Exploring the relationship between mpg and transmission in mtcars dataset

Anthony Liu

1/25/2020

Executive summary

The fuel economy of an automobile refers to how much fuel the engine consumes in travelling a certain distance. One common fuel economy metric is miles per gallon (MPG), which refers to the distance an automobile can travel using one gallon of fuel. The MPG rating is largely dictated by the physical specifications, design and build of the vehicle. In this report we use a dataset of 32 different car builds to attempt a statistical analysis on these two questions:

1. Is an automatic or manual transmission better for MPG?
2. What is the quantitative difference in MPG between automatic and manual transmissions?

We build a regression model for an inferential analysis regarding which factors affect MPG before turning to answer our questions. All analyses in this report were performed using R 3.6.1. The result was that we were unable to find any significant relationship between transmission and MPG.

Exploratory analysis (figures in Appendix A.1)

Our plot of the data suggests that the manual vehicles (`am = 1`) in our dataset have on average a higher MPG than automatic vehicles (`am = 0`). But we cannot yet draw any conclusions as this result may be confounded by other factors. That is, this is just a correlational relationship and it could be the case that the difference in MPG comes from other factors and vehicles with these factors just tend to have manual transmission. Creating a scatterplot matrix of the variables in the dataset indeed confirms there could potentially be some strong confounding. We take note that many of the other factors show strong correlation with each other and we will need to be careful later on when building our regression models as there may be high degrees of collinearity which will introduce uncertainty in our interpretations.

Model selection

Here we build many different regression models and find the one with the best fit. The strategy here is to first fit 10 different models in a nested fashion. That is, we start with a simple linear regression where `mpg` is the outcome and we only have a single predictor `am` (corresponding to transmission). We progressively add one extra predictor for each new model until all 10 predictors are used up. We then use analysis of variance (ANOVA) to see which predictors are significant (with level 0.05) in our model. We observe that the addition of `cyl`, `hp` and `wt` are all significant but we also cannot discount `disp` considering it has a high correlation with the former three variables. That is, there is a chance of collinearity between `cyl`, `hp`, `wt` and `disp` such that the p-values of `cyl`, `hp` and/or `wt` are inflated while the significance of `disp` is masked. When taking into consideration basic principles of automobile hardware and fuel consumption, it is intuitively obvious to us that the number of cylinders an engine has and the weight of the vehicle would have an impact on fuel economy. Thus, we choose to include both `cyl` and `wt` in our model regardless and experiment with whether to include the `hp` and `disp` variables which are less intuitive.

Further model fitting and analysis of variance shows that at level 0.05, neither adding `hp` nor `disp` is a significant improvement to fitting `cyl` and `wt` alone. The inclusion of `hp` does have a relatively low p-value of around 0.08 so we choose to err on the conservative side and include `hp` in our next model creation and testing phase which includes interaction terms. We find that the model with predictors `am`, `cyl`, `wt`, `hp` and interaction term `wt:hp` contributes the most significant improvement. Tests using square root and logarithmic terms showed no signficant improvement (not included in this document due to length constraints) so we settle with this as our final model. That is, our final model is $mpg = \beta_0 + \beta_1(am) + \beta_2(cyl) + \beta_3(wt) + \beta_4(hp) + \beta_5(wt \times hp) + \epsilon$.

```
fit17 <- lm(mpg ~ factor(am) + cyl + wt * hp, data = mtcars)
summary(fit17)$r.squared
```

```
## [1] 0.8869349
```

```
summary(fit17)$coef
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 49.68816328 5.341886192  9.3016140 9.363757e-10
## factor(am)1 -0.07095465 1.374980885 -0.0516041 9.592386e-01
## cyl         -0.37080011 0.529367367 -0.7004589 4.898620e-01
## wt          -7.68434678 1.901661507 -4.0408594 4.202228e-04
## hp          -0.10866912 0.030804795 -3.5276690 1.580972e-03
## wt:hp        0.02601008 0.008810073  2.9523110 6.605776e-03
```

Residual plots and diagnostics (figures in Appendix A.2)

