Transmission Type has no Impact on Car Fuel Consumption

Tim Wise

January 2016

Executive Summary

We looked at the Motor Trend cars data set to see if the type of transmission in a car had a significant effect on its fuel consumption, and if so, to quantify that effect.

At first glance, we found that cars with manual transmissions used less fuel, consuming on average 4.4 gallons per 100 miles (gpm) (+/- 1) compared to an average of 6.1 gpm (+/- 0.7) for cars with automatic transmissions. *But correllation is not causation* and just because there is a difference in fuel consumption doesn't mean transmission type is the cause of the difference.

Digging deeper, we found that a car's fuel consumption was primarily a function of its weight and that the type of transmission was, in the end, not significant. The reason why cars with manual transmissions used less fuel was because in this data set the transmission type and vehicle weight were confounders and were not independent. The high mileage cars were light and had manual transmissions, while the low mileage cars were heavy and had automatic transmissions.

Analysis Details

We did two transformations on the cars data set. First, we inverted and scaled miles per gallon (mpg) to get gallons per 100 miles. *Gallons per kilomile (gpm)* is a better metric for comparing the fuel consumption among cars. Also, we added a column, dividing horsepower (hp) by weight (wt) to get a the *horsepower per 1000 lbs*. As we will see, this gives us a a measure of power that is not correlated with weight.

First let's compare the fuel consumption of cars with automatic transmissions to those with manual transmissions. **Figure 1a** shows there is a sizable difference in the fuel consumption of manuals verses automatics. We verify the difference is significant by doing an intercept-only regression of fuel consumption as a function of transmission type. See **Figure 2** for an interpretation of the regression results. In short, manuals use less fuel, consuming on average 4.4 gpm (+/- 1) compared to an average of 6.1 gpm (+/- 0.7) for automatics.

Digging deeper, in **Figure 3**, we see that fuel consumption is most correlated to weight. Plotting fuel consumption verses weight in **Figure 4**, we see transmission type is not spread evenly over weight. (We see another view of that relationship in **Figure 1b**.) This reason for the significant difference in fuel consumption between automatics and manuals. We also see, looking at the two trend lines in **Figure 4**, that if we account for weight, there doesn't seem to be much difference in between the fuel consumption of automatics and manuals. Let's build a regression model to investigate that.

Starting with a base model of fuel consumption as a function of weight, we will iterate in a forward step-wise fashion, using the function `add1()` to help decide which regressors to include. `add1()` produces an F-statistic and a p-value (like an `anova()` comparison) for each possible regressor. We look for p-values < 0.05 and high F values.

Figure 5 shows the output of `add1()` given the base model (gpm ~ wt). Horsepower/weight (hpwt) has the highest F value, with horsepower (hp) and quarter mile time (qsec) next in line. From those three variables, we choose to add hpwt because 1) it is not correlated to weight like horsepower, so we avoid multicollinearity among regressors, and 2) it is design-time parameter and a cleaner definition of power than quarter mile time, which can only be measured after a car built and is a mash of power, drag, wind resistance, driver, etc.

Adding hpwt to the model and running `add1()` against the new model shows no more variables can improve the model. **Figure 6** shows a summary of the final fit. **Figure 7** shows diagnostic plots and a discussion of them. This seems a good, parsimonious model.

Now, to answer the question of whether transmission type affects fuel consumption. In **Figure 8**, we use `anova()` and find that transmission type would not be a significant factor in our model. So we conclude that transmission type does not impact fuel consumption.

Appendix

This section contains the figures referenced in the Analysis Details.

