Analysis of Automobile Fuel Economy using Different Transmissions

Elmar Langholz

May 21, 2015

Summary

Through this document we set out to try to determine whether using a manual or automatic transmission was better for fuel economy. Using the Motor Trend US magazine data set for 1973 - 1974 automobile models, we were able to determine, with a 95% confidence, that if we randomlly pick a car a manual transmission will perform on average better by reducing the consumption of gasoline by 7.24 gallons per mile. There are other confounding variables, besides transmission, which impact the fuel economy. Weight is one of these variables for which we were able to determine that with constant weight the fuel economy reduces on average 5.2984 faster from an automatic transmission to a manual transmission.

Exploratory data analysis

After the data is loaded and we are able to understand its structure, we proceed to process it by cleaning and validating to make sure that it contains appropriate values. Once processed and summarized, we are able to determine that the transmission type variable is not equally distributed across the observations which in turn means that *there could be a bias introduced just by the fact that there are more measurements from the automatic transmission type* (especially with such a low amount of observations). A proper follow up would be to understand how the observations where chosen and if they were indeed random. Due to time constraint, we will assume this has no reasonable effect on the data set and that it is truly random. Keeping that aside, if we perform a visual analysis through a scatterplot matrix, we are able to roughly determine that *the number of cylinders (`cyl`), displacement (`disp`), horse power (`hp`) and weight (`wt`) have a high negative correlation with the fuel economy (`mpg`)*. To confirm this, we look at the actual correlations and we ascertain the afore statement.

Impact of transmission on fuel economy

If we take a look at *Figure 1*, the fuel economy when compared with the transmission types, we notice that the fuel economy is better on average for manual transmission than that of an auatomated transmission. Performing statistical inference confirms (p-value: $0.00137 < 0.05$) this by accepting to reject the null hypothesis. With a 95% confidence interval the difference in mean (between automatic and manual) is of (-11.2801944, -3.2096842).

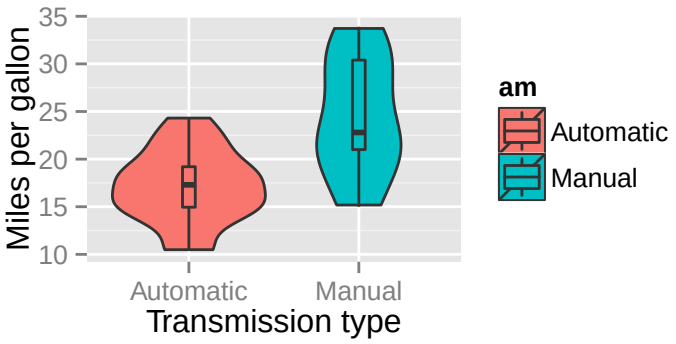

Figure 1: Average impact of fuel economy by transmission type

Quantification of fuel economy difference in transmissions

With the intent of quantifying the fuel ecnonomy we make use of single and multivariate regression models. We know, through the scatterplot matrix, that fuel economy is not only explained by transmission type but by other variables as well. Since the data set contains 11 columns and we are aware that we will fix the response `mpg` and include at least the `am` predictor, there are approximately $2^9 - 1 = 511$ different models that we would need to try in order to select the best one. However, we

decided to take a look at four models which would likely yield "parsimonious, interpretable representations of the data that enhance our understanding": 1. Model which includes only the baseline response and predictor 2. Model which includes all the variables 3. Model automatically generated by the stepwise regression 4. Model which includes the baseline and the variable with the maximum correlation

Baseline regression model

After creating the baseline regression model, we notice that manual transmission cars consume 7.2449 gallons per mile less fuel than an automatic transmission. However, this model only accounts for 33.85% of the variance.

All variables regression model

Comparing the regression model which contains all the variables with the baseline, through the p-value for the difference in the F statistic (1.77896752002729e-05), we confirm that including all variables is better model than just including the transmission. However, we question the outcome since we are including variables that have a low correlation to the response which in turn increases standard errors of the regression variables as we can see with all the coefficient p-values which are larger than 0.05. Also, R^2 increases monotonically as more regressors are included which explain why we the model accounts for an 80.66% variance.

Best stepwise AIC regression model

With the intent of automatically selecting the best model by iterating through different variables we make use of stepwise regression using AIC. The best regression model retrieved includes the car weight (`wt`), quarter mile time (`qsec`) and the transmission (`am`) as predictors. While it accounts for 83.36% of the variance, we see that the p-value for the intercept is not statistically significant (pvalue: $0.1779152 > 0.05$). As a point in hand, the quarter mile time makes it difficult to interpret.

