Which Has a Better MPG, Auto or Manual Transmission?

JLei

5/14/2020

Executive Summary

This data analysis on a dataset of a collection of cars explores the relationship between a set of variables and miles per gallon (MPG)(outcome). In particular, it addresses whether an automatic or manual transmission is better for MPG, and it quantifies the MPG difference between the two transmission types. Although a manual transmission yields more MPG than an automatic trnasmission with no other variables considered, the two transmissions are not very different in MPG.

Data Analysis

We'll begin by loading the *mtcars* data and perform some basic exploratory data analysis.

```
library(datasets)
data("mtcars")
dim(mtcars)
```

```
## [1] 32 11
```

```
head(mtcars)
```

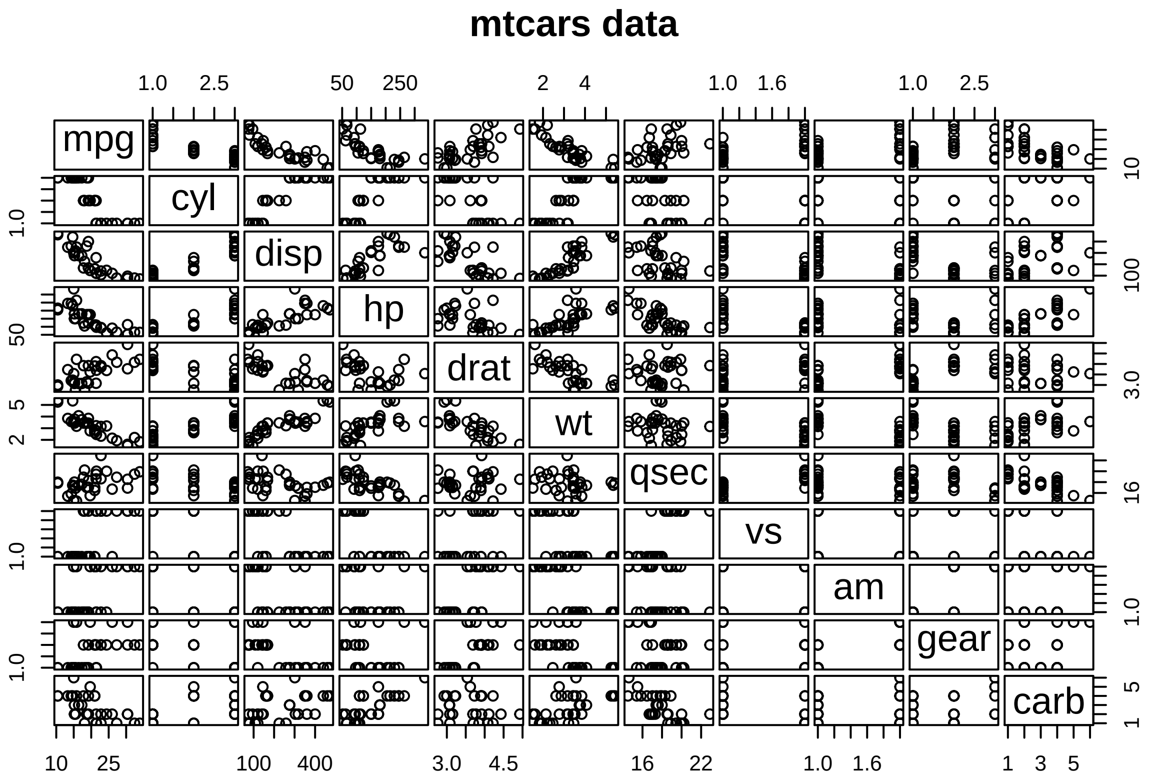
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
```

For some variables, it's more meaningful to convert their values to factors.

```
mtcars2 <- within(mtcars, {
        vs <- factor(vs, labels = c("V", "S"));
        am <- factor(am, labels = c("automatic", "manual"));
        cyl  <- factor(cyl);
        gear <- factor(gear);
        carb <- factor(carb)})
summary(mtcars2)
```

```
##       mpg        cyl         disp            hp             drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec       vs             am     gear   carb
##  Min.   :1.513   Min.   :14.50   V:18   automatic:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   S:14   manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                         5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1
##  Max.   :5.424   Max.   :22.90                                8: 1
```

```
pairs(mtcars2, main = "mtcars data", gap = 1/4)
```



Linear Regression Since the paired plots show that all other variables have an impact on car's MPG, let's fit a linear model with MPG as outcome to all the other variables.

