Motor Trend Data: MPG, Auto vs. Manual Transmission

Chris Gillespie

July 18, 2016

Executive Summary

By analyzing the "mtcars" data set in R, the analysis aims to answer two questions:

1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

Regression tests will be used to examine the relationship between MPG and transmission type.

Exploratory Analysis

First off, is there a MPG difference between auto and manual?

```
#Factor + rename transmission column
mtcars$am <- factor(mtcars$am)
levels(mtcars$am) <- c("automatic","manual")

#Run a basic regression
summary(lm(mpg ~ am -1, mtcars))$coeff
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## amautomatic 17.14737   1.124603 15.24749 1.133983e-15
## ammanual    24.39231   1.359578 17.94109 1.376283e-17
```

Yes, it is obvious that manual cars have better gas milage, as much as 7 mpg if you don't factor in other variables.

Is it because manual cars have less weight, horsepower, etc? Perhaps most manual transmission cars are built in a fundamentally different way.

With that mind, let's look into which other regressors may cause a difference.

```
manual <- subset(mtcars, am == 'automatic')
auto <- subset(mtcars, am == 'manual')

#Regressor differences
sapply(manual,mean,1)-sapply(auto,mean,1)
```

```
##    mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear
##  -5.50   4.00 155.50  66.00  -0.93   1.20   0.80  -1.00     NA  -1.00
##   carb
##   1.00
```

Let's start adding new regressiors to compare with the original fit.

```
#Run anova to see if extra variables are helpful or not
fit1 <- lm(mpg ~ am -1, mtcars)
fit2 <- lm(mpg ~ am -1 + wt, mtcars)
fit3 <- lm(mpg ~ am -1 + wt + hp + cyl, mtcars)
fit4 <- lm(mpg ~ am -1 + wt + hp + cyl + disp + gear + carb, mtcars)
anova(fit1,fit2,fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am - 1 + wt
## Model 3: mpg ~ am - 1 + wt + hp + cyl
## Model 4: mpg ~ am - 1 + wt + hp + cyl + disp + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 66.2444 2.323e-08 ***
## 3     27 170.00  2    108.32  8.1067  0.002042 **
## 4     24 160.34  3      9.65  0.4817  0.698091
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It looks like adding extra regressors is useful for most of the main differences (weight, horsepower, cylinders), but once you start adding too many of them, it no longer becomes helpful.

For the best model (fit3), the residual plot indicates there aren't any outliers causing skewed data. It seems like a good choice.

Final analysis

```
#What is the difference between manual and auto?
fit3$coefficients[2]-fit3$coefficients[1]
```

```
## ammanual
## 1.478048
```

Based on the regression model's summary (see appendix), the model is statistically significant:

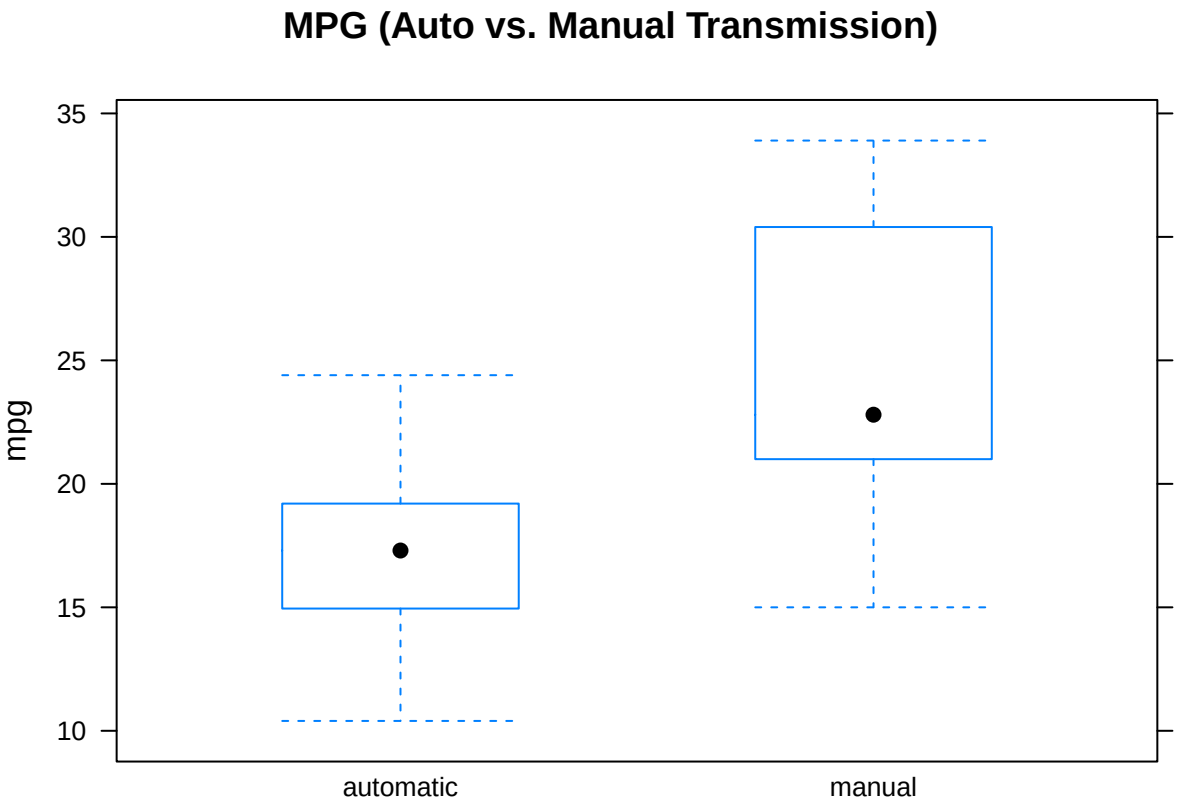
1. Coefficient p-values are less than .05 for transmission.
2. F-statistic for the model is basically zero.

Thus, based on our analysis, **manual transmission is better than auto by 1.48 MPG**.

Appendix

MPG: Auto vs. Manual

