The Relationship Between Miles per Gallon and Transmission Type

Regression Models Course Project

Annette Spithoven

15-11-2019

Introduction

Motor Trend, an automobile trend magazine, is interested in exploring the relationship between a set of variables and miles per gallon (MPG) outcome. Of especially interest are the questions:

- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG between automatic and manual transmissions?

Executive Summary

To answer the question whether automatic or manual transmission results in a better miles per gallon, we estimated a regression model. Our final model, which explained approximately 86% of the variance in miles per gallon, suggested that manual transmission cars have a higher miled per gallon than automatic transmission cars. The increase in miles per gallon is, when the weight, horsepower and number of cylinders are held constant, approximately 1.8 when switching from an automatic transmission car to a manual one.

Data Processing and Exploration

The Data

The current document describes the analyses conducted on the mtcars dataset, from the 1974 Motor Trend US magazine, to answer these questions. Documentation on this dataset can be found here. The variables in the dataset and their classes are:

- mpg: Miles/(US) gallon (numeric)
- cyl: Number of cylinders (factor: 4,6,8)
- disp: Displacement (cu.in.) (numeric)
- hp: Gross horsepower (numeric)
- drat: Rear axle ratio (numeric)
- wt: Weight (1000 lbs) (numeric)
- qsec: 1/4 mile time (numeric)
- vs: Engine type (factor: V-shaped, straight)
- am: Transmission type (factor: automatic, manual)
- gear: Number of forward gears (factor 3,4,5)
- carb: Number of carburetors (numeric)

Is an indicator of how long it took the car from a standing start to the end of a straight 1⁄4 mile (0.40 km) track. In other words, is a performance indicator and an outcome, just like MPG, rather than a predictor for MPG. Therefore, we will remove it from the dataset.

Loading and Preprocessing

The data is loaded and immediately transformed to the proper class.

```
data(mtcars)
suppressMessages(attach(mtcars))
```

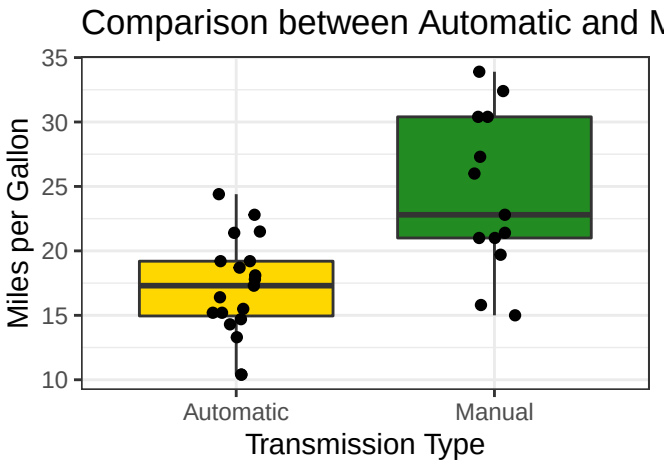
```r
mtcars <- mtcars %>%
  ## remove qsec
  select(-qsec) %>%
  ## set other variables to appropriate class
  mutate(am = factor(am,labels=c('Automatic','Manual')),
         vs = factor(vs, labels= c('V-shaped','Straight')),
         cyl = as.factor(cyl),
         gear = as.factor(gear)
  )

dim(mtcars)
```

```
## [1] 32 10
```

Exploratory Data Analysis

Our research questions concern MPG and transmission type. In the exploratory analyses it is determined whether transmission types is related to MPG.

```r
ggplot(mtcars,
       ## use fill to get nice colors
       aes(x=am, y=mpg, fill = am)) +
  ## outlier.shape is set to NA to prevent outliers to be plotted twice
  geom_boxplot(outlier.shape=NA) +
  ## add datapoints
  geom_jitter(position=position_jitter(width=.1, height=0)) +
  ## set nice theme/colors + informative titles
  theme_bw() +
  labs(x = "Transmission Type",
       y = "Miles per Gallon",
       title = "Comparison between Automatic and Manual Transmission on MPG") +
  scale_fill_manual(values = c("gold","forestgreen"))+
  theme(legend.position = "none")
```



