Estimating transmission impact on vehicle fuel economy

Syed Abdullah Hasan

7/15/2021

Executive Summary

This project studies the impact of vehicle transmission on fuel economy by examining a sample of 32 vehicles from 1973-74 model years studied by Motor Trend US Magazine. Initial exploratory analysis of the data set examines impact of transmission on fuel economy as well as underlying correlations between all variables. Subsequently, an iterative backward selection approach is undertaken to select the best-fit linear regression model, in which vehicle horsepower, weight and number of cylinders are identified as the key predictors for the outcome variable `mpg` alongside the transmission. Finally, the report concludes that manual transmission vehicles are generally better for fuel economy as compared with automatic transmission vehicles, and add 1.81 miles per gallon on average to fuel economy after adjusting for vehicle performance (modelled by vehicle horsepower, weight and cylinders).

Methodology

This report analyses data collected by Motor Trend US Magazine in 1974 to evaluate the impact of vehicle transmission fuel economy, by considering 10 aspects of automobile design & performance for a set of 32 automobiles (produced in the year 1973-74). The study will first explore key features of the data set, proceed to fit an appropriate regression model and subsequently evaluate the following questions:

1. Is an automatic or manual transmission better for fuel economy measured in MPG (miles per gallon)?
2. What is the difference in MPG between automatic and manual transmission vehicles?

Data Processing

The data set used for the report includes 32 observations on the following set of numeric variables:
1. `mpg` - Miles/(US) gallon
2. `cyl` - Number of cylinders
3. `disp` - Displacement (cu.in.)
4. `hp` - Gross horsepower
5. `drat` - Rear axle ratio
6. `wt` - Weight (1000 lbs)
7. `qsec` - 1/4 mile time
8. `vs` - Engine (0 = V-shaped, 1 = straight)
9. `am` - Transmission (0 = automatic, 1 = manual)
10. `gear` - Number of forward gears
11. `carb` - Number of carburetors

As seen in Chart 1 (Appendix), variables `cyl`, `vs`, `am`, `gear` and `carb` are factors with multiple levels. Therefore, the data is processed to convert each variable into a factor variable during model specification.

Exploratory Data Analysis

As seen in Chart 1 (Appendix), `mpg` is clearly correlated to each of the variables in the data set, hence all variables should be evaluated in model selection when analyzing the isolated impact of transmission on fuel economy. Chart 2 shows that automatic transmission vehicles have a better fuel economy on average as compared with manual transmission vehicles. From Chart 1, we can observe a high correlation between each of the variables in the data set with `mpg` and with each other. Correlated predictors may introduce multicollinearity in the model, confounding interpretation of potentially significant predictors. Hence, we need to be careful not to over-fit variables in the best-fit model.

Model Specification and Selection

Step-wise regression is employed to remove predictors in each iteration from a fully specified model until the best performing model with the lowest prediction error is determined. The Akaike Information Criterion (AIC) is used to evaluate the model with the best fit.

Base Model

By specifying a base regression model for `mpg` against `am` (Table 1), we can confirm that the average fuel economy (`mpg`) for manual transmission vehicles is 24.39 whereas that for automatic transmission vehicles is 17.15. Both coefficients in the model are statistically significant at an alpha level of 5%. Therefore, on first glance, it appears that the type of vehicle transmission impacts fuel economy by 7.24 miles per gallon, which is a very large difference.

However, this analysis does not account for the impact of additional variables in the data set on fuel economy. The base model has an adjusted R-squared of 33.85%, implying low predictive power. After adjusting for the impact of additional factors, the difference in fuel economy due to vehicle transmission may not be as great as seen in the base model.

Full Model

The full model includes `am`, `vs`, `carb` and `gear` as factor variables. In this model structure, the intercept term captures the expected value for fuel economy for a reference vehicle with automatic transmission, 1 carburetor, 3 gears, 4 cylinders and straight-line engine configuration.

The model has an adjusted R-squared of 77.90% and implies that the reference vehicle has an average fuel economy of 23.88 miles per gallon. A manual transmission would increase the fuel economy to 25.09 miles per gallon holding all other factors constant. After adjusting for all factors, a manual transmission improves fuel economy by 1.2121157 miles per gallon, which is lower than the value of 7.24 in the base model.

However, Table 2 shows that none of the predictors are statistically significant in this model at a 5% alpha level.

