Mileage Differences between Automatic and Manual Transmissions

T. Bennett

August 23, 2015

Executive Summary

This document describes an analysis of the U.S. National Oceanic and Atmospheric Administration (NOAA)'s Storm Events database from 1950 to November 2011. The goals of the analysis were to assess the 1) population health and 2) economic impact of severe storm events.

As one might expect based on geography, the types of storm events in each U.S. state vary widely. Government officials and planners need to account for local events in their use of this analysis.

Heat events cause the most deaths and injuries per event. Tropical depressions/storms and hurricanes cause nearly as many injuries per event, but fewer deaths.

Overall, flood events cause the most property damage, and are close behind drought events in causing the most crop damage. Tropical storms/depressions and hurricanes cause the most property and crop damage per storm event, and the second-most property damage overall, despite the fact that they are rare. Other rainstorm and tornado events calso cause significant property damage, and winter storms also cause major crop damage.

Data Processing

1) Loading packages

```
library(dplyr)
library(ggplot2)
library(stringr)
library(xtable)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.2.2
```

2) Loading the raw data

```
## load and inspect the data
data(mtcars)
str(mtcars)
```

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

```
glimpse(mtcars)
```

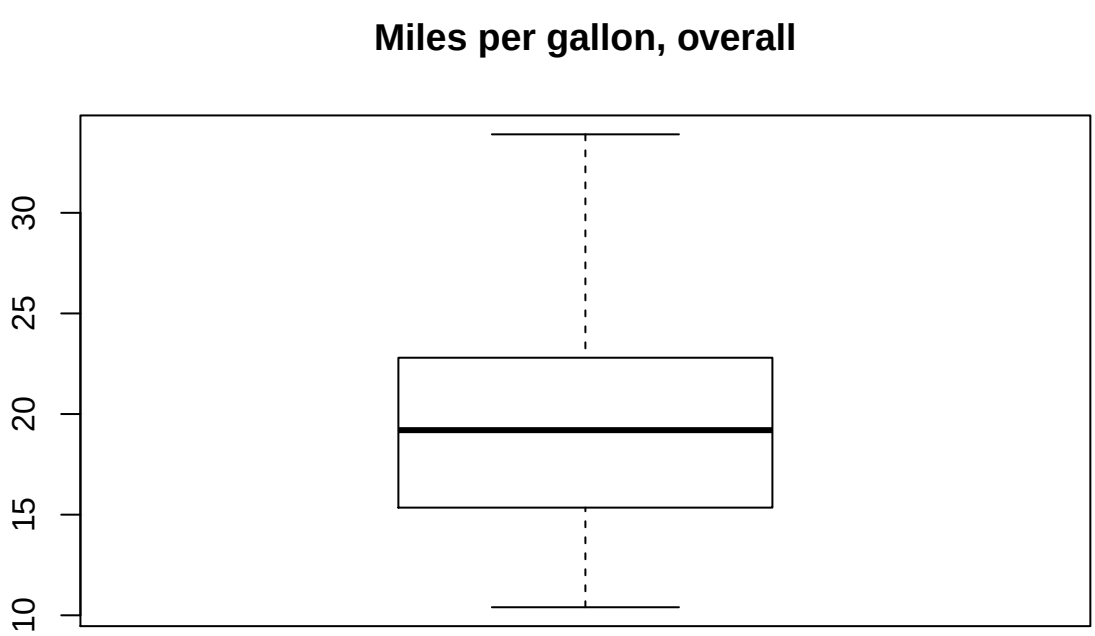
```
## Observations: 32
## Variables:
## $ mpg  (dbl) 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19....
## $ cyl  (dbl) 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, ...
## $ disp (dbl) 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 1...
## $ hp   (dbl) 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, ...
## $ drat (dbl) 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.9...
## $ wt   (dbl) 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3...
## $ qsec (dbl) 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 2...
## $ vs   (dbl) 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, ...
## $ am   (dbl) 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ...
## $ gear (dbl) 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, ...
## $ carb (dbl) 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, ...
```

This analysis is based on the `mtcars` dataset included in the `datasets` package in base R. It contains data on the 1973-74 models of 32 different automobiles including fuel consumption and 10 aspects of automobile design. Total 32 rows and 11 fields, no missing data.