The plot shows that manual (transmission) cars have a higher MPG than automatic (transmission) cars. The plot also shows that there is a greater variation in MPG for manual cars than for automatic cars (see Appendix Plot 1 for an even clearer representation of the differences in variation). To test whether this difference between automatic and manual cars is significant, a t-test was conducted.

```r
t.test(mtcars$mpg[which(mtcars$am == "Automatic")], mtcars$mpg[which(mtcars$am == "Manual")])
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[which(mtcars$am == "Automatic")] and mtcars$mpg[which(mtcars$am == "Manual")]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

As the p-value is smaller than 0.05 we reject the null-hypothesis indicating there is a difference in MPG between manual and automatic cars. Additional analyses for the relation between MPG and the other variables that might have explanatory importance can be found in the Appendix (see Plot 2).

Results

We already showed in the exploratory analyses that there is a difference between automatic and manual cars in MPG. We estimate a simple regression model to quantify this difference further

```r
Basic_model <- lm(data = mtcars, mpg ~ am )
summary(Basic_model)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

The R^2 is 0.3597989, which means that transmission explains 35.98% of the variance in MPG. Furthermore, the analysis indicates that on average, a car has 17.147 MPG with automatic transmission, if it is manual transmission, it is increased with 7.245 mpg.

The relative low R^2 leaves room for improvement. A new model is estimated using stepwise selection. Stepwise selection (or sequential replacement) is a method in which you start with a model with no predictors. Sequentially the most contributive predictors are added (i.e., forward selection). After adding each new variable, any variables that no longer provide an improvement in the model fit is removed (i.e., backward selection).

In order to select the best model we will look at the fit statics. The **adjusted R^2**, which is (already mentioned above) the proportion of variation in the outcome that is explained by the predictor variables while correcting

for the number of variables in the model. The higher the R-squared, the better the model.The **CP** (Mallows Cp) and **BIC** (Bayesian information criteria) are fit inidices that penalize the inclusion of additional variables to a model. In other words, the CP and BIC address the issue of overfitting and favor the most parsimonuous model. The lower the CP and BIC, the better the model.

In short, we would like to select the model with the highest **adjusted R^2** and the lowest **CP** and **BIC**. Overall, the performance of the various models is shown below.

```
model <- regsubsets(data = mtcars, mpg ~ . , method = "seqrep")
tibble(Model = 1:8, Adjusted_R2 = summary(model)$adjr2, CP = summary(model)$cp, BIC = summary(model)$bi
```

```
## # A tibble: 8 x 4
##   Model Adjusted_R2     CP   BIC
##   <int>       <dbl>  <dbl> <dbl>
## 1     1       0.745 12.6   -37.8
## 2     2       0.815  2.44  -45.7
## 3     3       0.829  1.31  -46.0
## 4     4       0.841  0.777 -45.9
## 5     5       0.848  0.954 -45.1
## 6     6       0.844  2.63  -42.1
## 7     7       0.839  4.43  -39.0
## 8     8       0.829  6.80  -34.9
```

The best fitting models are 3 (according to the BIC), 4 (according to CP) and 5 (according to the adjusted R^2). As the best fitting model we select model four, as it contains the same variables as model three while it added our variable of interest (transmission) and still more parsimonuous than the fifth model.

```
Final_model <- lm(data = mtcars, mpg ~ am + cyl + wt + hp )
summary(Final_model)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + hp, 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 ***
## amManual     1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## wt          -2.49683    0.88559  -2.819  0.00908 **
## hp          -0.03211    0.01369  -2.345  0.02693 *
## ---
## 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
```

