# Transmission Type and MPG in *Motor Trend* Data

*Justin Dallmann*

*8/6/2017*

## Executive summary—a tale of two models

In this analysis, we examine 1974 *Motor Trend* data to establish whether transmission type has an effect on fuel milage. Due to the sample size and nature of the data, no (reliable) conclusion can be drawn about the relationship of transmission type to mpg. In particular, the natural strategy of using a linear model is difficult to justify (most of the variables are strongly correlated and the residuals of the models that they give rise to have non-normal residuals).

In what follows, I present the model that best meets the constraints of linear regression given the data. The best model takes mpg to be dependant on transmission type (*trans*), quarter mile time in seconds (*q.mile*), and number of carburator barrels (*carb*). This model suggests that a manual transmission is better for mpg. *However*, I also show the effects of weight as a possible confounder diminishing our confidence in the model.

**Given the nature of the data, we cannot be confident that mpg is predicted by transmission type (and, thus, we cannot place much weight on our best model's conclusion that with all of the other variables fixed, switching from automatic transmission to manual yeilds an `8.435` increase in mpg).**

## Exploratory work

In order to examine the relationship between transmission type (trans) and miles per gallon (mpg), I began by looking at the distribution of mpg by manual and auto transmission (see figure 1 in § Appendix). On first glance, it looks like automatic transmissions have worse mpg ratings.

However, of the cars in the data set, it was largely economy cars that had manual transmissions (which you might reasonably expect to have better mpg ratings). This might reasonably be screened off by several of the correlated variables.

Setting that aside for the moment, to pick the best extra regressor for a linear model, we examine the correlation between transmission and the other variables. Some of the measurements for the least correlated (in terms of Spearman's rho), and the most highly correlated variables are as follows:

Table 1: Spearman's rho correlations with trans type

| | rho | p-value |
|---|---|---|
| Carburator barrels | $-6.437 \times 10^{-2}$ | $7.264 \times 10^{-1}$ |
| Quarter mile | $-2.033 \times 10^{-1}$ | $2.644 \times 10^{-1}$ |
| Horsepower | $-3.623 \times 10^{-1}$ | $4.156 \times 10^{-2}$ |
| Displacement | $-6.241 \times 10^{-1}$ | $1.352 \times 10^{-4}$ |
| Weight | $-7.377 \times 10^{-1}$ | $1.454 \times 10^{-6}$ |

Note that number of carburator barrels, quarter mile time, and horsepower are low correlation variables that could screen off the economy/non-economy confounder described above on preliminary visual examination. See figure 1 in § Appendix.

# Modeling

Preliminary examination of the relationship between mpg and transmission type suggests a relationship at $\alpha = .005$, with manual transmission predicting better mileage:

```
t.test(mpg ~ trans, data = mtcars, conf.level = .99)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by trans
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -12.769128  -1.720751
## sample estimates:
##   mean in group auto mean in group manual
##             17.14737             24.39231
```

In order to assess possible confounders, we construct a linear model and extend it, first by adding quarter mile time, then number of carburator barrels. The three (nested) linear regression models are as follows:

1. mpg as predicted by transmission type alone;
2. mpg as predicted by transmission type and quarter mile time;
3. mpg as predicted by transmission type, quarter mile time, and number of carburator barrels.

Finally an extended model to test for confounding of the relationship between transmission type and mpg is constructed using weight as possible confounder.

### mpg = f(transmission)

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## transmanual  7.244939   1.764422  4.106127 2.850207e-04
```

### mpg = f(transmission, q.mile (qsec), carb)

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  0.3252711  8.6306964  0.0376877 9.702041e-01
## transmanual  8.4353493  1.1492472  7.3398913 5.423664e-08
## qsec         1.1332876  0.4246104  2.6690057 1.251399e-02
## carb        -1.3828531  0.4579391 -3.0197318 5.349452e-03
```

### mpg = f(transmission, weight)

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## transmanual -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

The nested model suggests that manual transmissions are better for fuel efficeny in mpg than manual ones. On the other hand, the correlated confounder model suggests that weight screens off the effect of transmission on mpg rating, but that model risks violating the assumptions of regression (see diagnostics below).

## Diagnostics

Issues to examine for sucessful linear regression are:

1. that the relationship be approximately linear/the influence of outliers is minimal;
2. multivariate normality of the residuals;
3. no or little multicollinearity/variance inflation;
4. No auto-correlation;
5. homoscedasticity.

### 1. Outliers

The cook's distances are all less than 1, so outliers are not a substantial problem. For example, the cooks distance for the final model is `0.135`.

### 2. multivariate normality

In order to see if the linear regression modeling assumptions are met, I looked at Q-Q plots to check for normality, and a plot of residuals vs. predictions to check for independance/constant variance. See Figure 2 of the Appendix. The plot looks fairly normal for the best model, but might violate it for the confounding model.