Exploratory Data Analyses

Miles per gallon is the primary outcome.

```
with(mtcars, boxplot(mpg, main = "Miles per gallon, overall"))
```



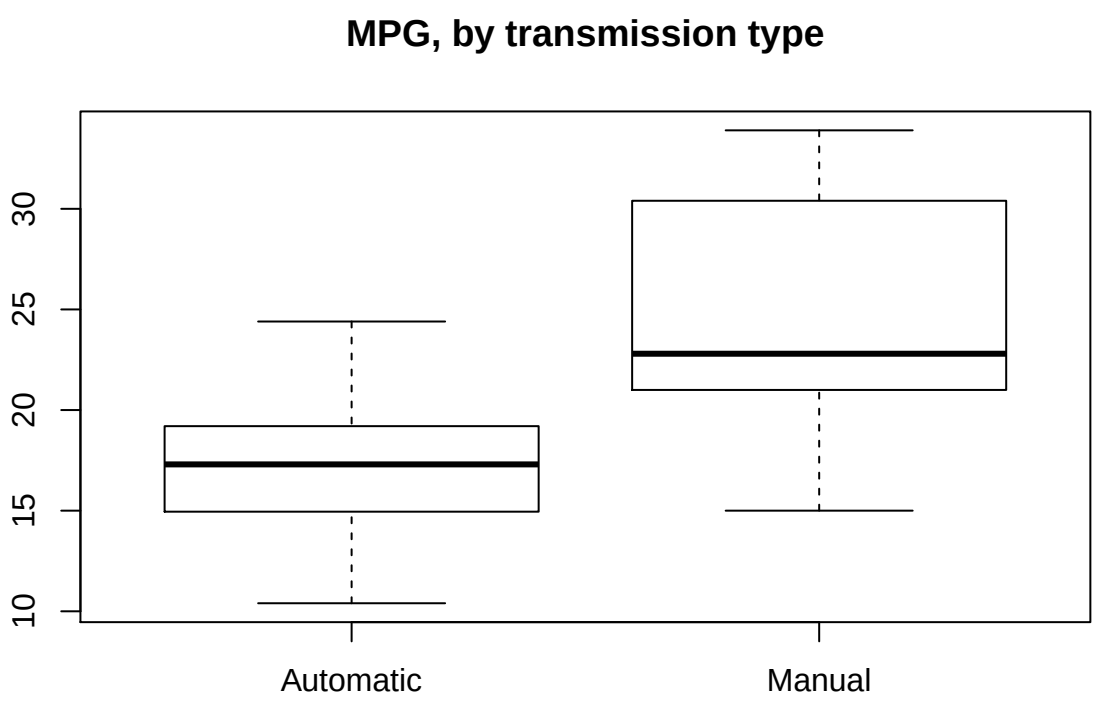

```
mtcars <- mtcars %>%
    mutate(trtype = factor(am, labels = c("Automatic", "Manual"))) %>%
    select(-am)

tab <- xtable(table(mtcars$trtype))
tab
```

% latex table generated in R 3.2.1 by xtable 1.7-4 package % Fri Aug 21 15:43:49 2015

	V1
Automatic	19
Manual	13

Transmission type is the primary predictor of interest. 0.40625 have a manual transmission and `19/32` have a manual transmission.

```
boxplot(mpg ~ trtype, main = "MPG, by transmission type", data = mtcars)
```



A bivariate plot does suggest that manual transmissions are associated with better gas mileage.