```r
#Plot out the difference
bwplot(mpg ~ am, data = mtcars,
       main = "MPG (Auto vs. Manual Transmission)")
```

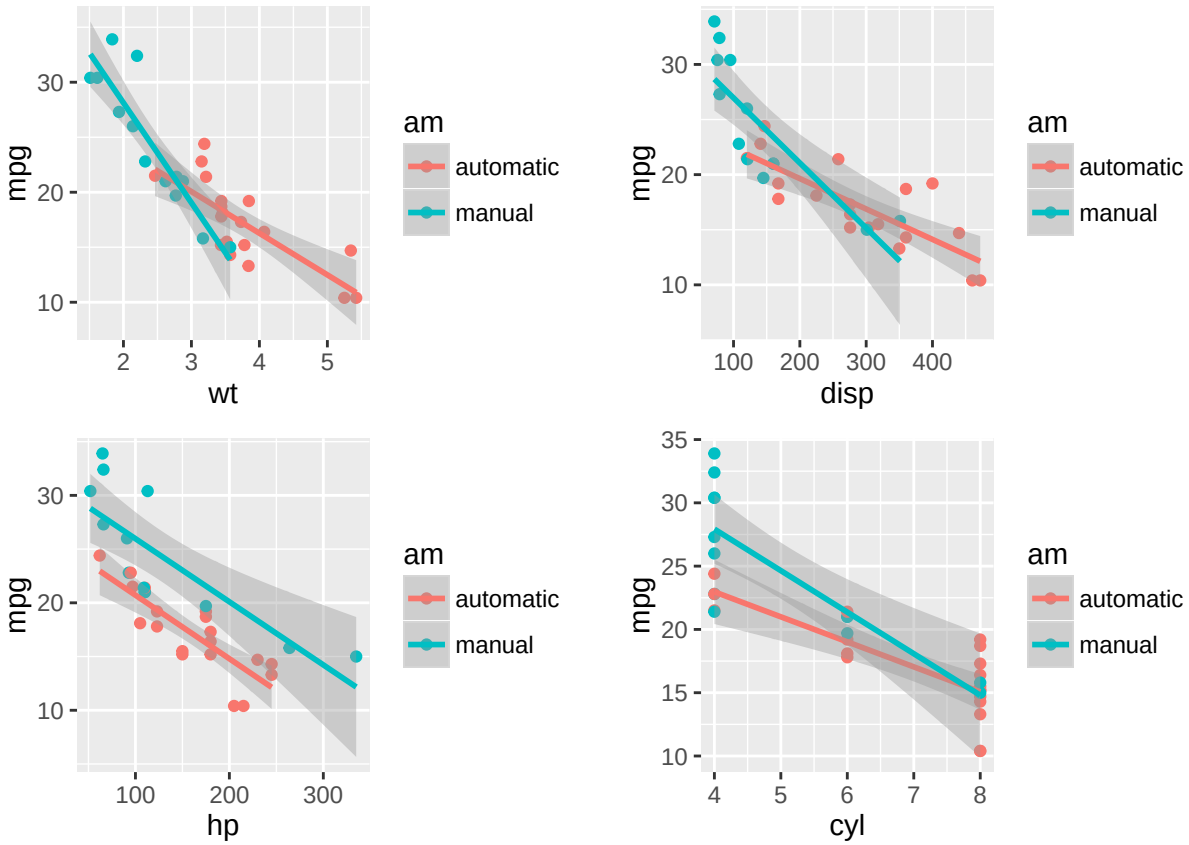


Regressor Analysis

