fit. However the **Cook's distance** line shows some outliers towards the tail of the plot that have high potential to influence the prediction.

Conclusion

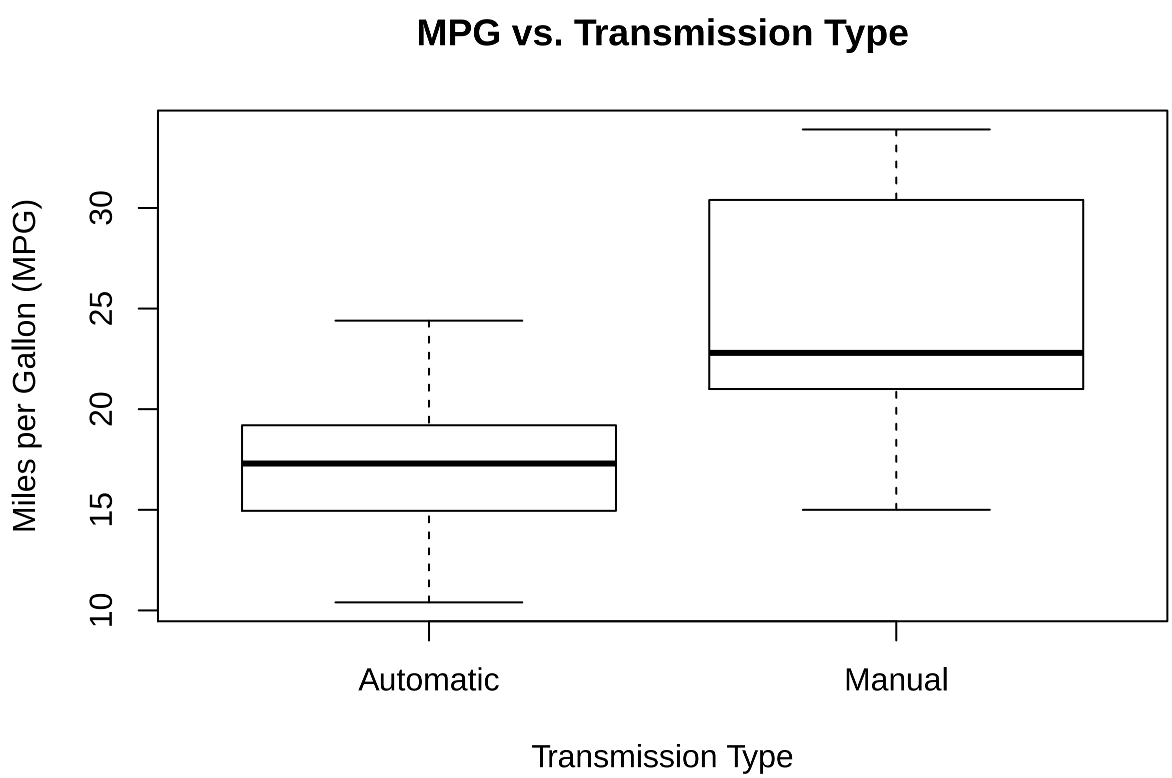
Overall, manual transmission is better by **1.809** MPG. We acknowledge that transmission type is not the only factor accounting for better MPG. Cylinders, horsepower, and weight affect MPG the most.

Appendix

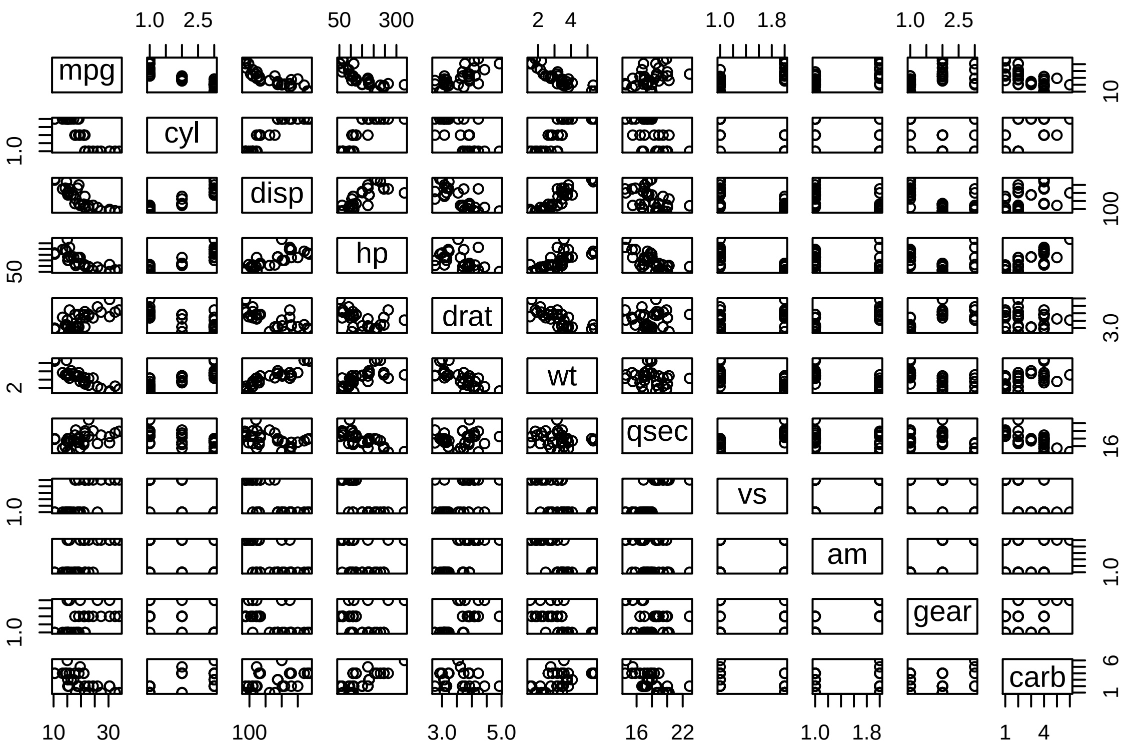
Appendix A

```
## '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 ...
```