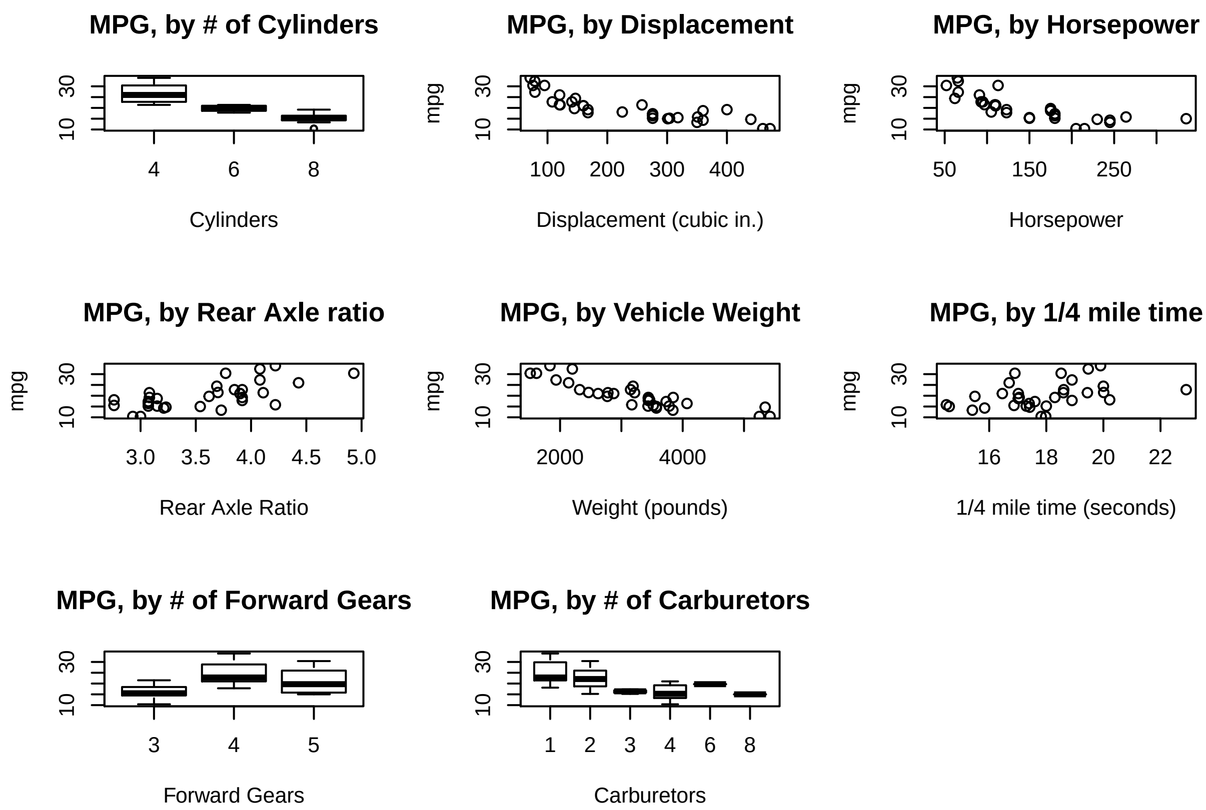
```
res <- t.test(mpg ~ trtype, var.equal=FALSE, data = mtcars)
```

A t-test of `mpg` using unequal variances bears that out: cars with automatic transmissions get 17.1473684 miles per gallon and cars with manual transmissions get 24.3923077 miles per gallon, t-test $p = 0.0013736$.

Potential Confounders

Several other variables might confound the Transmission Type - Mileage relationship. Based on my working knowledge of automobile engines, any of the other 9 variables in the `mtcars` dataset might be hypothesized to be important. The below bivariate plots of the 9 variables against `mpg` suggest that any of them could reasonably be a predictor.

