Transmission Type has no Impact on Car Gas Mileage

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 had better fuel economy, consuming on average 4.4 gallons per 100 miles (gpm) (+/- 1.7) 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 economy 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 not significant. The reason why cars with manual transmission had better mileage was because in this data set the transmission type and vehicle weight were confounders and were not independent. Given a car's weight you could reasonably predict its transmission type -- 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* (gpm), or kilomiles. It is a better metric for comparing the fuel consumption among cars (see ...). 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 metric that is 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, see **Figure 1**. There is a sizable difference in the median mileage, with manual transmissions consuming 4.39 gpm compared to 5.78 gpm for automatics.

Next, let's test whether the difference is significant by doing an intercept-only regression of fuel consumption as a function of transmission type, see **Figure 2**. We see:

* In the Coefficients, the p-value (Pr(>|t|) for manual transmissions (transManaul) is 0.00143 and less than 0.05. That means the difference between automatics and manuals is indeed significant.
* In the Coefficients, the Estimate column says that automatics (Intercept) consume on average 6.14 gpm, and manuals (transManual) consume -1.78 gpm less, for an average of 4.4 gpm.
* In the Confint, we see the 95th percentile interval for gpm, with automatics (Intercept) consuming 5.49 to 6.8 gpm and manuals (transManual) consuming -2.81 to -0.74 gpm *less*, for an interval of 2.68 to 6.06 gpm.

Let's investigate further looking at what variables affect fuel consumption the most. **Figure 3** shows the correlation between variables in the data set. We see:

* In the first row, fuel consumption (gpm) is correlated strongest to weight (wt) (0.89).
* Weight (wt) is correlated to transmission type (am) (-0.69). This is a hint that weight and tranmission may be confounders.
* Weight (wt) is strongly correlated to number of cylinders (cyl), displacement (disp), horsepower, but not horsepower/weight (hpwt).

Let's explore the relationship between fuel consumption and weight, see **Figure 4**. Here we plainly see that 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.

Let's build a multivariate regression starting with fuel consumption as a function of weight and see if transmission type is significant to the equation. We consider potential regressors in a forward stepwise fashion, using the function add1() to help decide which regressors to add. 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.

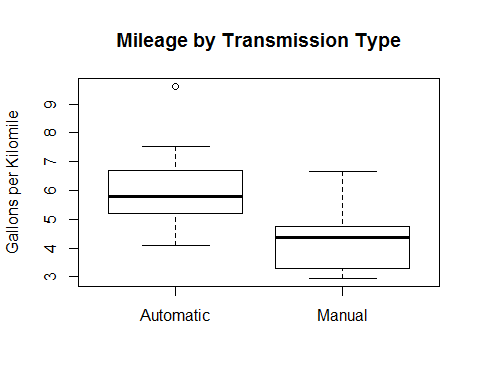
**Figure 5** show the output of add1() given the base model (gpm ~ wt). Horsepower/weight (hpwt) has the most influence, with horsepower (hp) and quarter mile time (qsec) next in line. From those three variables, we choose to add hpwt because 1) it is less correlated to weight than horsepower, so we avoid multiliearity among regressors, and 2) it is a cleaner definition of power than quarter mile time which is a mish mash of power, drag, wind resistance, etc.

We add hpwt to the model. Running add1() against the new model shows no more variables can improve the model. **Figure 6** shows the summary of the final fit.

### Appendix

This section contains the figures referenced in the Analysis Details.

boxplot(gpm ~ trans, data=D,   
 main="Mileage by Transmission Type", ylab="Gallons per Kilomile")



**Figure 1**: Cars with manual transmissions use less fuel than those with automatic transmissions. But is the difference significant?

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

summary(fit.gpm.vs.trans)$coef; round(confint(fit.gpm.vs.trans), 2)

## 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

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

**Figure 2**: A regression fit of mpg ~ trans shows the difference between automatic and manual transmissions is statistically significant (Pr(>|t|) = 0.00143 < 0.05) and that manuals consume on average -1.78 less gpm than automatics, with a range of -2.81 to -0.74 gpm less.

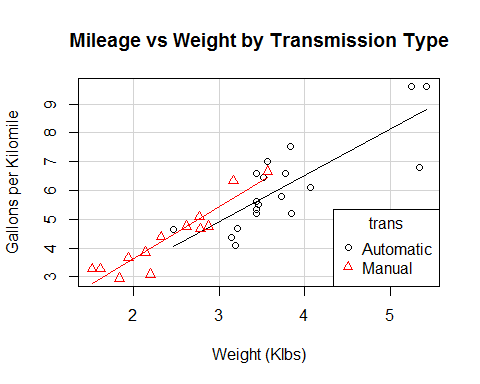
corr.matrix <- round(cor(D[,!names(D) %in% c('trans')]), 2)  
corr.matrix[lower.tri(corr.matrix,)] <- ""

kable(as.data.frame(corr.matrix), digits=2, align='r', format='markdown')

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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, the first row, is correlated most to weight (0.89) . It is also correlated to number of cylinders, displacment, and horsepower, all of which are highly correlated amongst themselves. Note that weight and horsepower/weight are not correlated ().

scatterplot(gpm ~ wt | trans, data=D,  
 main="Mileage vs Weight by Transmission Type",   
 ylab="Gallons per Kilomile", xlab="Weight (Klbs)",   
 labels=row.names(D), smoother=FALSE, legend.coords='bottomright')



**Figure 4**: Here we plainly see that 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.

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 with 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

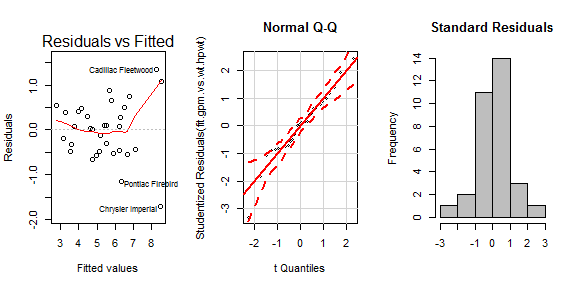
round(sqrt(vif(fit.gpm.vs.wt.hpwt)), 4)

## 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 multicolliniarity among the regressors. The Shapiro-Wilk test, with a p-value > 0.05, says that the residuals of the fit are normally distributed.



**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 like a normal distribution. I am satisfied with this fit.