Baseline with maximum correlation variable regression model

Looking at the previously defined models and the correlation analysis, we notice that weight (`wt`) is not only present in the stepwise regression but is also the variable with the maximum correlation with a value of -0.86766. Assesing the regression model with the baseline and including weight yields a model for which the p-value of the transmission coefficient is not statistically significant (pvalue: $0.001621 > 0.05$). Nonetheless, if we analyze at the regression model which includes interaction between the weight and the transmission we realize that this model encompasses both a valid set of p-values for the coefficients ($\forall p_i < 0.05$) and that the model accounts for 81.5149% of the variance.

Conclusion

Regarding the impact of the tranmisson on the fuel economy, we can say that a randomlly picked car with a manual tranmission will have a better fuel economy than one with an automatic transmission. Specifically, we notice that manual transmission cars consume 7.2449 gallons per mile less fuel than an automatic transmission on average.

As far as the quantification aspect of the selected regression model (`lm(mpg ~ am * wt)`) which accounts for the weight confounding effect, we can say that constant weight (with a p-value of 0.001) the fuel economy reduces on average 5.2983605 faster from an automatic transmission to a manual transmission (which can be depicted by an effects plot. There is room for improvement in this model, likely including other correlated variables. Sepecifically, after performing diagnostics, through the *Residuals vs Fitted* plot we are able to determine heteroscedasticity. In the *Normal Q-Q* plot, when looking at the end of the line, we notice that the current model is not fully linear yet. As a point in hand, the press residuals identify that the *Fiat 128* is the car that deviates the most from others leading us think that further investigation on it should be performed.

Appendix

1. Loading data

The mtcars data set is provided by the datasets package.

```
library(datasets); data(mtcars); mtcarsOrig <- mtcars
```

2. Structure of data

```
str(mtcarsOrig)
```

3. Clean data

```
mtcars$am <- factor(mtcars$am); levels(mtcars$am) <- c("Automatic", "Manual")
mtcars$vs <- factor(mtcars$vs); levels(mtcars$vs) <- c("Straight", "V")
```

4. Validate data

```
sum(data.frame(
    mpg = sum(mtcars$mpg <= 0), disp = sum(mtcars$disp <= 0), hp = sum(mtcars$hp <= 0),
    drat = sum(mtcars$drat <= 0), wt = sum(mtcars$wt <= 0, qsec = sum(mtcars$qsec <= 0)),
    cyl = sum(mtcars$cyl <= 0), gear = sum(mtcars$gear <= 0), carb = sum(mtcars$carb <= 0)))
```

```
## [1] 0
```

5. Summary of data

Variables	Description	Values Summary
mpg	Miles/(US) gallon	Real-valued: $\mu = 20.09$, $\sigma = 6.03$, [10.4, 33.9]
cyl	Number of cylinders	Count: 4, 6, 8
disp	Displacement (cu.in.)	Real-valued: $\mu = 230.72$, $\sigma = 123.94$, [71.1, 472]
hp	Gross horsepower	Real-valued: $\mu = 146.69$, $\sigma = 68.56$, [52, 335]
drat	Rear axle ratio	Real-valued: $\mu = 3.6$, $\sigma = 0.53$, [2.76, 4.93]
wt	Weight (lb/1000)	Real-valued: $\mu = 3.22$, $\sigma = 0.98$, [1.513, 5.424]
qsec	Quarter (1/4) mile time	Real-valued: $\mu = 17.85$, 1.79, [14.5, 22.9]
vs	V/S (engine type)	Categorical: Straight, V
am	Transmission	Categorical: Automatic, Manual
gear	Number of forward gears	Count: 3, 4, 5
carb	Number of carburetors	Count: 1, 2, 3, 4, 6, 8

Table 1: Summary of data

6. Frequency of the transmission variable

Table 2: Frequency of the values of am

am	Freq
Automatic	19

am	Freq
Manual	13

7. Scatter plot matrix for all the cars data set

```
scatterplotMatrix(mtcars)
```

8. Correlation analysis of all variables with respect to fuel economy

```
varCor <- cor(mtcars$mpg, mtcars[, !names(mtcars) %in% c("mpg", "am", "vs")])
```