Here we first create a plot comparing our fitted values (x-axis) from our regression model against the actual values (y-axis). For comparison, a red line representing perfect fit is drawn. It can be seen that our model fits the data quite well. We also create a barplot to compare the variance inflation factors of each predictor in the model. The variable `am` shows a variance inflation factor of only 2.98 (5 or above is typically problematic), so we can reasonably assume that there is little collinearity against the other variables affecting its reported significance in the model. The other variables do seem to show problematic levels of variance inflation, but much of this is actually due to structural collinearity from adding an interaction term. Centering the variables around their means and refitting the model is able to account for this and demonstrates acceptable levels of variance inflation (not shown in this report due to length constraints). In either case, variance inflation among these variables are of little concern for the questions in this report which involve only the `am` variable. We next plot some residual and diagnostic plots. The residual plots do not display any obvious pattern and so we can conclude that the bias in our model is minimal. The Q-Q plot suggests some non-normality which does not seem to be significant. There also doesn't seem to be any significant outliers or high-leverage points. In summary, we have quite a good fit.

Answering our questions of interest

```
confint(fit17)
```

```
##                     2.5 %      97.5 %
## (Intercept)  38.707758956 60.66856761
## factor(am)1  -2.897268335  2.75535904
## cyl          -1.458930317  0.71733010
## wt          -11.593267994 -3.77542557
## hp           -0.171989283 -0.04534896
## wt:hp         0.007900712  0.04411944
```

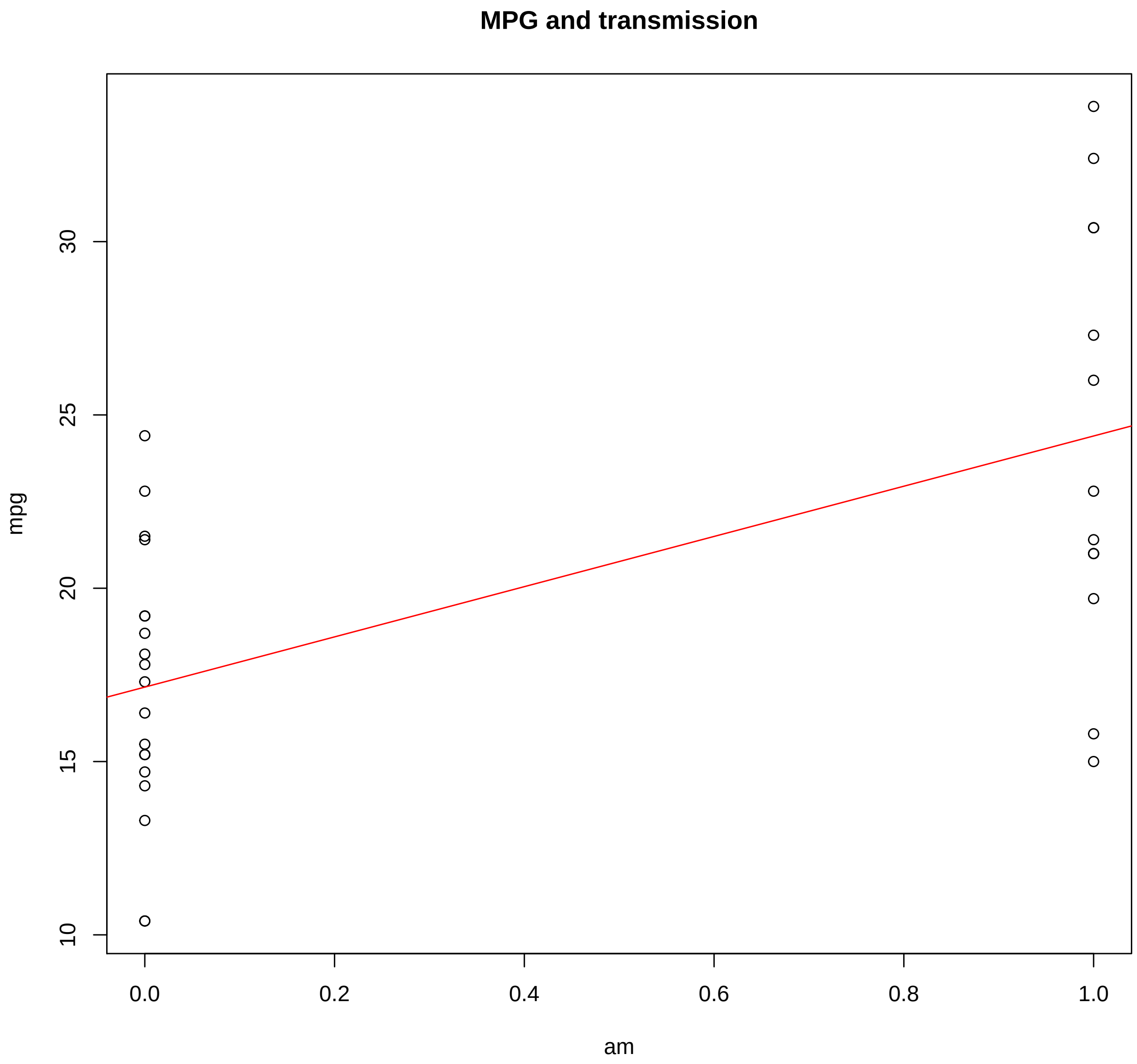
Let us return now to our questions proposed in the Executive summary. Our answers are:

1. We fail to reject the null hypothesis that there is a relationship between transmission and MPG (p-value of 0.96). That is, there is little evidence suggesting a relationship and we are unable to conclude whether automatic or manual transmission is better for MPG.
2. Holding all other variables constant, the MPG difference between automatic and manual transmissions (manual minus automatic) is between -2.90 and 2.76 with 95% confidence. That is, we expect to see in 95% of vehicles that ones with manual transmission will have between 2.90 less and 2.76 more MPG than ones with automatic transmission (keeping all other variables constant).

Appendix A

A.1 - Exploratory analysis

```r
data(mtcars)
plot(mtcars$am, mtcars$mpg, xlab = 'am', ylab = 'mpg', main = 'MPG and transmission')
abline(lm(mpg ~ factor(am), data = mtcars), col = 'red')
```

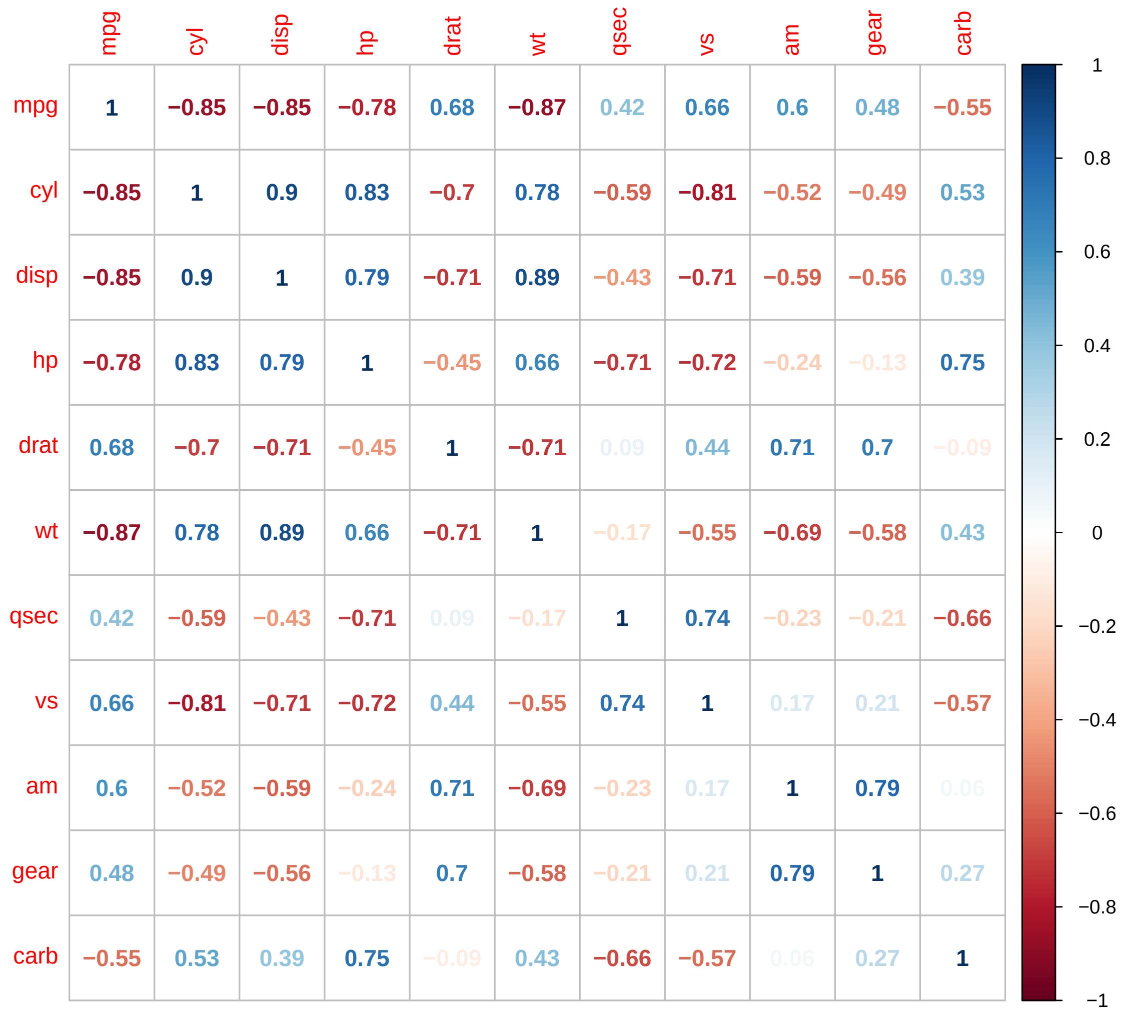


```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
```

```
## corrplot 0.84 loaded
```

```r
corrplot(cor(mtcars), method = "number")
```



