Analysis of the Mileage performance and Transmission relationship: insights from the 1974 Motor Trend data

Ozan Aygun

3/7/2017

Executive Summary

Here we present our conclusions to answer the question whether the manual or automatic transmission designs has any significant relationship with the mileage performance. By using 1974 Motor Trend Data, with 95% confidence we concluded that manual cars give a mean of 2.92 miles ($p < 0.05$) more mileage compared to automatic cars, holding weight, horse power and quarter mile time constant.

Exploratory data analysis: Developing expectations from the data

Initial examination of the data gives the impression that manual cars provide better mileage compared to automatic transmission cars (**Figure 1, Appendix**). However, we have measured 9 other aspects of the automobiles, which challenges this initial observation. Most importantly, we notice that most of the manual transmission cars in our data set also have lower number of cylinders and lower weight compared to automatic cars (**Figure 2, Appendix**). Both of these variables are directly related to fuel consumption. Therefore, the engine size (or car's overall power) as well as the weight can confound the relationship between the transmission design and mileage. We need to build a model that adjusts for these major confounders.

In order to include the right confounders into our model, we needed to account for the collinear variables in our data. There are several covariates in our data set that can be collectively considered as a measure of overall automobile engine power.We expect that the cars that have higher number of cylinders would have more horse power, have higher `engine displacement`. All these covariates are highly correlated with each other and anti-correlated with fuel consumption (**Figure 2, Appendix**). Therefore, we expect that only one of these covariates will be useful to include in our model.Weight is a major contributor for fuel consumption and is a clear confounder in our question.Although we anticipate that more powerful cars have bigger engine and therefore are heavier, this alone can not explain the overall weight of the car. We will include weight in our model by all means.

Model building

We attempted to explain the relationship between fuel mileage and transmission category by fitting our outcome variable mpg into a multiple linear regression model with our primary predictor variable am, as well as the major confounder wt (weight). Therefore, the base model will be: **mpg = B0 + B1 * am + B2 * wt**

In order to decide between cyl,disp and hp variables, we used a nested model selection:

```
fit1 <- lm(mpg ~ factor(am) + wt, data = mtcars)
fit2 <- lm(mpg ~ factor(am) + wt + hp, data = mtcars)
fit3 <- lm(mpg ~ factor(am) + wt + hp +cyl, data = mtcars)
fit4 <- lm(mpg ~ factor(am) + wt + hp +cyl +disp, data = mtcars)
fit.test <-anova(fit1,fit2,fit3,fit4)[,6] # Gets the p-values from Anova()
names(fit.test) <- c("base","+hp","+hp+cyl","+hp+cyl+disp");round(fit.test,5)
```

```
##         base          +hp      +hp+cyl +hp+cyl+disp
##           NA      0.00053      0.21154      0.30472
```

The p-values suggest that only addition of hp significantly advances the base model. Therefore, we included hp as the only measure of engine power and possible confounder. The two variables gear and drat are highly correlated, but somewhat orthogonal to other variables we measured. Therefore, we tested their impact on the model that includes horsepower.

```
fit5 <- lm(mpg ~ factor(am) + wt + hp + gear, data = mtcars)
fit6 <- lm(mpg ~ factor(am) + wt + hp + gear + drat, data = mtcars)
fit.test2 <- anova(fit2,fit5,fit6)[,6] # Gets the p-values from Anova()
names(fit.test2) <- c("base+hp","base+hp+gear","base+hp+gear+drat");round(fit.test2,5)
```

```
##          base+hp      base+hp+gear base+hp+gear+drat
##               NA           0.71154           0.55064
```

The p-values suggest that neither of these variables add significantly to the model we established. We also suspected that qsec and carb variables might have some independent impact on mileage contribution, in addition to their relationship with the engine power. Therefore, we tested their potential contributions on the model we carried over.

```
fit7 <- lm(mpg ~ factor(am) + wt + hp + qsec, data = mtcars)
fit8 <- lm(mpg ~ factor(am) + wt + hp + qsec + carb, data = mtcars)
fit.test3 <- anova(fit2,fit7,fit8)[,6] # Gets the p-values from Anova()
names(fit.test3) <- c("base+hp","base+hp+qsec","base+hp+qsec+carb");round(fit.test3,5)
```

```
##           base+hp      base+hp+qsec base+hp+qsec+carb
##                NA           0.07858           0.47106
```

Neither of these covariates seem to be necessary for our model.According to the `original paper` the vs variable encodes the V or straight (S) engine types. We test the impact of engine type in our model as a binary variable:

```
fit9 <- lm(mpg ~ factor(am) + wt + hp + factor(vs), data = mtcars)
fit.test4 <- anova(fit2,fit9)[,6] # Gets the p-values from Anova()
names(fit.test4) <- c("base+hp","base+hp+vs");round(fit.test4,5)
```

```
##    base+hp base+hp+vs
##         NA    0.18969
```

The engine type does not seem to be adding significant information to our model either. Therefore, our finalized model includes weight and horsepower in addition to the transmission type.