```
library(broom)
lmfit <- lm(mpg ~ ., data = mtcars2)
tidy(lmfit)
```

```
## # A tibble: 17 x 5
##    term        estimate std.error statistic p.value
##    <chr>          <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)  23.9      20.1        1.19    0.253
##  2 cyl6         -2.65      3.04      -0.871   0.397
##  3 cyl8         -0.336     7.16      -0.0470  0.963
##  4 disp          0.0355    0.0319     1.11    0.283
##  5 hp           -0.0705    0.0394    -1.79    0.0939
##  6 drat          1.18      2.48       0.476   0.641
##  7 wt           -4.53      2.54      -1.78    0.0946
##  8 qsec          0.368     0.935      0.393   0.700
##  9 vsS           1.93      2.87       0.672   0.512
## 10 ammanual      1.21      3.21       0.377   0.711
## 11 gear4         1.11      3.80       0.293   0.773
## 12 gear5         2.53      3.74       0.677   0.509
## 13 carb2        -0.979     2.32     -0.423   0.679
## 14 carb3         3.00      4.29      0.699   0.495
## 15 carb4         1.09      4.45      0.245   0.810
## 16 carb6         4.48      6.38      0.701   0.494
## 17 carb8         7.25      8.36      0.867   0.399
```

The coefficient for variable *am* when it's manual transmission is **1.212** with **p=0.71**, indicating that there is no significant evidence to reject that an automatic and manual transmission yields similar MPG. The manual transmission produces 1.212 MPG more than the automatic transmission, but the p-value of 0.71 sugests that it's not significant with 95% confidence level.

Then, let's investigate which covariate is inflating the variance and can be excluded in the next model fit.

```
library(car)
```

```
## Loading required package: carData
```

```
sqrt(vif(lmfit))
```

```
##           GVIF       Df GVIF^(1/(2*Df))
## cyl  11.319053 1.414214        1.834225
## disp  7.769536 1.000000        2.787389
## hp    5.312210 1.000000        2.304823
## drat  2.609533 1.000000        1.615405
## wt    4.881683 1.000000        2.209453
## qsec  3.284842 1.000000        1.812413
## vs    2.843970 1.000000        1.686407
## am    3.151269 1.000000        1.775181
## gear  7.131081 1.414214        1.634138
## carb 22.432384 2.236068        1.364858
```

```
anova(lmfit)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 51.3766 1.943e-07 ***
## disp       1  57.64   57.64  7.1813   0.01714 *
## hp         1  18.50   18.50  2.3050   0.14975
## drat       1  11.91   11.91  1.4843   0.24191
## wt         1  55.79   55.79  6.9500   0.01870 *
## qsec       1   1.52    1.52  0.1899   0.66918
## vs         1   0.30    0.30  0.0376   0.84878
## am         1  16.57   16.57  2.0639   0.17135
## gear       2   5.02    2.51  0.3128   0.73606
## carb       5  13.60    2.72  0.3388   0.88144
## Residuals 15 120.40    8.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

cyl, *disp*, *hp*, *gear* are considered variance inflation factor; however, analysis of variance anova() of the model fit *lmfit* shows that *cyl*, *disp*, and *wt* are significant in cars' MPG. Hence we will include *cyl*, *disp*, and *wt* in the next model fit to better investigate the impact of transmission (*am*).

Next, let's fit a linear model with MPG regressing on transmission type (*am*), *cyl*, *disp*, and *wt*.

```
library(broom)
fit1 <- lm(mpg ~ cyl+disp+wt+am, data = mtcars2)
tidy(fit1)
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 33.8        2.91      11.6   8.79e-12
## 2 cyl6        -4.30       1.49      -2.88  7.77e- 3
## 3 cyl8        -6.32       2.65      -2.39  2.46e- 2
## 4 disp         0.00163    0.0138     0.119 9.06e- 1
## 5 wt          -3.25       1.25      -2.60  1.51e- 2
## 6 ammanual     0.141      1.33       0.106 9.16e- 1
```

Holding constant the number of cylinders (*cyl*), displacement (*disp*), and weight (*wt*), we don't see much effect of transmission type on MPG. A manual transmission car would have **0.14** MPG more than an automatic tranmission car, and **p=0.91** suggests it's not significant to reject that a manual transmission is not different from an automatic when it comes to MPG.

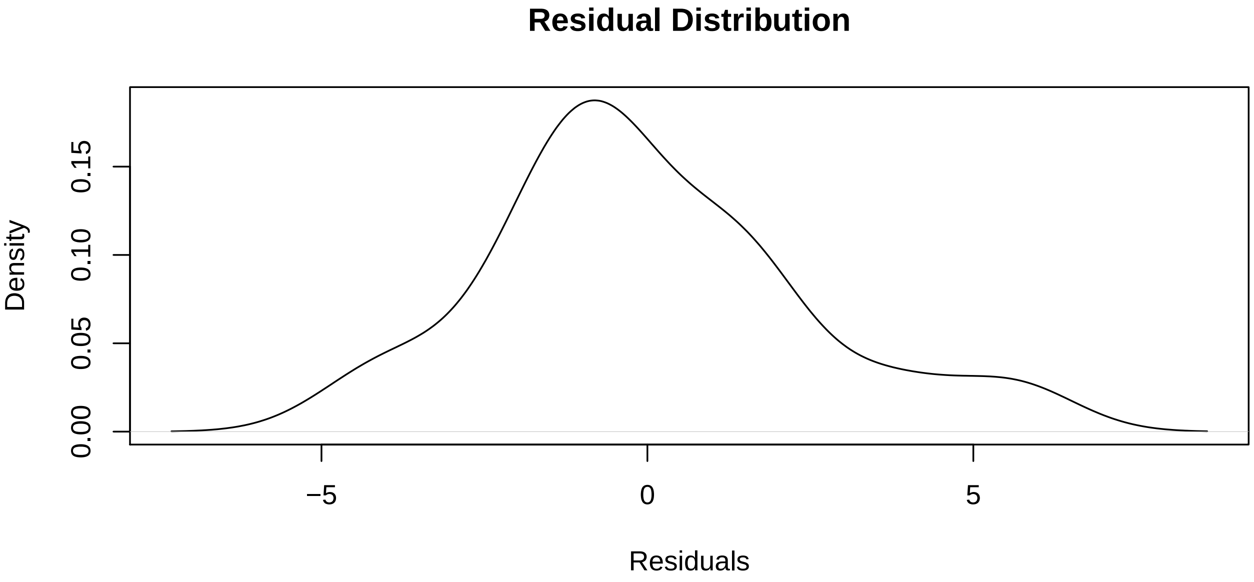
```
anova(fit1, lmfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + wt + am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 182.87
## 2     15 120.40 11    62.467 0.7075 0.7153
```