This is also borne out by a Shapiro-Wilks test with null = normality, which provides weak evidence for normality of residuals in the best model.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm(mpg ~ trans + qsec + carb, data = mtcars))
## W = 0.96368, p-value = 0.3452
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm(mpg ~ trans + wt, data = mtcars))
## W = 0.94478, p-value = 0.1024
```

### 3. No auto-correlation

Checks for indepenence of residuals using the Durbin Watson test reveal little auto-correlation.

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1016508      1.672734   0.178
##  Alternative hypothesis: rho != 0
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3628498      1.252011    0.01
##  Alternative hypothesis: rho != 0
```

### 4. multicolinearity and variance inflation

In addition to the correlation checks performed when narrowing in on the models considered above, we also checked the models' variance inflation factors:

Table 2: Varience inflation factors

| | trans | q.mile | carb | weight |
|---|---|---|---|---|
| mpg = f(trans, 1/4 mile) | 1.056 | 1.056 | | |
| mpg = f(trans, 1/4 mile, carb) | 1.073 | 1.879 | 1.785 | |
| mpg = f(trans, weight) | 1.921 | | | 1.921 |

While the VIFs are not particularly large there is a worry that the three variable model inflates the varience too much. Still, it is worth noting that the choice of quarter mile time and carburator barrels make up a triad of variables with lower overall VIFs.

**5. homoscedasticity**

Checked by a plot of residuals vs. fitted values in the diagnositics section of the Appendix.

## Anova

Since the diagnostic Q-Q plots suggest that the studentized residuals are approximately normal for the nested models, we proceed with an anova analysis to assess whether or not the extensions help prediction:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ trans
## Model 2: mpg ~ trans + qsec
## Model 3: mpg ~ trans + qsec + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 352.63  1    368.26 38.7640 9.982e-07 ***
## 3     28 266.00  1     86.63  9.1188  0.005349 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the anova analysis suggest that each extension of the model captures an important aspect of prediction compared to the previous at $\alpha = .01$.

Likewise, the anova for the confounding model suggests that each variable is predictive (though in this case the normality assumption is more questionable). In this case the p-value for the added predictive power of weight beyond transmission type is `1.867e-07`.

# Appendix

**Figure 1: Exploratory analyses**

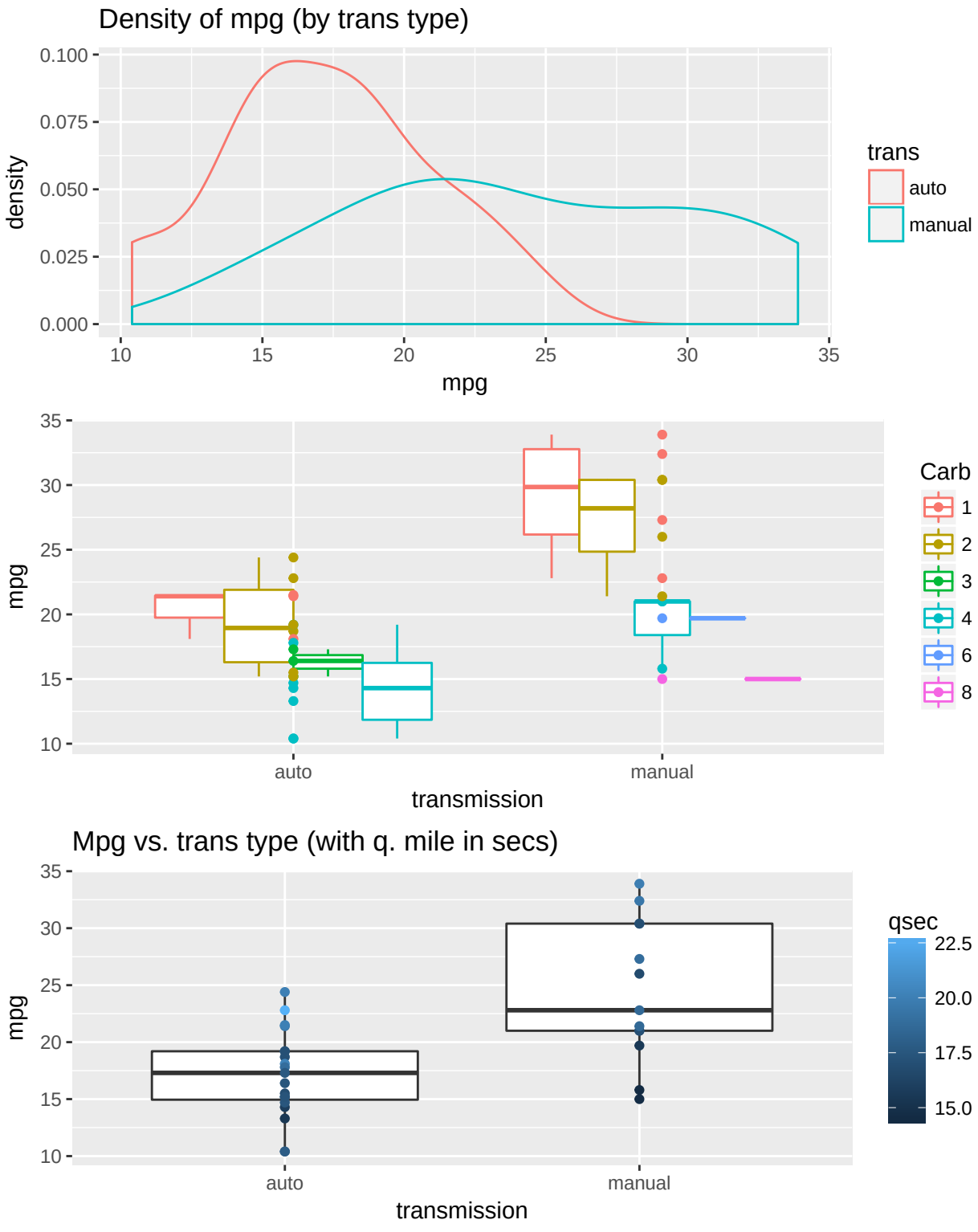

**Figure 2: Diagnostics**