Challenging the model: AIC and residual diagnostics

AIC is another useful approach for model selection.When comparing a series of meaningful models, the model that provides the lowest AIC is considered as the best fit. In order to challenge our existing model we also extracted AIC information from each of the models we tested:

```
model.list <- list(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9)
fitAIC.summary <-sapply(model.list, extractAIC)[2,]; names(fitAIC.summary)<-c("fit1","fit2","fit3","fit4","fit5","
```

```
##    fit1    fit2    fit3    fit4    fit5    fit6    fit7    fit8    fit9
## 75.2171 63.3228 63.4416 64.1200 65.1536 66.7067 61.5153 62.8635 63.2463
```

Therefore, the model that contains qsec (quarter mile time) has lower AIC compared to our earlier model, suggesting that this variable, although does not seem to be significantly adding to the model, still explains some of the variance. Therefore, we will keep it in the final model, which now includes weight,horsepower and quarter mile time in addition to the transmission type.

The residual diagnostics of the finalized model suggest that we don't have gross departures from multiple linear regression assumptions. Residuals seem to have near equal variance across the fitted values(homoscedastic) and reasonably normally distributed (**Figure 3,Appendix**).

Conclusions and Inference

The summary of the final model, estimated coefficients, p values and confidence intervals are:

```
round(summary(fit7)$coefficients,5); confint(fit7)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 17.44019    9.31887  1.87149  0.07215
## factor(am)1  2.92550    1.39715  2.09391  0.04579
## wt          -3.23810    0.88990 -3.63873  0.00114
## hp          -0.01765    0.01415 -1.24705  0.22309
## qsec         0.81060    0.43887  1.84702  0.07573
```

```
##                    2.5 %     97.5 %
## (Intercept) -1.68054821 36.5609304
## factor(am)1  0.05879488  5.7922130
## wt          -5.06401789 -1.4121758
## hp          -0.04668118  0.0113881
## qsec        -0.08988486  1.7110899
```

With 95% confidence, we therefore concluded that manual cars give a mean of 2.92 miles more mileage compared to automatic cars, holding weight, horse power and quarter mile time constant. We also concluded that this difference is statistically significant ($p < 0.05$).

Appendix

Figure 1: Relationship between mpg with other covariates

```
library(GGally); library(ggplot2)
mtcars$trans <- ifelse(mtcars$am == 0,"automatic","manual")
mtcars$trans <- factor(mtcars$trans)
ggduo(mtcars,columnsX = 2:11,columnsY = 1, mapping = aes(col=trans), showStrips = FALSE, legend =1)+theme_bw()
```

