Transmission Choice And Its Impact on MPG

Barb Dornseif - Saoirsegirl

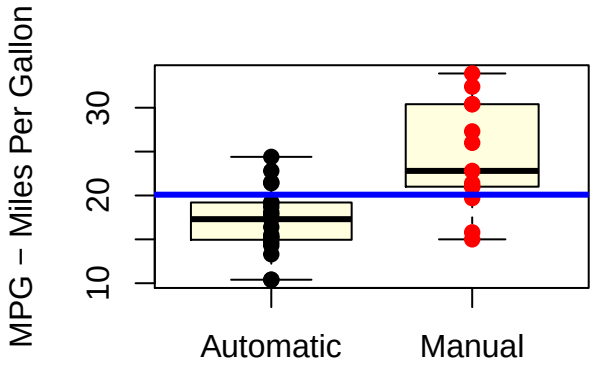
Executive Summary

Does the choice of transmission in your car impact the Miles Per Gallon (MPG) you should expect? And by how much? Given a data set of 11 elements describing ten performance related design features of 32 automobiles and their associated MPG we must conclude that choice of transmission does has a reliable impact on MPG with slightly more than 90% of confidence that a configuration with preduce this expected result. The following analysis will break down the data and process used to make this detirmination. To add readability all the code and tabular data outputs are attached in the appendix.

The Data Set

The data set 'mtcars' was compiled from an issue of Motor Trend in 1974 and included as a standard dataset in the R package. Let's load the data and take a quick look at the simple relationships between MPG (with a mean of 20.09) and transmission type as well as the other features to see how correlated they may be.

We see that all of the features are numeric and their respective correlations to MPG span a wide range both positive and negative. Given that transmission is a type and has no intrinsic numeric value, let's transform the data into a Factor variable so we can separate the transmission types into two separate variables in our subsequent models. This will help us identify the distribution of mpg grouped by the transmission type and compare each against the mean of all cars together.



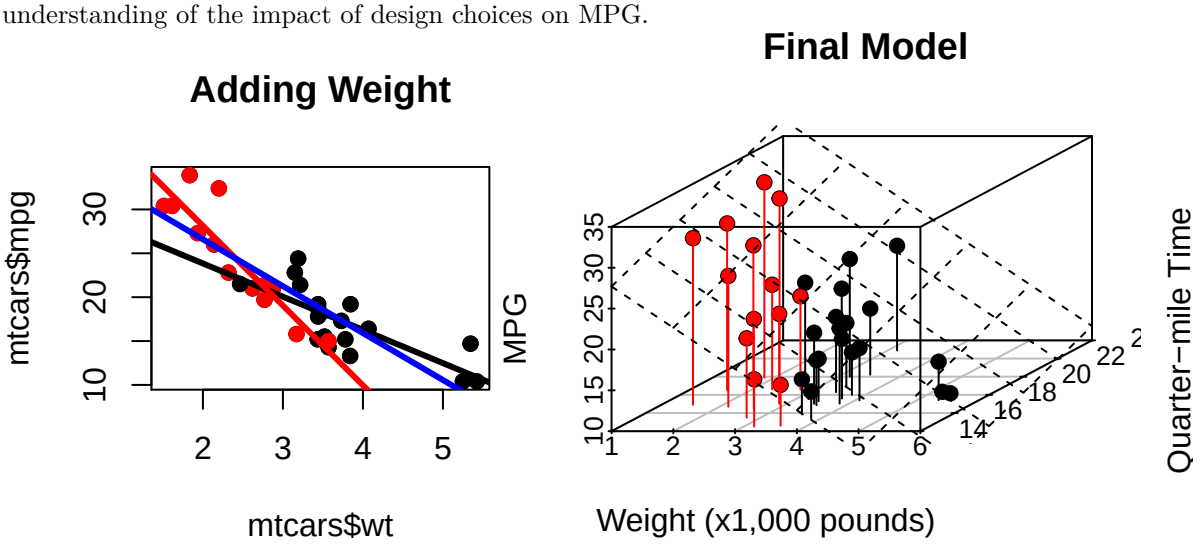
We can clearly see that while there is a distinct difference in the means for each transmission type, straddling the mean of all cars. The 25% of values that are greater than or less than the mean - the interquartile range - overlap significantly. This suggests that one or more of the nine other design features must have an impact on MPG. But it is unlikely that all variables are descriptive and the more design choices we consider, the more difficult it is to understand which design choices are most impactful on MPG. Running a model on all variables will tell us alot, so let's start there.