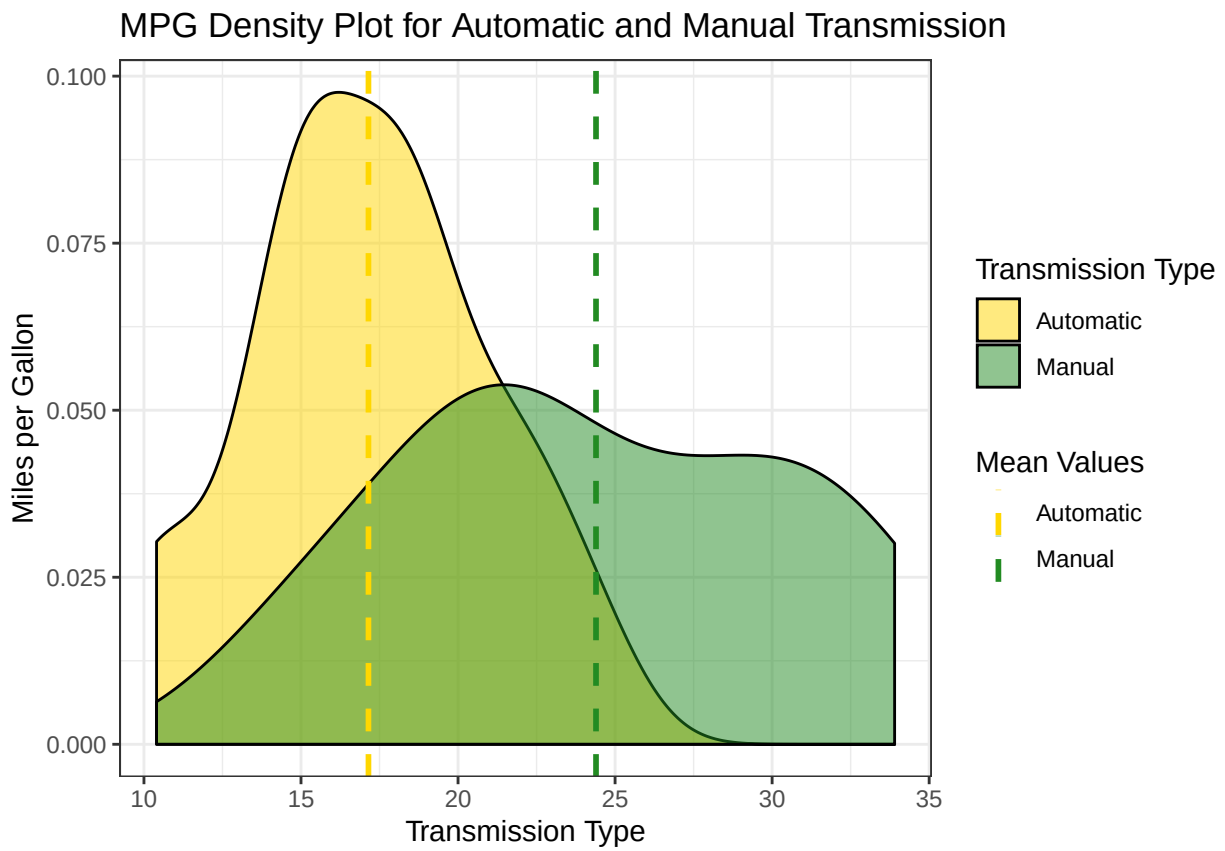
The R^2 is 0.8658799, which means that transmission explains 86.59% of the variance in MPG. See the Appendix for more information on this model. The results from this model suggest that manual transmission cars have a higher miled per gallon than automatic transmission cars. The increase in MPG is approximately 1.8 MPG when switching from an automatic transmission car to a manual one, with the weight, horsepower

and cylinders held constant. The MPG is reduced by an increased horsepower and weight. Having 6 cylinders in comparison to 2 cylinders also decreases the MPG, while having 8 cylinders in comparison to 2 cylinders does not.