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	−0.85	−0.85	−0.78	0.68	−0.87	0.42	0.66	0.6	0.48	−0.55
cyl	−0.85	1	0.9	0.83	−0.7	0.78	−0.59	−0.81	−0.52	−0.49	0.53
disp	−0.85	0.9	1	0.79	−0.71	0.89	−0.43	−0.71	−0.59	−0.56	0.39
hp	−0.78	0.83	0.79	1	−0.45	0.66	−0.71	−0.72	−0.24	−0.13	0.75
drat	0.68	−0.7	−0.71	−0.45	1	−0.71	0.09	0.44	0.71	0.7	−0.09
wt	−0.87	0.78	0.89	0.66	−0.71	1	−0.17	−0.55	−0.69	−0.58	0.43
qsec	0.42	−0.59	−0.43	−0.71	0.09	−0.17	1	0.74	−0.23	−0.21	−0.66
vs	0.66	−0.81	−0.71	−0.72	0.44	−0.55	0.74	1	0.17	0.21	−0.57
am	0.6	−0.52	−0.59	−0.24	0.71	−0.69	−0.23	0.17	1	0.79	0.06
gear	0.48	−0.49	−0.56	−0.13	0.7	−0.58	−0.21	0.21	0.79	1	0.27
carb	−0.55	0.53	0.39	0.75	−0.09	0.43	−0.66	−0.57	0.06	0.27	1

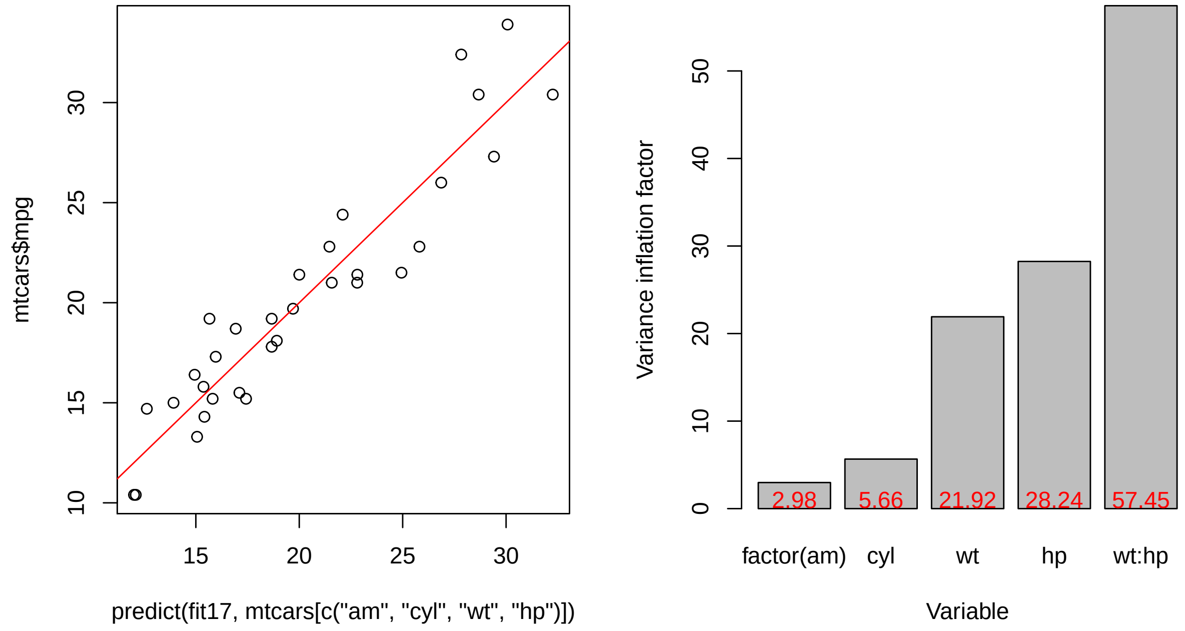
A.2 - Residual plots and diagnostics

```r
library(car)
```