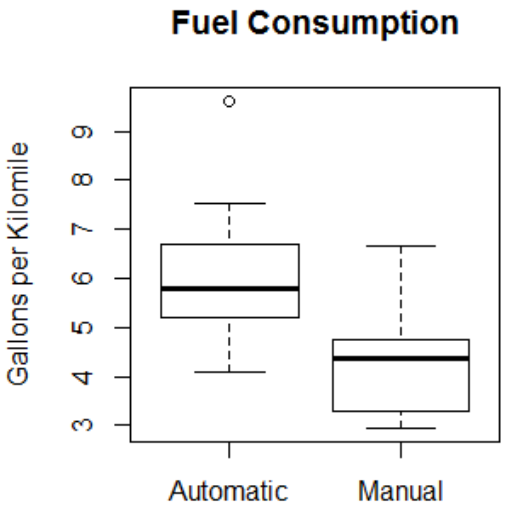

Figure 1a

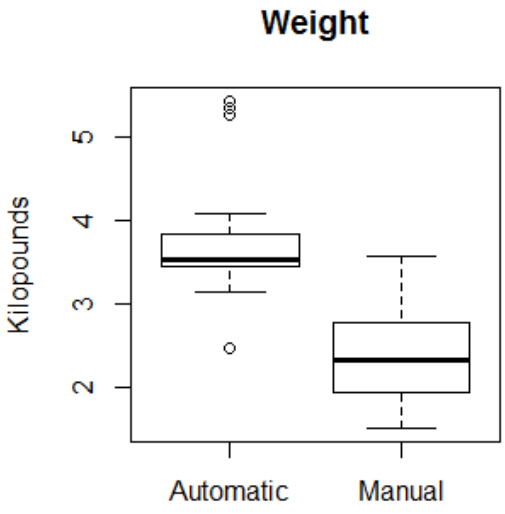

Figure 1b

Figures 1a and 1b: Cars with manual transmissions use less fuel (4.39 gpm) than automatics (5.78 gpm) and are lighter (2.32 Klbs) than automatics (3.52 Klbs). But are the differences significant?

```
fit.gpm.vs.trans <- lm(gpm ~ trans, data=D)

summary(fit.gpm.vs.trans)$coef;

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  6.144642  0.3224107 19.058429 2.588825e-18
## transManual -1.777029  0.5058396 -3.513028 1.426570e-03

round(confint(fit.gpm.vs.trans), 2)

##             2.5 % 97.5 %
## (Intercept)  5.49   6.80
## transManual -2.81  -0.74
```

Figure 2: An intercept-only regression fit of mpg ~ trans shows the difference between automatic and manual transmissions is statistically significant (for transManual, Pr(>|t|) = 0.00143 < 0.05). Automatics consume on average 6.14 gpm, with a confidence interval of 5.49 to 6.8 gpm. Manuals consume on average -1.78 less gpm than automatics, with a confidence interval of -2.81 to -0.74 gpm less. Adding the deltas for manuals with the 6.14 average for automatics gives, for manuals, an average of 4.37 gpm, with a confidence interval of 3.33 to 5.4 gpm.

---

	gpm	hpwt	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
gpm	1	0.29	0.81	0.88	0.76	-0.64	0.89	-0.39	-0.64	-0.54	-0.48	0.53
hpwt		1	0.45	0.32	0.77	-0.06	0.05	-0.8	-0.49	0.24	0.35	0.63
cyl			1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp				1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp					1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat						1	-0.71	0.09	0.44	0.71	0.7	-0.09
wt							1	-0.17	-0.55	-0.69	-0.58	0.43
qsec								1	0.74	-0.23	-0.21	-0.66
vs									1	0.17	0.21	-0.57
am										1	0.79	0.06
gear											1	0.27
carb												1

Figure 3: Fuel consumption (gpm), the first row, is correlated most to weight (wt) (0.89). Weight is also correlated to displacement (disp), number of cylinders (cyl), and horsepower (hp), all of which are also correlated with fuel consumption. Weight is not correlated to horsepower/weight (hpwt) (0.05).