Appendix

Plot 1 - Density Plot of MPG by Transmission

```
ggplot(mtcars,
       aes(x = mpg, fill = am)) +
  geom_density(alpha = 0.5) +
  geom_vline(data = mtcars %>%
               group_by(am) %>%
               summarise(mean = mean(mpg)),
             aes(xintercept = mean, color = am),
             linetype="dashed",
             lwd = 1
             ) +
  ## set nice theme/colors + informative titles
  theme_bw() +
  labs(x = "Transmission Type",
       y = "Miles per Gallon",
       title = "MPG Density Plot for Automatic and Manual Transmission",
       fill = "Transmission Type",
       color = "Mean Values") +
  scale_fill_manual(values = c("gold","forestgreen"))+
  scale_color_manual(values = c("gold","forestgreen"))
```

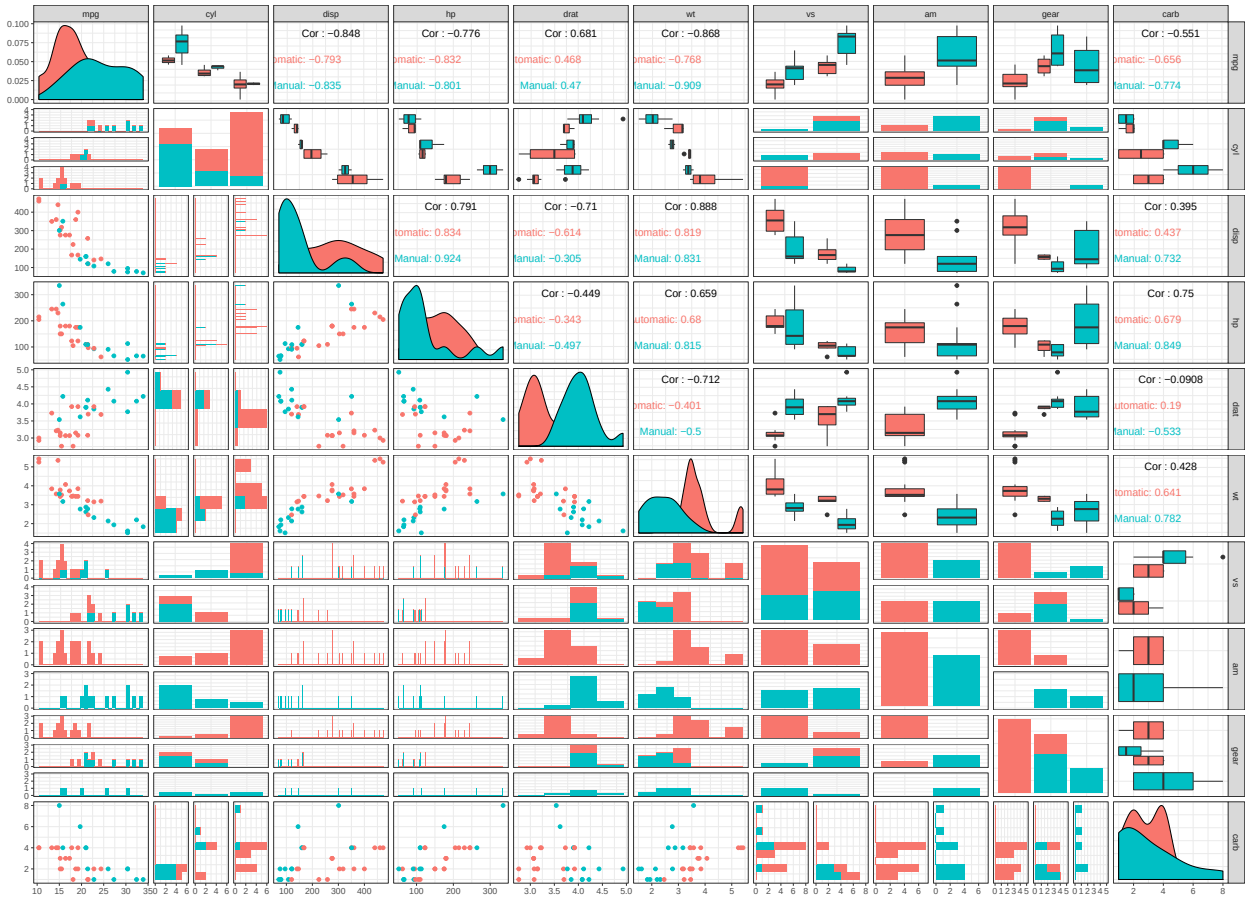


As you can see, the automatic and manual cars start at the same minimum, while the manual cars have a higher maximum MPG.