Best-fit Model

We therefore proceed to remove predictors from the model until a best-fit model can be narrowed down. Table 3 shows the summary statistics for the best-fit model with the lowest AIC for the data, for which the formula is as follows:

$$mpg = \beta_0 + \beta_1 * hp + \beta_2 * wt + \beta_3 * am_1 + \beta_4 * cyl_6 + \beta_5 * cyl_8$$

where

$$\beta_0$$

is the intercept capturing impact of transmission `am` on fuel economy `mpg` for a reference vehicle with automatic transmission and a four cylinder engine.

The F-statistic p-value for the best-fit model shows that all variables are collectively significant and all but one predictors are individually significant at the 5% alpha level. The model has an adjusted R-squared value of 84.01%, implying strong fit with the data set.

Analysis of Residuals

The selected model residuals are now screened to rule out any evidence of heteroskedasticity and influential outliers. The following observations confirm that the variance of residuals is constant:

- The plot of residuals against fitted-vales (Chart 3) for the best-fit model shows a largely random distribution.
- The Normal Q-Q plot shows majority of points falling along the dotted line, indicating that the residuals are normally distributed.
- The scale-location plot confirms a normal distribution.
- Despite the presence of two outliers (Chart 4), the Cook's D-bar plot shows that all distances are significantly less than 1 and below the threshold.
- We can formally test for heteroskedasticity using the Breush Pagan test (Table 4). At an alpha level of 5%, the null hypothesis cannot be rejected. We therefore select the current model and confirm that the variance of residuals in the model is constant.

Model Interpretation

The following conclusions may be drawn from the best-fit model selected in this analysis:

- The model explains 84.01 percent of variation in the outcome variable `mpg`.
- The model predicts that a reference vehicle witn automatic transmission has an expected fuel economy (`mpg`) of 33.71 whereas this improves by 1.81 miles per gallon, after adjusting for the impact of weight, horsepower and number of cylinders across sampled cars.
- Based on the model results, automatic transmission vehicles are generally worse for fuel economy as compared with manual transmission vehicles.

Appendix

Dataset Summary

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Tables

Table 1 - Summary Table - Full Regression Model

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.3923 -3.0923 -0.2974  3.2439  9.5077 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385 
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Table 2 - Summary Table - Full Regression Model

```
##
```

```
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + as.factor(am) +
##     as.factor(vs) + as.factor(carb) + as.factor(gear) + as.factor(cyl),
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.87913   20.06582   1.190   0.2525
## disp              0.03555    0.03190   1.114   0.2827
## hp               -0.07051    0.03943  -1.788   0.0939 .
## drat              1.18283    2.48348   0.476   0.6407
## wt               -4.52978    2.53875  -1.784   0.0946 .
## qsec              0.36784    0.93540   0.393   0.6997
## as.factor(am)1    1.21212    3.21355   0.377   0.7113
## as.factor(vs)1    1.93085    2.87126   0.672   0.5115
## as.factor(carb)2 -0.97935    2.31797  -0.423   0.6787
## as.factor(carb)3  2.99964    4.29355   0.699   0.4955
## as.factor(carb)4  1.09142    4.44962   0.245   0.8096
## as.factor(carb)6  4.47757    6.38406   0.701   0.4938
## as.factor(carb)8  7.25041    8.36057   0.867   0.3995
## as.factor(gear)4  1.11435    3.79952   0.293   0.7733
## as.factor(gear)5  2.52840    3.73636   0.677   0.5089
## as.factor(cyl)6  -2.64870    3.04089  -0.871   0.3975
## as.factor(cyl)8  -0.33616    7.15954  -0.047   0.9632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Table 3 - Summary Table - Best-fit Regression Model

```
##
## Call:
## lm(formula = mpg ~ hp + wt + as.factor(am) + as.factor(cyl),
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.70832    2.60489  12.940 7.73e-13 ***
## hp              -0.03211    0.01369  -2.345  0.02693 *
## wt              -2.49683    0.88559  -2.819  0.00908 **
## as.factor(am)1   1.80921    1.39630   1.296  0.20646
## as.factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *
```

```
## as.factor(cyl)8 -2.16368    2.28425  -0.947  0.35225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Table 4 - Breusch-pagan test for Heteroskedasticity