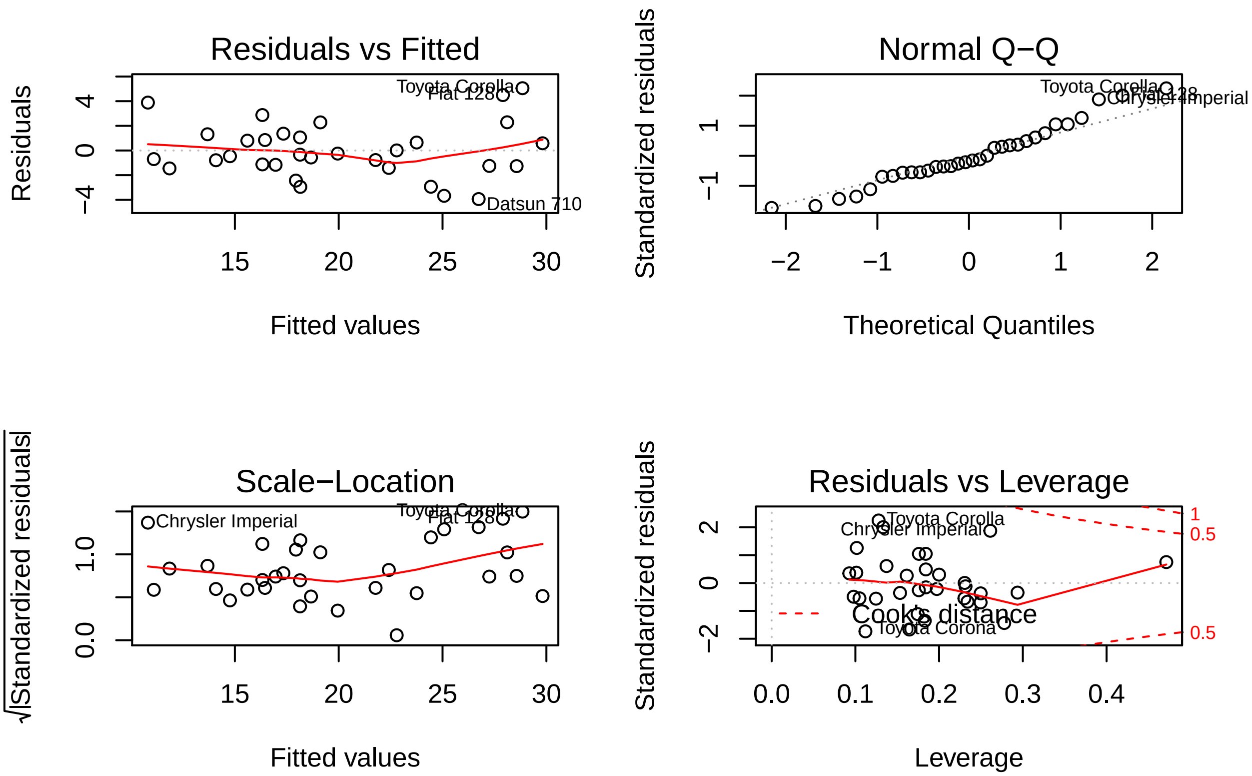
Appendix B



Appendix C


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
1.0
2.5
50
300
2
4
1.0
1.8
1.0
2.5
10
30
100
3.0
5.0
16
22
1.0
1.8
1
4
10
100
3.0
16
1.0
1 6
1.0
50
2
1.0
1.0

Appendix D



Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5