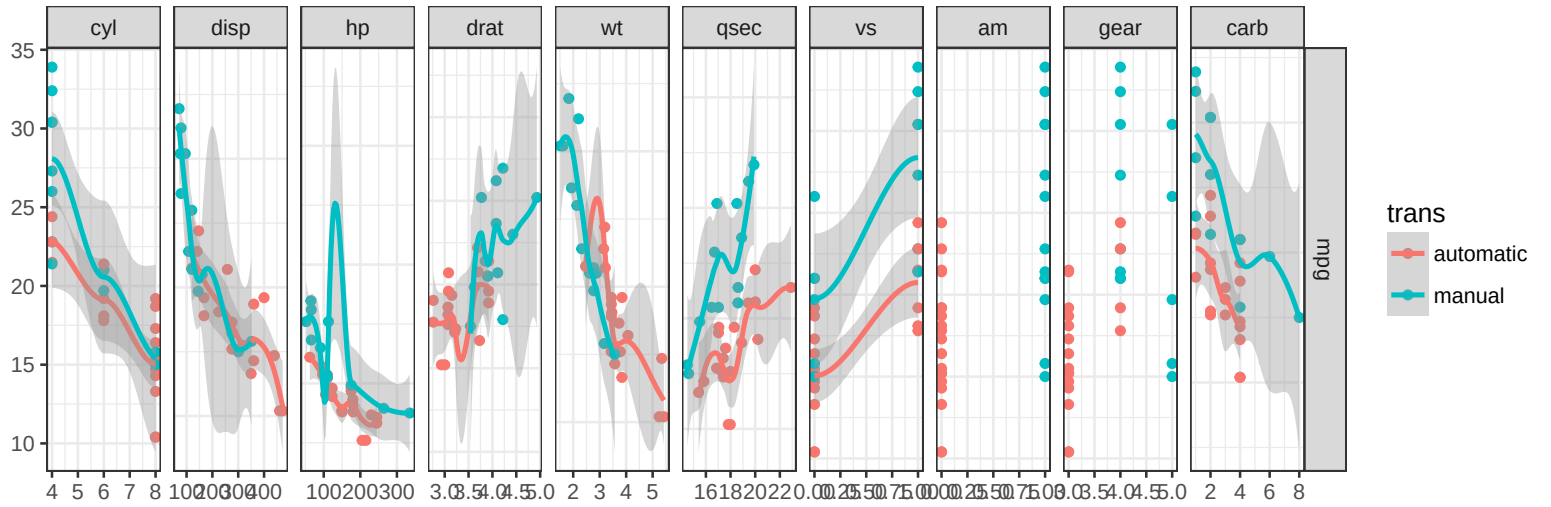

Figure 2: Pairs plot and correlation coefficients between mtcars variables

```
ggpairs(mtcars,c(1:7,10:11),mapping = ggplot2::aes(color = "navy"))+theme_bw()
```

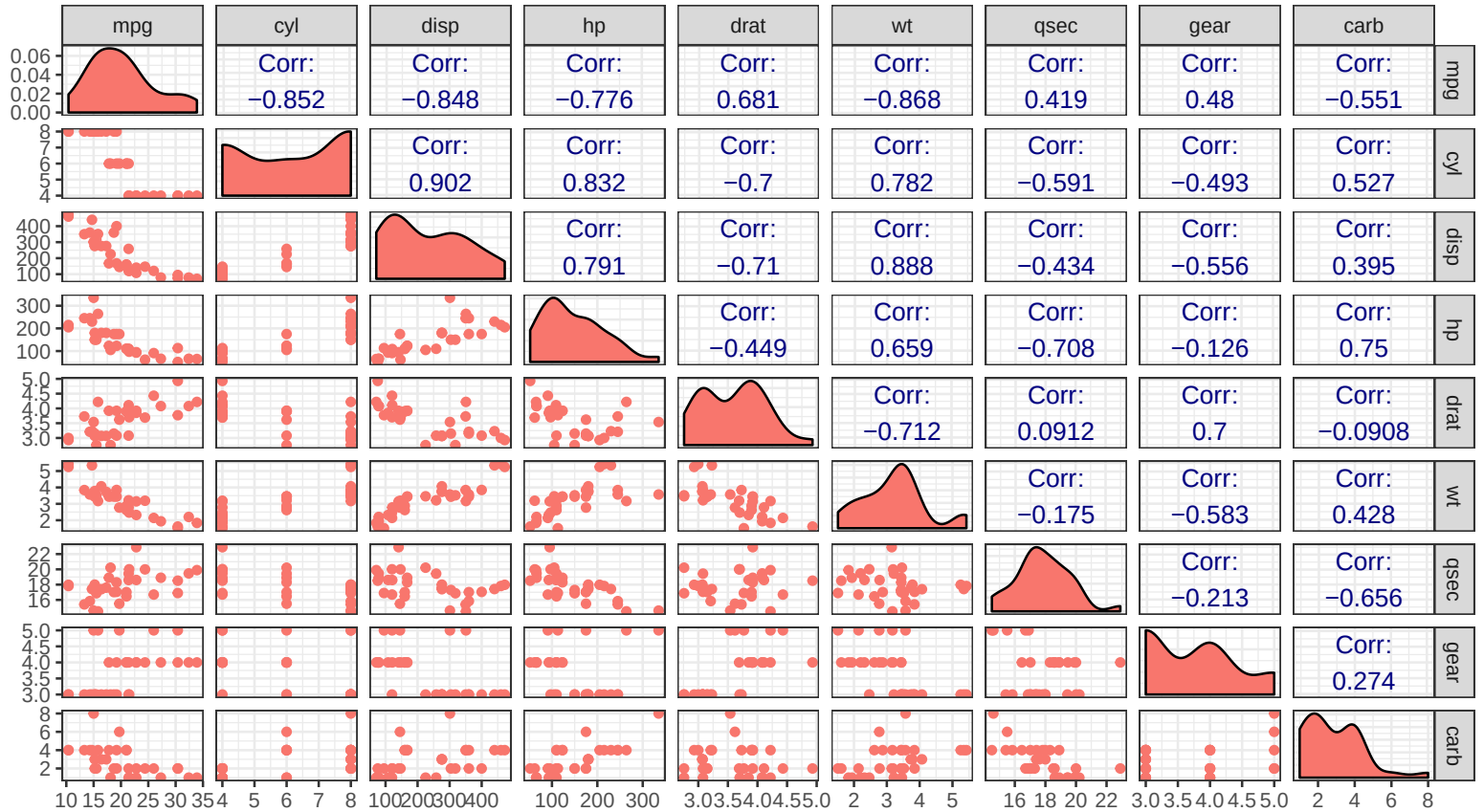

Figure 3: residual diagnostics of the final model

```
par(mfrow=c(2,2))
plot(fit7,cex = 0.4,pch =19, col = "navy")
```