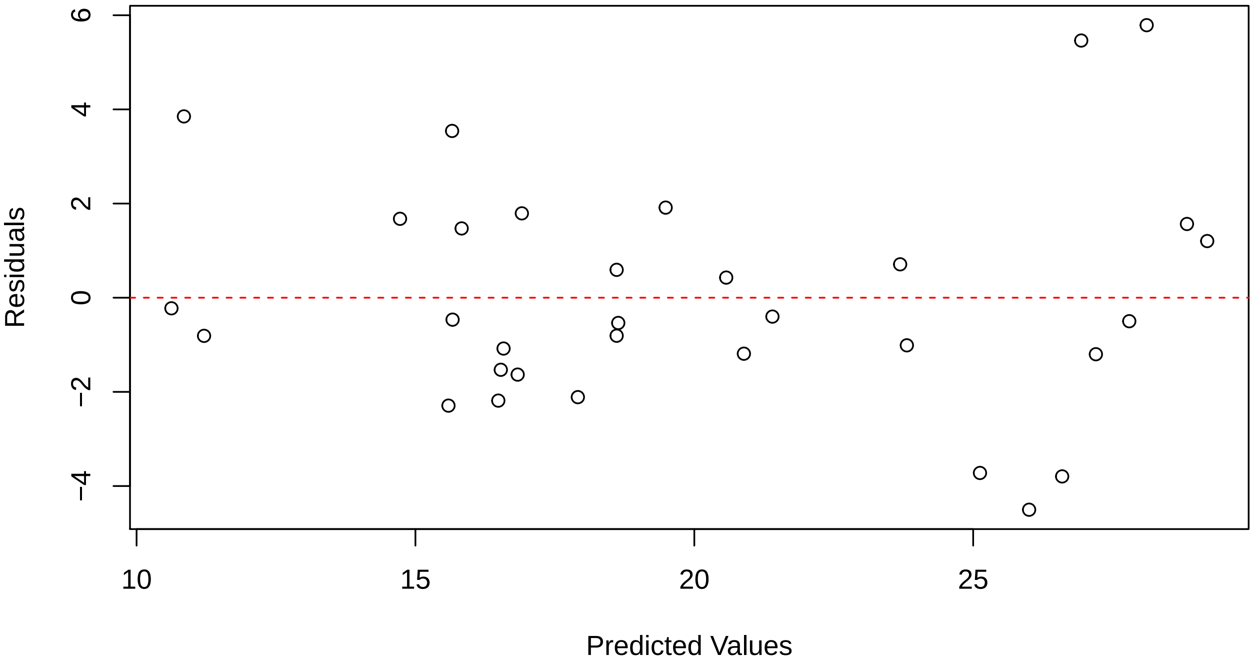
p=0.72 suggests that the additional variables in *lmfit* is not necessary to include for analyzing the impact of transmission on MPG.

Now, let's check if regression assumptions are met with some diagnostic plotting:

```
resid <- residuals(fit1)
fitted <- fitted.values(fit1)
plot(density(resid), xlab="Residuals", ylab="Density", main="Residual Distribution")
```



```
plot(fitted, resid, xlab="Predicted Values", ylab="Residuals")
abline(h=0, col="red", lty="dashed")
```



Normality assumptions don't seem far off, and heteroskedasticity doesn't seem to be an issue.

Overall, there doesn't seem to be an effect of transmission type on MPG with this dataset of a collection of cars.

[**Optional Read**] If we only fit MPG to transmission type:

```
fit2 <- lm(mpg ~ am, data = mtcars2)
tidy(fit2)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    17.1       1.12     15.2  1.13e-15
## 2 ammanual        7.24      1.76      4.11 2.85e- 4
```

The coefficient for variable *ammanual* when it's manual transmission is **7.24** with **p=0.00029**, indicating that there is significant evidence to reject that an automatic and manual transmission yields similar MPG. The manual transmission produces 7.24 MPG more than the automatic transmission, and the p-value of 0.00029 sugests that it's significant with 95% confidence level.

BUT, WATCH OUT FOR SIMPSON'S PARADOX!