Fitting The Linear Model

First, let's remove the intercept from our model so we can see the change in MPG for each variable more clearly. *lm(mpg ~ . -1, data = mtcars)* The resulting table of coefficients shows that one variable, wt (weight in 1,000's of pounds), has the highest t-value in the model -1.96. The t-value is a measure of how untrue the claim that the coeffiecient is zero - or not impactful. A t-value greater than or equal to ~2 are considered impactful with 95% confidence. Looking at the list (in the Appendix) we see that only one more variable

shows a t-value that is worth considering for inclusion (1.12), so we will take an additive approach. Adding two is easier than subtracting seven to finding our best model.

Adding wt to a simple model gives us the formula *lm(mpg ~ wt * am -1 , data = mtcars)*, and we will compare it to only using weight *lm(mpg ~ wt , data = mtcars)* . Let's see if the resulting model improves our understanding of the impact of design choices on MPG.



The plot on the left shows us that when a car has a manual transmission (red line and dots), it is much more impacted by increases in weight with a decrease of ~9 MPG per 1,000 pounds. Whereas an automatic (in black) will lose only 3.8 MPG per 1,000 pounds. The blue line represents the change related only to weight. This graph clearly shows that larger cars will naturally want to having automatic transmissions at around 3,000 pounds to decrease this effect and thus their MPG as a group will be lower based on their size more than the type of transmission they have.

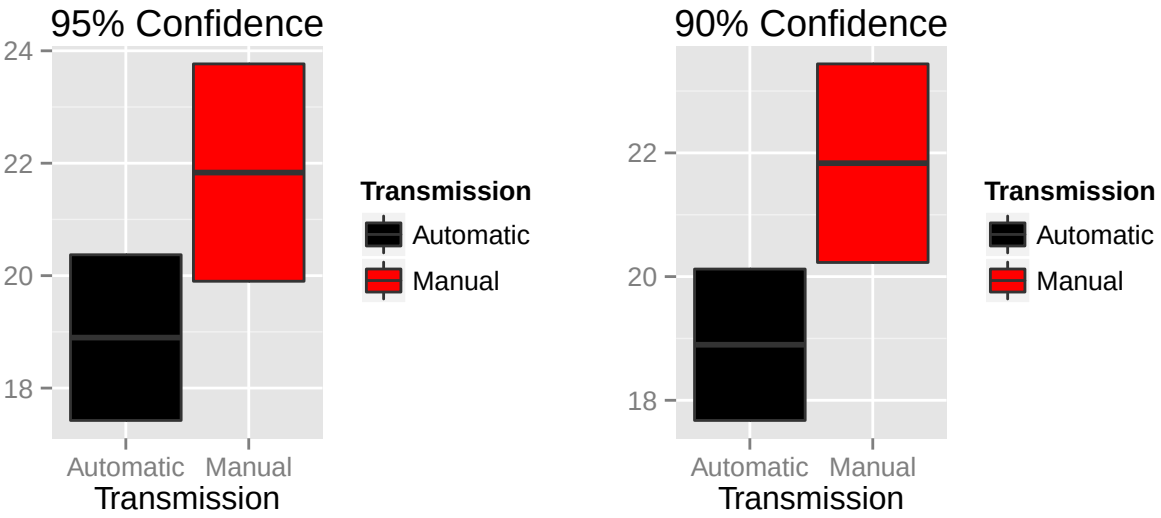
The second variable *qsec* is a measurement of time it takes to cover a quarter mile of road. This is a complex variable as it takes into account a host of design choices - but basically faster means more energy is required to cover the distance and thus MPG is lower overall. or more simply Slower = Better MPG. By adding this variable we lower the residual variance of the model and explain a few of the outliers in our simpler am + wt model. Smaller fast cars vs. small economy cars and heavy sports cars vs. big trucks.

The graph on the right shows the three numeric variable with the data points colored by transmission type. The 3-dimensional slope plane is akin to the blue line on the left with the *qsec* impact added to it. This model has a residual sum of squares - the measure of how our model explains the data in total - that is better than our transmission only or weight + transmission models. In fact it is more descriptive than our overly complex ten variable model (which doesn't graph well!) So we will stick with this model to evaluate the impact of transmission on MPG.

Confidence of Our Prediction

If we use our model to predict MPG given a set of values for weight and quarter-mile time (we used the mean of both), we will get an answer and a range of values that explains the range of possible MPG values within the expected variance of the model. Depending on how confident we want to be that the range will include the truth, the range will be narrower or wider. For us to declare that transmission does have a clear impact on MPG, the ranges should NOT overlap.

We see that our model says that at a 95% confidence transmission fails the overlap test - so we must conclude that tranmission type does not reliably impact MPG. However, if we lower our expectations, we see that at 90% confidence, it does pass the test - with no overlap to the estimated MPG for our test Automobile.