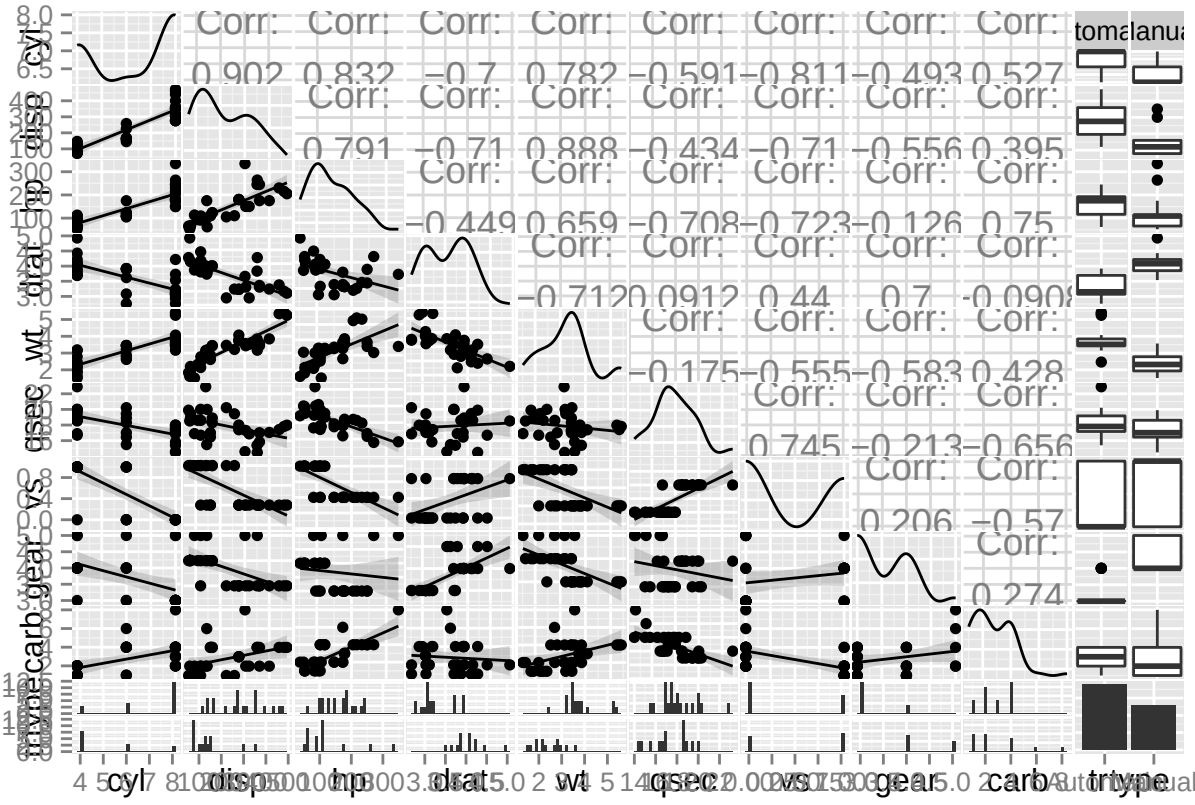
```
par(mfrow = c(3,3))
boxplot(mpg ~ cyl, main = "MPG, by # of Cylinders", data = mtcars, xlab = "Cylinders")
with(mtcars, plot(disp, mpg, main = "MPG, by Displacement", xlab = "Displacement (cubic in.)"))
with(mtcars, plot(hp, mpg, main = "MPG, by Horsepower", xlab = "Horsepower"))
with(mtcars, plot(drat, mpg, main = "MPG, by Rear Axle ratio", xlab = "Rear Axle Ratio"))
with(mtcars, plot(wt*1000, mpg, main = "MPG, by Vehicle Weight", xlab = "Weight (pounds)"))
with(mtcars, plot(qsec, mpg, main = "MPG, by 1/4 mile time", xlab = "1/4 mile time (seconds)"))
boxplot(mpg ~ gear, main = "MPG, by # of Forward Gears", data = mtcars, xlab = "Forward Gears")
boxplot(mpg ~ carb, main = "MPG, by # of Carburetors", data = mtcars, xlab = "Carburetors")
par(mfrow = c(1,1))
```



To avoid including two covariates that are highly correlated, I created the `pairs` plot below.

```
p <- ggpairs(mtcars, columns = 2:11, lower = list(continuous = "smooth"), params = c(method = "loess"))
p
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Number of cylinders, horsepower, and displacement are highly correlated (which makes sense, I suppose). I decided to include displacement in the model selection process and exclude the other two variables because displacement is a continuous variable that reflects a physical property of the car rather than a performance property of the car. Weight is also highly correlated with displacement, but not quite as highly correlated with the other two variables, so I will consider weight during model selection.

Linear Regression including all potential predictors

```
##
## Call:
## lm(formula = mpg ~ I(1 * (trtype == "Manual")) + disp + drat +
##     wt + qsec + vs + gear + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8664 -1.1040 -0.2339  1.2073  4.3409
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.274748  13.092785   0.556   0.5838
## I(1 * (trtype == "Manual"))  2.578423   1.944461   1.326   0.1978
## disp                          0.002155   0.013415   0.161   0.8738
## drat                          1.069650   1.523091   0.702   0.4895
## wt                           -3.177158   1.776312  -1.789   0.0869 .
## qsec                          0.963394   0.678444   1.420   0.1690
```

```
## vs                          -0.077889   1.900199  -0.041   0.9677
## gear                         0.673222   1.368025   0.492   0.6273
## carb                        -0.710393   0.635666  -1.118   0.2753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.595 on 23 degrees of freedom
## Multiple R-squared:  0.8624, Adjusted R-squared:  0.8145
## F-statistic: 18.02 on 8 and 23 DF,  p-value: 3.374e-08
```

Model Selection