Transmissions and Fuel Efficiency

Benjamin Milks

Overview of the Data and Project

I looked at a data set mtcars and learned about the MPG in manual and automatic transmissions. I used many different models and ran an anova test. The T value for manual transmission is not high enough to conclude there is a difference between the two transmission types when other factors are held constant. The estimated difference is 1.8 MPG, but the t value was .206 which prevents us from having a claim of better fuel efficiency. To explore this further I would recommend data for new cars and a understanding of cars enough to know how to understand and quantify the outliers.

```
head(mtcars)
```

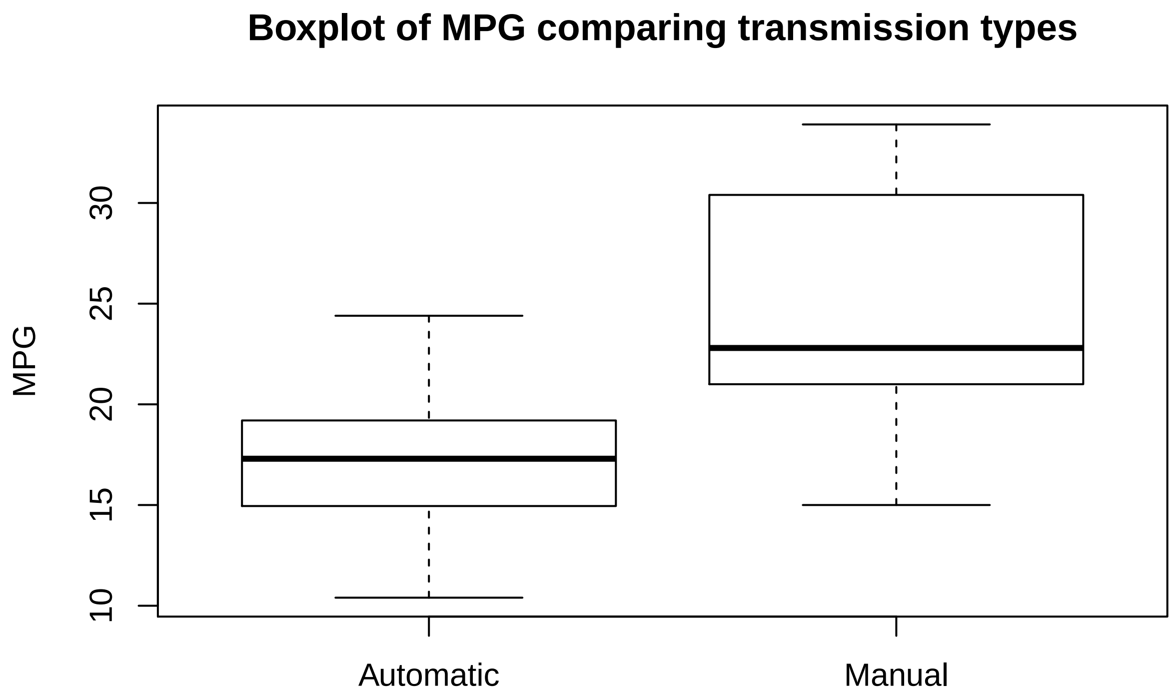
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Initial Plots to Explore

```
cor(mtcars$mpg, mtcars[-1])
```

```
##             cyl       disp         hp      drat         wt     qsec
## [1,] -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##              vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
boxplot(mpg ~ am, data = mtcars, ylab = "MPG", main = "Boxplot of MPG comparing transmission types")
```



We can see the mean mpg is higher for manual transmissions, but we can have models help us see if there are any other factors at work.

Making the Data Functional

```
mtcars$cyl <- factor(mtcars$cyl, labels = c("4 cyl", "6 cyl", "8 cyl"))
mtcars$vs <- factor(mtcars$vs, labels = c("V engine", "Straight engine"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

I changed the data that were whole numbers into different factors so R wouldn't think it is continuous data.

Initial Model

```
simplemodel <- lm(mpg ~ am, data = mtcars)
summary(simplemodel)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

Here it appears that manual transmission is siginificantly better, but this is only factoring in transmission type. It is ignoring factors like weight and horse power, which could make the plot misleading.

Looking Further

```r
allvarsmodel <- lm (mpg ~ ., data = mtcars)
summary(allvarsmodel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       23.87913   20.06582   1.190   0.2525
## cyl6 cyl          -2.64870    3.04089  -0.871   0.3975
## cyl8 cyl          -0.33616    7.15954  -0.047   0.9632
## disp               0.03555    0.03190   1.114   0.2827
## hp                -0.07051    0.03943  -1.788   0.0939 .
## drat               1.18283    2.48348   0.476   0.6407
## wt                -4.52978    2.53875  -1.784   0.0946 .
## qsec               0.36784    0.93540   0.393   0.6997
## vsStraight engine  1.93085    2.87126   0.672   0.5115
## amManual           1.21212    3.21355   0.377   0.7113
## gear4              1.11435    3.79952   0.293   0.7733
## gear5              2.52840    3.73636   0.677   0.5089
## carb2             -0.97935    2.31797  -0.423   0.6787
## carb3              2.99964    4.29355   0.699   0.4955
## carb4              1.09142    4.44962   0.245   0.8096
## carb6              4.47757    6.38406   0.701   0.4938
## carb8              7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