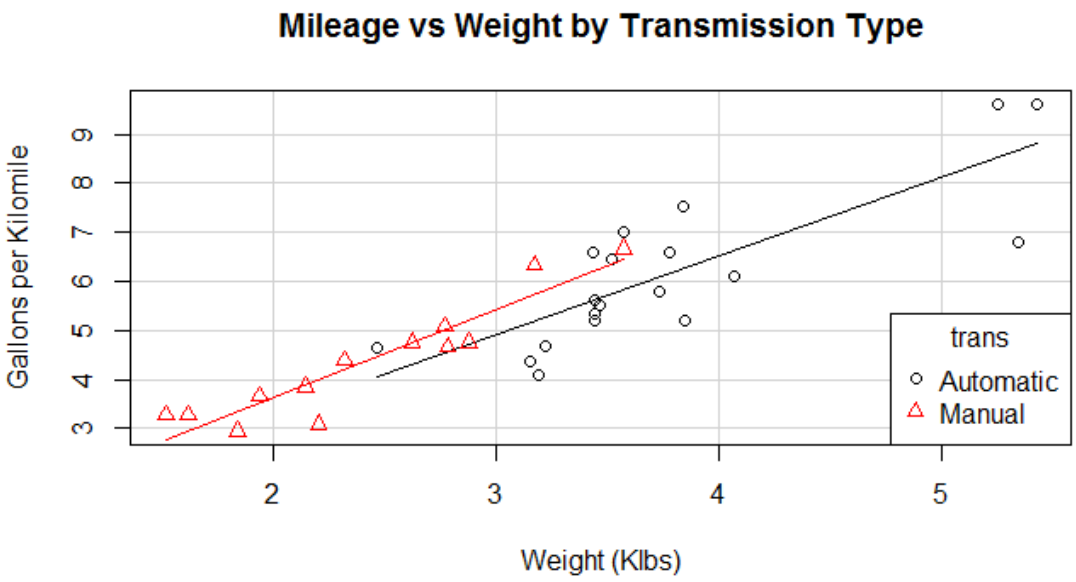

Figure 4: Here we plainly see transmission type is not spread evenly across weight. The cars with the lowest fuel consumption are the lightest and have manual transmissions, and the cars with the highest fuel consumption are the heaviest and have automatic transmissions. It also looks like once we account for weight, there is not that much difference in fuel consumption between manuals and automatics.

```
fit.gpm.vs.wt <- lm(gpm ~ wt, data=D);
add1(fit.gpm.vs.wt, D, test="F")
```

```
## Single term additions
##
## Model:
## gpm ~ wt
##        Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>              17.402 -15.493
## hpwt    1    4.7220 12.680 -23.624 10.7998 0.002659 **
## cyl     1    2.9735 14.428 -19.490  5.9766 0.020809 *
## disp    1    3.1754 14.226 -19.941  6.4731 0.016538 *
## hp      1    4.6144 12.787 -23.353 10.4649 0.003034 **
## drat    1    0.0028 17.399 -13.499  0.0047 0.946036
## qsec    1    4.5777 12.824 -23.262 10.3521 0.003173 **
## vs      1    2.5790 14.823 -18.627  5.0458 0.032471 *
## am      1    0.9370 16.465 -15.265  1.6504 0.209076
## gear    1    0.2011 17.201 -13.865  0.3390 0.564922
## carb    1    2.1760 15.226 -17.768  4.1445 0.050997 .
## trans   1    0.9370 16.465 -15.265  1.6504 0.209076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: With a base model of fuel consumption as a function of weight, use `add1()` to evaluate which variable to add next. Horsepower/weight (hpwt) has the highest F value and would have the most influence.

```
fit.gpm.vs.wt.hpwt <- lm(gpm ~ wt + hpwt, data=D)

summary(fit.gpm.vs.wt.hpwt)
```

```
##
## Call:
## lm(formula = gpm ~ wt + hpwt, data = D)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69714 -0.46822  0.05312  0.42744  1.35097
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.401534   0.512044  -0.784  0.43929
## wt           1.472176   0.121554  12.111 7.24e-13 ***
## hpwt         0.023997   0.007302   3.286  0.00266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6612 on 29 degrees of freedom
## Multiple R-squared:  0.8484, Adjusted R-squared:  0.8379
## F-statistic: 81.13 on 2 and 29 DF,  p-value: 1.322e-12
```