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##              Data
##  -------------------------------
##  Response : mpg
##  Variables: fitted values of mpg
##
##         Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    3.693268
##  Prob > Chi2   =    0.05463247
```

Charts

Chart 1 - Scatter plot of MPG against all other variables

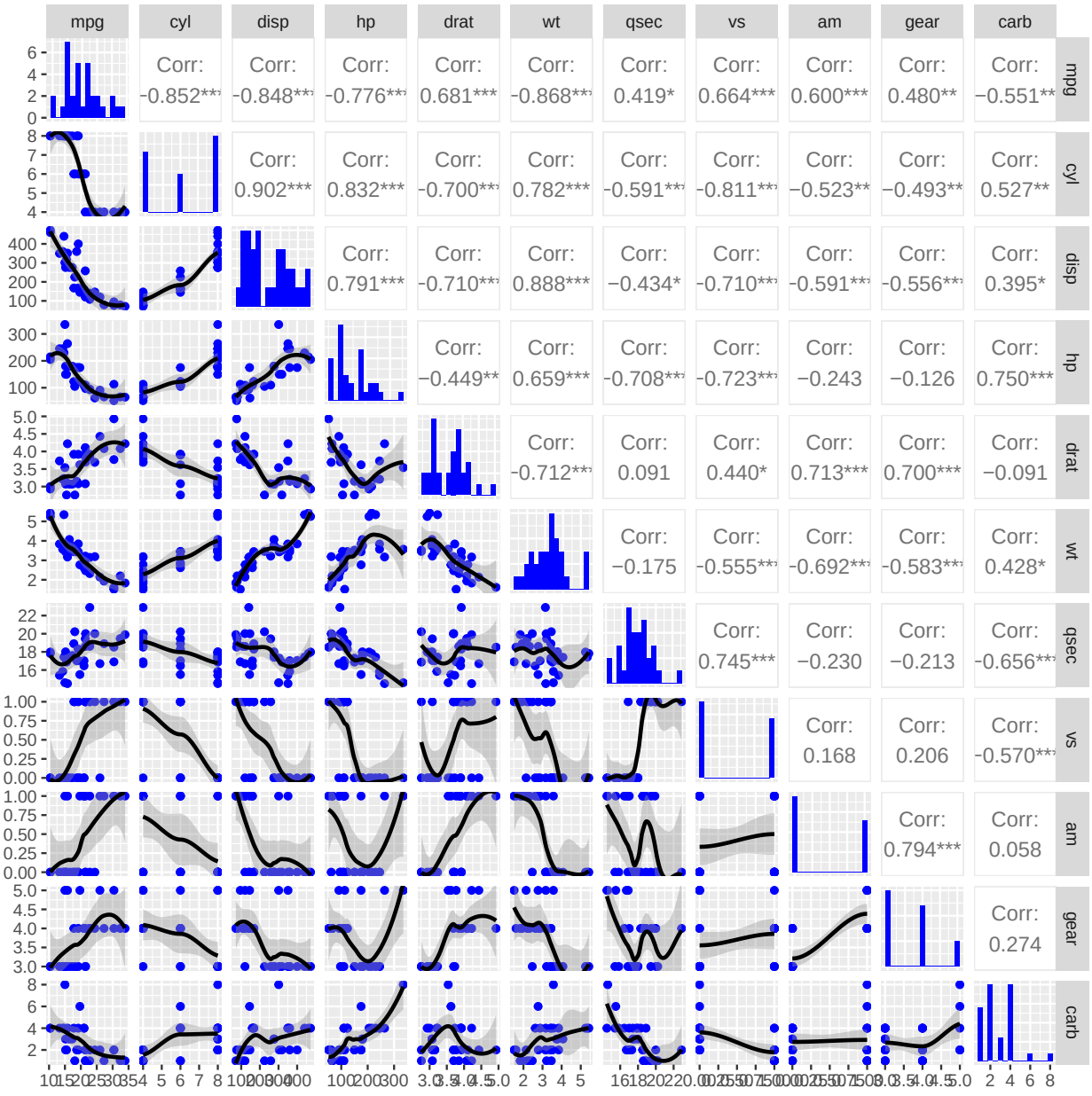

Chart 2 - Boxplot of MPG against AM

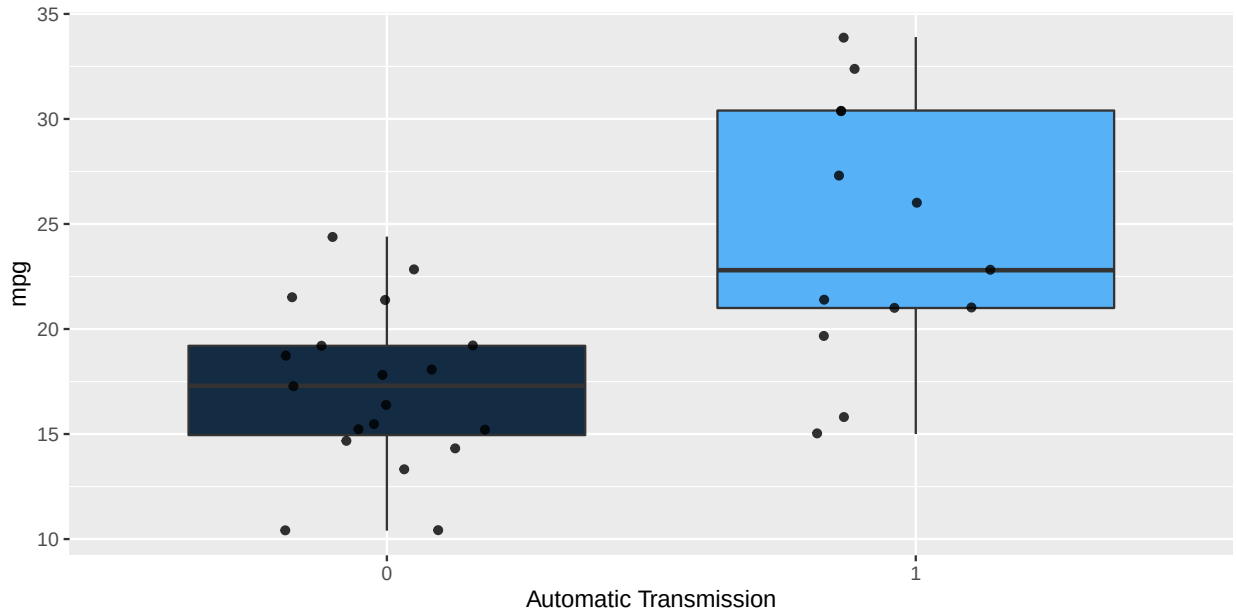

Chart 3 - Residual plots for Best-fit Model

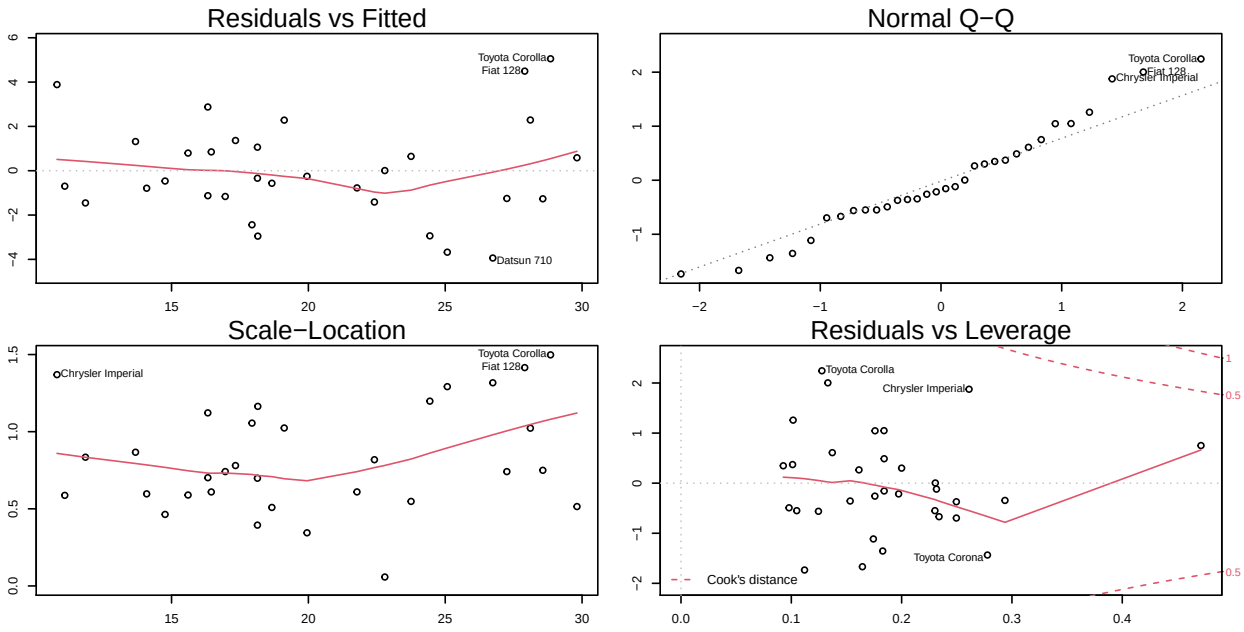

Chart 4 - Cooks D-bar plot for Best-fit Model