```r
fit1 <- lm(mpg ~ am -1, data = mtcars)
fit2 <- lm(mpg ~ am + cyl -1, data = mtcars)
fit3 <- lm(mpg ~ am + cyl +  wt -1, data = mtcars)
fit4 <- lm(mpg ~ am + cyl + wt + hp -1, data = mtcars)
fit5 <- lm(mpg ~ am + cyl +  wt + hp + vs -1, data = mtcars)
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + cyl - 1
## Model 3: mpg ~ am + cyl + wt - 1
## Model 4: mpg ~ am + cyl + wt + hp - 1
## Model 5: mpg ~ am + cyl + wt + hp + vs - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 39.7065 1.737e-08 ***
## 3     27 182.97  1     81.53 14.1856 0.0009005 ***
## 4     26 151.03  1     31.94  5.5580 0.0265310 *
## 5     25 143.68  1      7.35  1.2782 0.2689680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model gets better as the fit number gets higher. This is until fit 5 where it loses its significance code. I should not include V engine/Straight engine.

```r
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + hp - 1, 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|)
## amAutomatic 33.70832    2.60489  12.940 7.73e-13 ***
## amManual    35.51754    2.03171  17.482 6.81e-16 ***
## cyl6 cyl    -3.03134    1.40728  -2.154  0.04068 *
## cyl8 cyl    -2.16368    2.28425  -0.947  0.35225
## wt          -2.49683    0.88559  -2.819  0.00908 **
## hp          -0.03211    0.01369  -2.345  0.02693 *
## ---
## 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:  0.9892, Adjusted R-squared:  0.9868
## F-statistic: 398.6 on 6 and 26 DF,  p-value: < 2.2e-16
```

```r
sum(resid(fit4))
```

```
## [1] 2.664535e-15
```

```r
confint(fit4)
```

```
##                   2.5 %       97.5 %
## amAutomatic 28.35390366 39.062744138
## amManual    31.34129487 39.693775694
## cyl6 cyl    -5.92405718 -0.138631806
## cyl8 cyl    -6.85902199  2.531671342
## wt          -4.31718120 -0.676477640
## hp          -0.06025492 -0.003963941
```

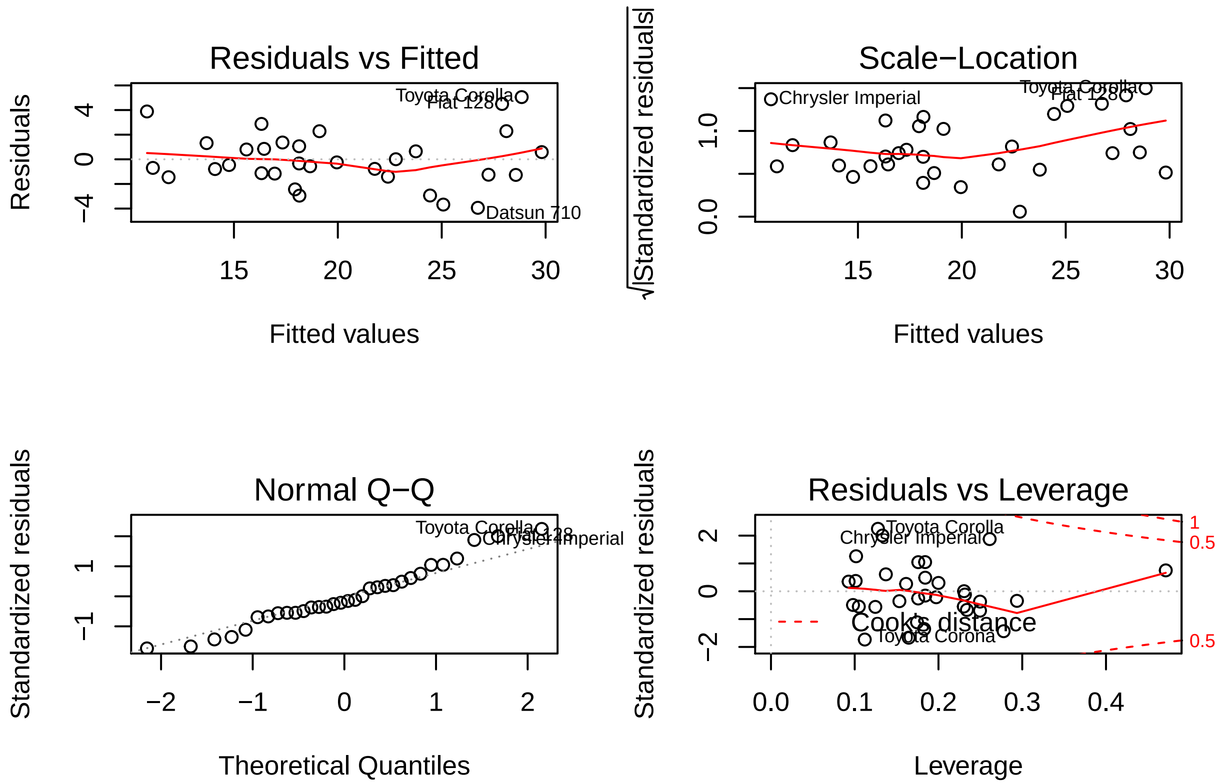
The sum of residuals is close to zero, which is a good sign for the model. Now I want to see the model with the intercept.

```r
modelwithintercept <- lm(mpg ~ am + cyl + wt + hp, data = mtcars)
summary(modelwithintercept)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + hp, 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 ***
## amManual     1.80921    1.39630   1.296  0.20646
## cyl6 cyl    -3.03134    1.40728  -2.154  0.04068 *
## cyl8 cyl    -2.16368    2.28425  -0.947  0.35225
## wt          -2.49683    0.88559  -2.819  0.00908 **
## hp          -0.03211    0.01369  -2.345  0.02693 *
## ---
## 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
```

Extra Figures

```r
par(mfcol=c(2,2))
plot(fit4)
```


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