Plot 2 - Relation Between MPG and All Other Variables

Coloring is by transmission type.

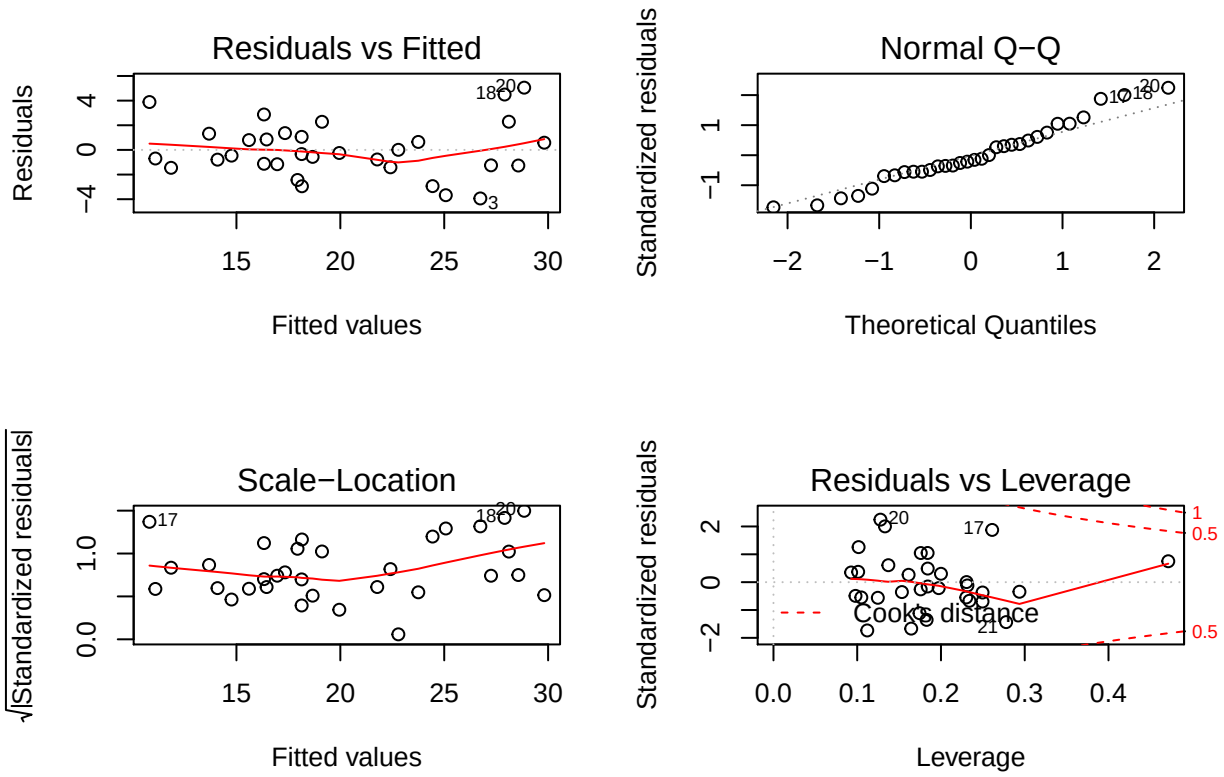
```
ggpairs(data = mtcars,
        aes(color = am),
        ## to prevent the stat_bin error from occuring
        lower=list(combo=wrap("facethist", binwidth=0.8))
        ) +
  ## set nice theme/colors + informative titles
  theme_bw()
```



Note that the correlations between the variables are rather high, this might cause problems with multicolinearity in the model.

Final Model diagnostics

```
par(mfrow = c(2,2))
plot(Final_model)
```



The plot above shows that the underlying assumptions of regression are being met:

1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.