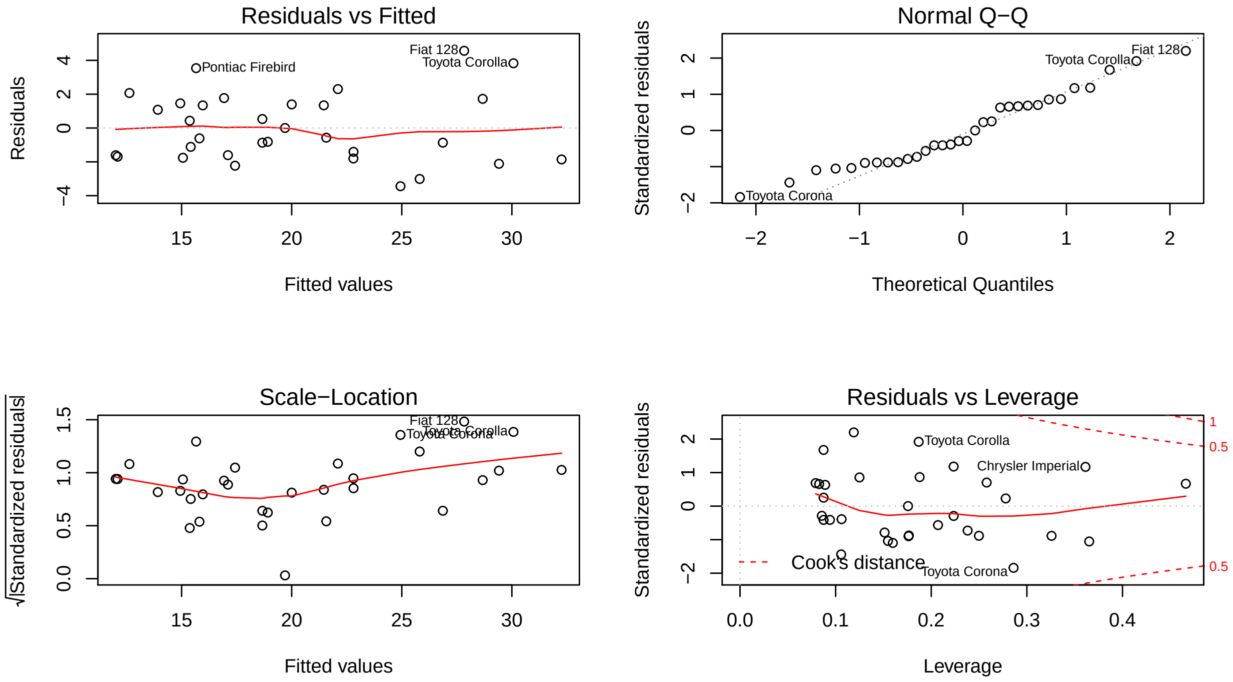
```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```r
par(mfrow = c(1, 2))
plot(predict(fit17, mtcars[c('am', 'cyl', 'wt', 'hp')]), mtcars$mpg)
abline(a = 0, b = 1, col = 'red')
VIF <- vif(fit17)
barplot(VIF, xlab = 'Variable', ylab = 'Variance inflation factor')
for(i in 1:5) text(0.7 + 1.2 * (i - 1), 1, round(VIF[i], 2), col = 'red')
```



```r
par(mfrow = c(2, 2))
plot(fit17)
```


Residuals vs Fitted
Residuals
−4
0
2
4
Pontiac Firebird
Fiat 128
Toyota Corolla
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
−2
0
1
2
Toyota Corolla
Fiat 128
Toyota Corona
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
Fiat 128
Toyota Corona
Toyota Corolla
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
−2
0
1
2
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
0.0
0.1
0.2
0.3
0.4
Leverage