Appendix

Here are the code blocks that generated the above paper, with the echo=FALSE and/or eval=FALSE removed, and with the # notation removed to allow tabled data to be presented. Where a code block will reprint a graph, the code has been ## - commented out below to avoid redundancy.

```
data(mtcars);  x <- vector(mode = "numeric"); r <- vector(mode = "numeric")
for (i in 1:11) x <- c(x, class(mtcars[,i]))
for (j in 1:11) r <- c(r, round(cor(mtcars[j], mtcars$mpg),4))
dataSum <- rbind(names(mtcars), x, r) # # removed for Appendix
```

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
par(mfrow = c(1,1))
## boxplot(mpg ~ am, data = mtcars, ylab = "MPG - Miles Per Gallon",
        varwidth = TRUE, col = "light yellow")
points(mtcars$am, mtcars$mpg, pch=19, col= mtcars$am)
abline(h = mean(mtcars$mpg), col='blue', lwd = 3)
fitTrans <- lm(mpg ~ am, data = mtcars)
```

```
fitAll <- lm(mpg ~ . -1, data = mtcars)
summary(fitAll)$coefficients #used during developement
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## amAutomatic 12.30337416 18.71788443  0.6573058 0.51812440
## amManual    14.82360104 18.35265170  0.8077090 0.42831299
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```r
fitwt <- lm(mpg ~ wt, data = mtcars)
fit169 <- lm(mpg ~ wt * am -1 , data = mtcars) # index values for variables used
summary(fit169)$coefficients # used for dev and Appendix
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## amAutomatic 31.416055  3.0201093 10.402291 4.001043e-11
## amManual    46.294478  3.0101489 15.379465 3.488923e-15
## wt:amManual -5.298360  1.4446993 -3.667449 1.017148e-03
```

```r
## plot(mtcars$wt, mtcars$mpg, pch=19, main = "Adding Weight")
## points(mtcars$wt, mtcars$mpg, pch=19, col= mtcars$am)
## abline(c(fit169$coeff[2], fit169$coeff[1]), col="Black", lwd=3)
## abline(c(fit169$coeff[3], fit169$coeff[1] + fit169$coeff[4]), col="red", lwd=3)
## abline(c(fitwt$coeff[1], fitwt$coeff[2]), col="blue", lwd=3)
```

```r
fit1679.res = resid(fit1679)
## plot(mtcars$mpg, fit1679.res, ylab="Residuals", xlab="MPG", main="Model lm(mpg ~ wt + qsec + am -1)"
## abline(0, 0)                  # the horizon
```

```r
wtAve <- mean(mtcars$wt)
qsecAve <- mean(mtcars$qsec)
newdataM = data.frame(am="Manual", wt=wtAve, qsec=qsecAve)
newdataA = data.frame(am="Automatic", wt=wtAve, qsec=qsecAve)
manRange95 <- predict(fit1679, newdataM, interval="confidence", level = .95)
autoRange95 <-predict(fit1679, newdataA, interval="confidence", level = .95)
manRange90 <-predict(fit1679, newdataM, interval="confidence", level = .90)
autoRange90 <- predict(fit1679, newdataA, interval="confidence", level = .90)
par(mfcol = c(1,2))
d95 <- data.frame(rbind(manRange95[1:3], autoRange95[1:3]))
d95 <- cbind(c("Manual","Automatic"), d95)
colnames(d95) <- c("Transmission", "Fit","Min","Max")
print(d95)
```

```
##   Transmission      Fit      Min      Max
## 1       Manual 21.83378 19.90054 23.76702
## 2    Automatic 18.89794 17.42441 20.37147
```

```r
library(ggplot2) ; library(gridExtra)
plot1 <- ggplot(d95, aes(x=Transmission, ymin=Min, lower=Min,fill=Transmission, middle=Fit, upper=Max, y
    geom_boxplot(stat="identity") + scale_fill_manual(values=c("black", "red")) +
    labs(title="95% Confidence")

d90 <- data.frame(rbind(manRange90[1:3], autoRange90[1:3]))
d90 <- cbind(c("Manual","Automatic"), d90)
colnames(d90) <- c("Transmission", "Fit","Min","Max")
print(d90)
```

```
##   Transmission      Fit      Min      Max
## 1       Manual 21.83378 20.22829 23.43927
## 2    Automatic 18.89794 17.67423 20.12166
```

```
plot2 <- ggplot(d90, aes(x=Transmission, ymin=Min, lower=Min,fill=Transmission, middle=Fit, upper=Max,
    geom_boxplot(stat="identity") + scale_fill_manual(values=c("black", "red")) +
    labs(title="90% Confidence")
## grid.arrange(plot1, plot2, ncol=2)
```