```
round(confint (fit.gpm.vs.wt.hpwt),  2)
```

```
##             2.5 % 97.5 %
## (Intercept) -1.45   0.65
## wt           1.22   1.72
## hpwt         0.01   0.04
```

```
round(sqrt(vif(fit.gpm.vs.wt.hpwt)), 4)
```

```
##     wt   hpwt
## 1.0015 1.0015
```

```
shapiro.test(rstandard(fit.gpm.vs.wt.hpwt))

##
##  Shapiro-Wilk normality test
##
## data:  rstandard(fit.gpm.vs.wt.hpwt)
## W = 0.97685, p-value = 0.7041
```

Figure 6: The summary of the final model. Both regressors are significant. The Adjusted R-squared value says the model accounts for 83.8% of the variance of the fuel consumption. The variance inflation factors (vif) are near 1 and indicate there is no multicollinearity among the regressors. The Shapiro-Wilk test, with a p-value > 0.05, says that the residuals of the fit are normally distributed. Note that the Intercept term is not significant, it's p-value is > 0.05. That means the reported Intercept value is not significantly different from 0, which makes sense for this regression.

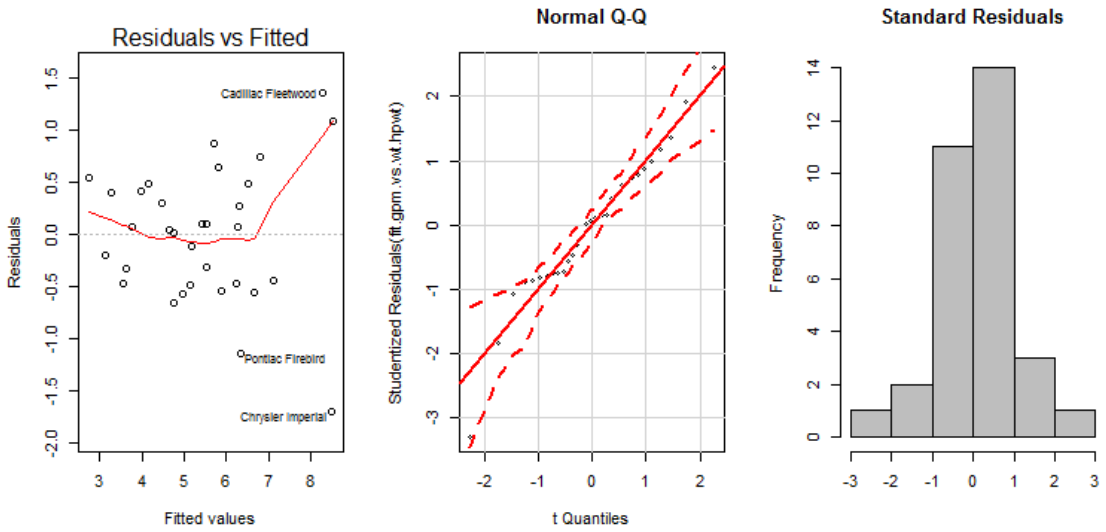

Figure 7: The diagnostic plots for gpm ~ wt + hpwt. In Residual vs Fitted, the values look reasonably randomly distributed about 0. In Normal Q-Q, the points lie reasonably along the diagonal. And the histogram of standardized residuals looks very much like a normal distribution.

```
## Analysis of Variance Table
##
## Model 1: gpm ~ wt + hpwt
## Model 2: gpm ~ wt + hpwt + trans
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     29 12.680
## 2     28 12.656  1  0.024087 0.0533 0.8191
```

Figure 8: Using `anova()`, we compare our base model (Model 1) to one that includes transmission type (Model 2). The Pr(>F) for Model 2 is greater than 0.05 which says transmission type would not make a significant difference to the model.

A note on formatting: To get to five pages, I knitr'd to a Word docx file, then in Word, changed margins, reduced overall font size, and inserted page breaks to prevent figures from spanning pages.