Does transmission style affect your mpg?

Sid Reddy

January 21, 2015

Executive Summary

In this paper, we will examine the mtcars dataset, and determine if the transmission style impacts the miles per gallon (mpg) delivered by an automobile. In particular, we will address if automatic or manual transmission is better for mpg, and quantify the difference. Our analysis of the given data shows that the impact of transmission style on mpg is not statistically significant.

Building a model for mpg

We will begin by examining two models for mpg: one involving all of the variables, and another involving tranmission style alone.

```r
library(datasets)
fit1 <- lm(mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs)
           + factor(am) + factor(gear) + factor(carb), data = mtcars)
s1 <- summary(fit1)
c1 <- s1$coef['factor(am)1', ]

fit2 <- lm(mpg ~ factor(am), data = mtcars)
s2 <- summary(fit2)
c2 <- s2$coef['factor(am)1', ]

rbind(c1, c2)
```

```
##    Estimate Std. Error   t value     Pr(>|t|)
## c1 1.212116   3.213545 0.3771896 0.7113157275
## c2 7.244939   1.764422 4.1061270 0.0002850207
```

We notice that in either case, manual transmission delivers higher mpg (the coefficient for am is positive). The p-value (Pr(>|t|)) in c2 being very low (less than an alpha level of 0.05) indicates that manual transmission contributes to higher mpg. On the other hand, the high p-value in c1 suggests that when other variables (like wt, disp, hp etc.) are taken into consideration, the transmission style might not impact mpg after all.

We note that having more variables could result in overfitting. So, we will try to identify the minimal variables that impact mpg. We will choose weight, number of cylinders, horsepower, and transmission style to build a model. Why only these variables? Here are the reasons:

1. This choice of variables reflects how cars work.
2. This combination of variables maximizes the regression variance heuristically (I have tried several other combinations).
3. The ANOVA test below shows that the inclusion of these additional variables is justified (the p-value is very low, favoring fit3).

```
fit3 <- lm(mpg ~ wt + factor(cyl) + hp + factor(am), data = mtcars)
s3 <- summary(fit3)
a3 <- s3$adj.r.squared
c3 <- s3$coef['factor(am)1', ]
anova(fit2, fit3)$'Pr(>F)'[2] # p-value for choosing between fit2 and fit3
```

```
## [1] 1.688435e-08
```

I have also excluded transmission style (see fit4 below), and the loss in regression variance is negligible (see a4 vs a3 below). Also, given that the p-value (c3[‘Pr(>|t|)’] = 0.2064597) of transmission style is high in fit3, we infer that transmission stlye does not impact mpg in a statistically significant way.

```
fit4 <- lm(mpg ~ wt + factor(cyl) + hp, data = mtcars)
s4 <- summary(fit4)
a4 <- s4$adj.r.squared # Regression variance (dropping transmission style)
cbind(a3, a4)
```

```
##             a3        a4
## [1,] 0.8400875 0.8360668
```

Analysis of residuals, and VIF tests

We will now analyze the residuals to see if that provides more information (please see the residual plots below).

1. There are no systematic patterns in the residual vs fitted, indicating the model has not left out obvious variables. The smoothed curve is impacted by 3 outliers (Imperial, Corolla, Fiat 128), but that does not change our conclusion.
2. The errors also look to be normally distributed from the Q-Q plot below (the deviations from normal are minimal at the edges).
3. The scale-location plot shows that there are no issues like heteroskedasticity.
4. The residuals-leverage plot is fairly normal, with all points having a Cook's distance < 0.5.

We will also evaluate the variance inflation factor (VIF) to check if we are including more variables than required. From looking at the GVIF^(1/(2*Df)) values and given that they are all < sqrt(5), we conclude that the variables are all good to include. Note that a VIF value > 5 typically represents a redundant variable.

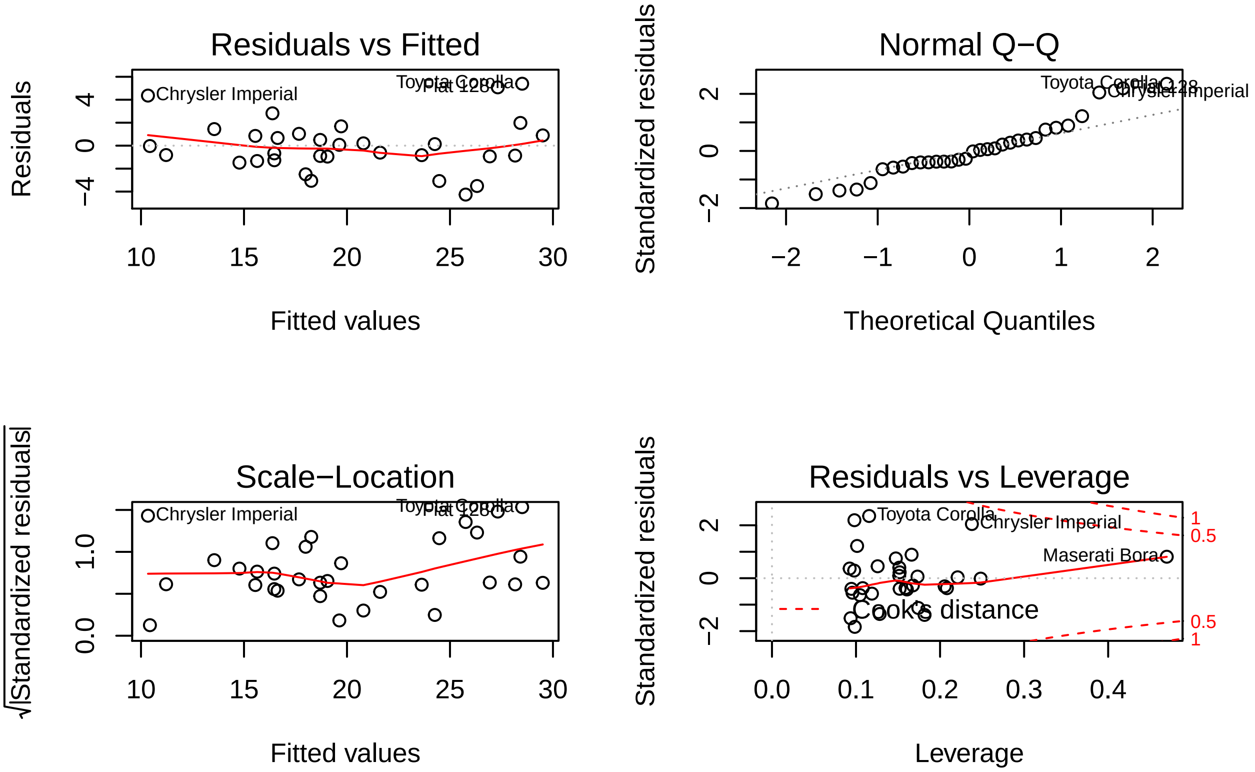
```
library(car)
vif(fit4)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## wt          2.580877  1        1.606511
## factor(cyl) 5.105811  2        1.503198
## hp          3.496014  1        1.869763
```

Conclusion

Given the above, the model we proposed (fit4) explains mpg pretty well, and the residual diagnostics and VIF tests do not indicate any serious flaws. Thus, our conclusion is that transmission style does not impact mpg in a statistically significant manner.

```r
par(mfrow = c(2, 2))
plot(fit4)
```



Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Maserati Bora
Cook's distance
1
0.5
0.5
1