```r
#Visualize the most obvious differences as a 2x2 grid
require(gridExtra)
```

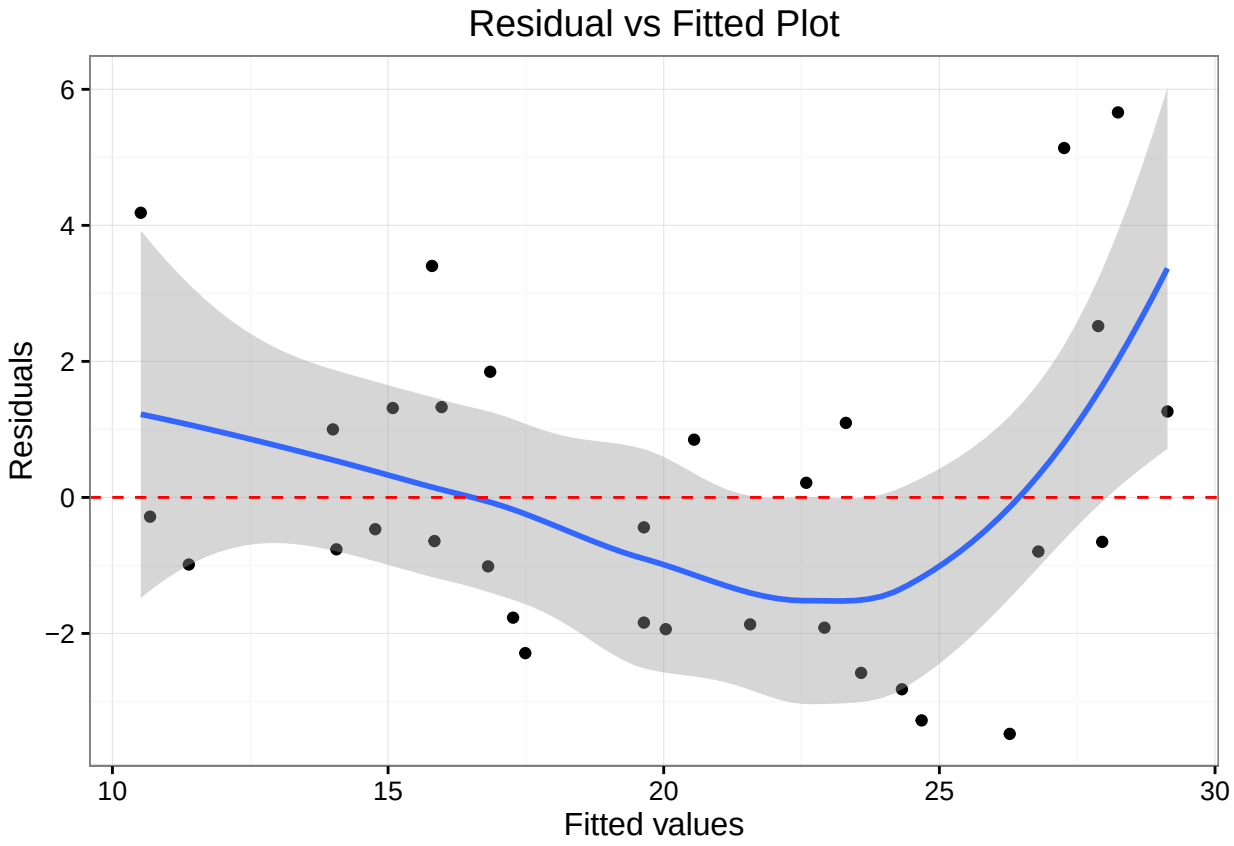
```
## Loading required package: gridExtra
```

```r
g1 <- ggplot(mtcars, aes(x=wt, y=mpg, color = am))+geom_point()+
  geom_smooth(method = lm)
g2 <- ggplot(mtcars, aes(x=disp, y=mpg, color = am))+geom_point()+
  geom_smooth(method = lm)
g3 <- ggplot(mtcars, aes(x=hp, y=mpg, color = am))+geom_point()+
  geom_smooth(method = lm)
g4 <- ggplot(mtcars, aes(x=cyl, y=mpg, color = am))+geom_point()+
  geom_smooth(method = lm)
grid.arrange(g1,g2,g3,g4, ncol = 2)
```



Residual Plot

```
ggplot(fit3, aes(.fitted, .resid))+geom_point()+
  stat_smooth(method="loess")+geom_hline(yintercept=0, col="red", linetype="dashed")+
  xlab("Fitted values")+ylab("Residuals")+ggtitle("Residual vs Fitted Plot")+theme_bw()
```



Regression Summary

```
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am - 1 + wt + hp + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## amautomatic 36.14654    3.10478  11.642 4.94e-12 ***
## ammanual    37.62458    2.09641  17.947  < 2e-16 ***
## wt          -2.60648    0.91984  -2.834   0.0086 **
## hp          -0.02495    0.01365  -1.828   0.0786 .
## cyl         -0.74516    0.58279  -1.279   0.2119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.9879, Adjusted R-squared:  0.9857
## F-statistic: 440.7 on 5 and 27 DF,  p-value: < 2.2e-16
```