9. Student's t-Test of fuel economy by transmission type

```
testMpgByAm <- t.test(mtcars$mpg ~ mtcars$am)
```

10. Covariate model selection I: Baseline

```
fitMpgWithAm <- lm(mpg ~ am, data = mtcars)
```

11. Covariate model selection II: All variables

```
fitAll <- lm(mpg ~ ., data = mtcars)
```

12. Model comparison: Baseline vs. All

```
devianceBaselineVsAll <- anova(fitMpgWithAm, fitAll)
```

13. Covariate model selection III: Stepwise using AIC

```
k <- qchisq(0.05, 1, lower.tail = FALSE) # adjust single variable at a time p-value < 0.05 (5%)
fitAIC <- step(fitAll, direction = "both", trace = 0, k = k, steps = 10000)
```

14. Covariate model selection IV: Baseline with maximum correlation

```
fitMpgWithAmPlusWt <- lm(mpg ~ am + wt, data = mtcars)
```

15. Covariate model selection IV: Baseline with maximum correlation with interaction

```
fitMpgWithAmTimesWt <- lm(mpg ~ am * wt, data = mtcars)
```

16. PRESS residual analysis

```
press <- resid(fitMpgWithAmTimesWt) / (1 - hatvalues(fitMpgWithAmTimesWt)); maxDev <- which.max(press)
```

17. Effect plot of transmission with weight on fuel economy

```
plot(effect(term = "am:wt", mod = fitMpgWithAmTimesWt, default.levels=2), multiline=TRUE)
```

18. Diagnostics for lm(mpg ~ am * wt)

```
par(mfrow = c(2, 2))
plot(fitMpgWithAmTimesWt)
par(mfrow = c(1, 1))
```

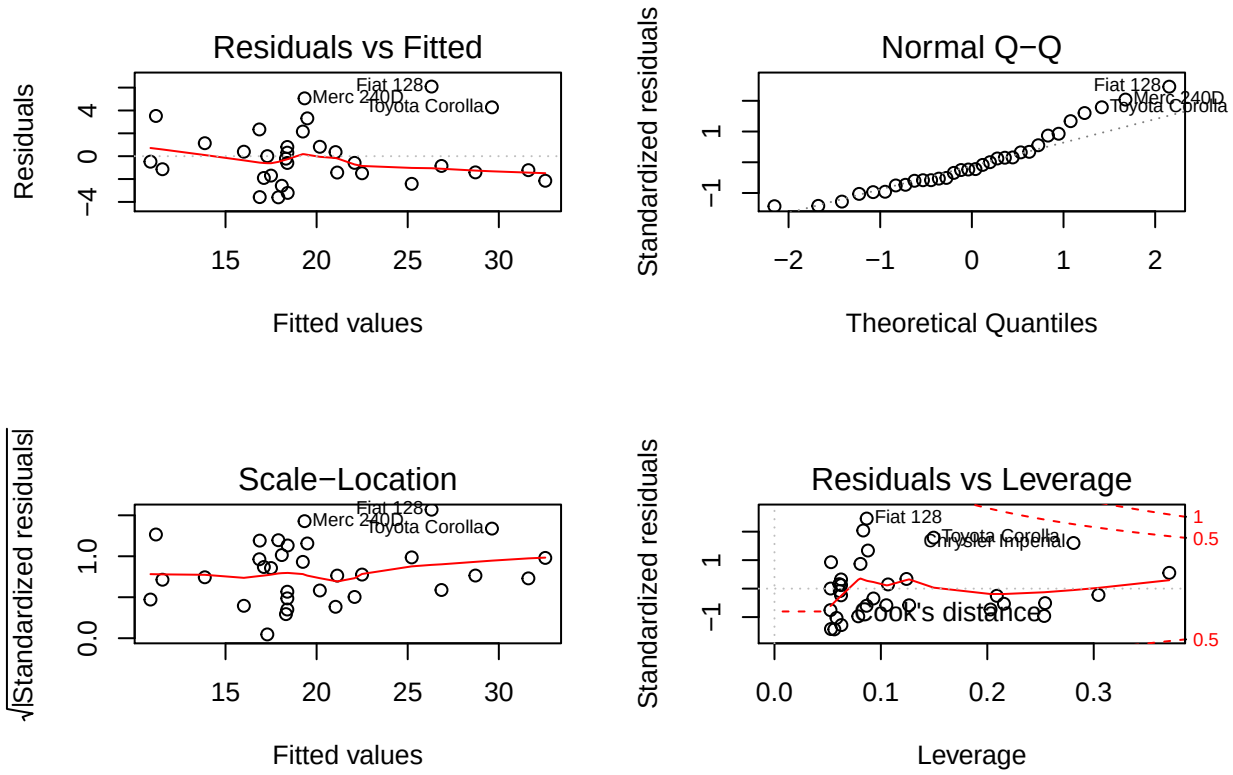

Figure 2: Diagnostics: lm(mpg ~ am * wt)