Coursera Regression Models Course Project

Anne Strader

8. December 2020

Executive Summary

The automobile industry magazine *Motor Trend* is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

The analysis shows that a manual transmission is better for MPG. According to the final selected linear model, using a manual transmission instead of automatic increases the MPG by 4.16.

Dataset

The structure of the dataset "mtcars" is given here, with categorical variables converted to factors:

```
##                   mpg cyl disp  hp drat    wt  qsec       vs        am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46 V-shaped    manual    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02 V-shaped    manual    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61 straight    manual    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44 straight automatic    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02 V-shaped automatic    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22 straight automatic    3
##                   carb
## Mazda RX4            4
## Mazda RX4 Wag        4
## Datsun 710           1
## Hornet 4 Drive       1
## Hornet Sportabout    2
## Valiant              1
```

Exploratory Data Analysis

To get a basic idea of how the distribution of mpg depends on transmisison type, the five-number summaries and means of mpg for manual and automatic transmission types are shown here:

Automatic transmission:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

Manual transmission:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

These results are visualized as boxplots in Appendix Section A.1. MPG appears to be substantially higher for cars with manual transmissions. The mean MPG for cars with automatic transmissions is 17.15, whereas the mean MPG for cars with manual transmissions is 24.39.

To determine if this difference is statistically significant, a two-sample, two-sided T-test is conducted (see Appendix section A.2). The p-value, 0.000285, is lower than 0.025. Therefore, the null hypothesis can be rejected at the 0.05 significance level, and the mean MPG for cars using manual transmissions is significantly higher than the MPG for cars using automatic transmissions.

Using a pairplot, the correlations between all variables in the dataset are visualized in Appendix Section A.3. Although transmission type significantly impacts MPG, several other variables appear highly correlated with MPG as well.

Regression Models

Based on the pairplot observed in Appendix Section A.3, multiple variables, including transmission type, appear to be strongly correlated with MPG. Therefore, a set of nested linear models is evaluated to determine the extent to which the addition of each variable reduces the residual sum of squares (RSS). First, MPG is only modeled as a function of transmission type; then, each other variable is added individually to the model.

ANOVA test results (see Appendix Section A.4) suggest that transmission type is not the only variable that considerably affects MPG. Reductions in the RSS when adding the number of cylinders (cyl) and gross horsepower (hp) as predictors are significant, with P(>F) values lower than 0.05.

The final model is defined as: MPG as a function of transmission type (am), number of cylinders (cyl) and gross horsepower (hp). Justification for the model selection as well as a summary of the coefficients and their corresponding uncertainties is provided in Appendix Section A.5. The model fits the data well, with an adjusted R-squared value of 0.7989.

Residual Analysis

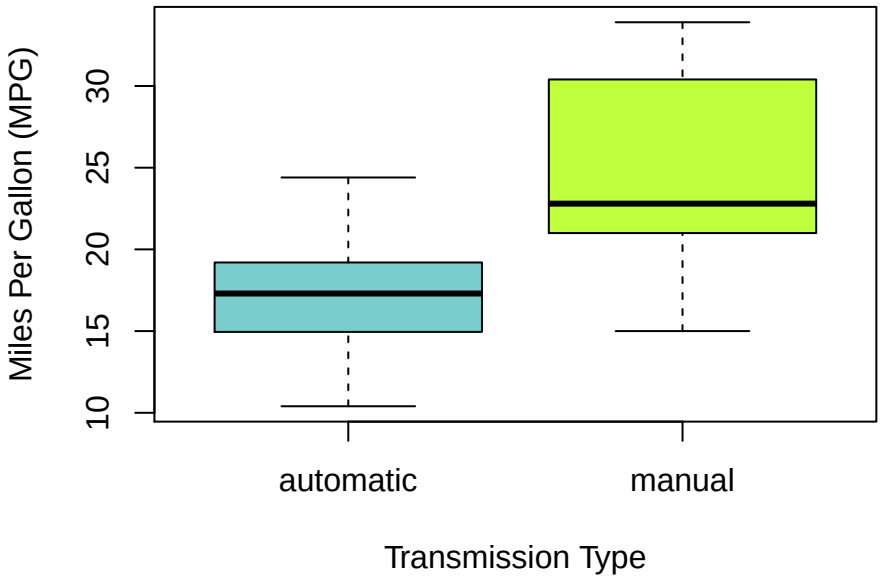
Residual diagnostics are shown in Appendix Section A.6. The distribution of residuals appears to be normal and homoscedastic. With the exception of a few outliers which could be analyzed in more detail, the linear model fits the data effectively.

Conclusions

By using a manual transmission instead of automatic, the MPG is increased by 4.16, with a standard error of 1.26. However, factors such as the number of cylinders and gross horsepower also significantly influence MPG. Furthermore, the effects of confounding variables (for example, number of cylinders and displacement in cubic inches) should be taken into account when quantifying these effects. Considering these predictors also reduces the apparent effect of transmission type on MPG, though not to the extent that the effect becomes insignificant.

Appendix

A.1: Boxplots of MPG by transmission type



A.2: T-test: Mean MPG by transmission type

The hypotheses are defined as follows:

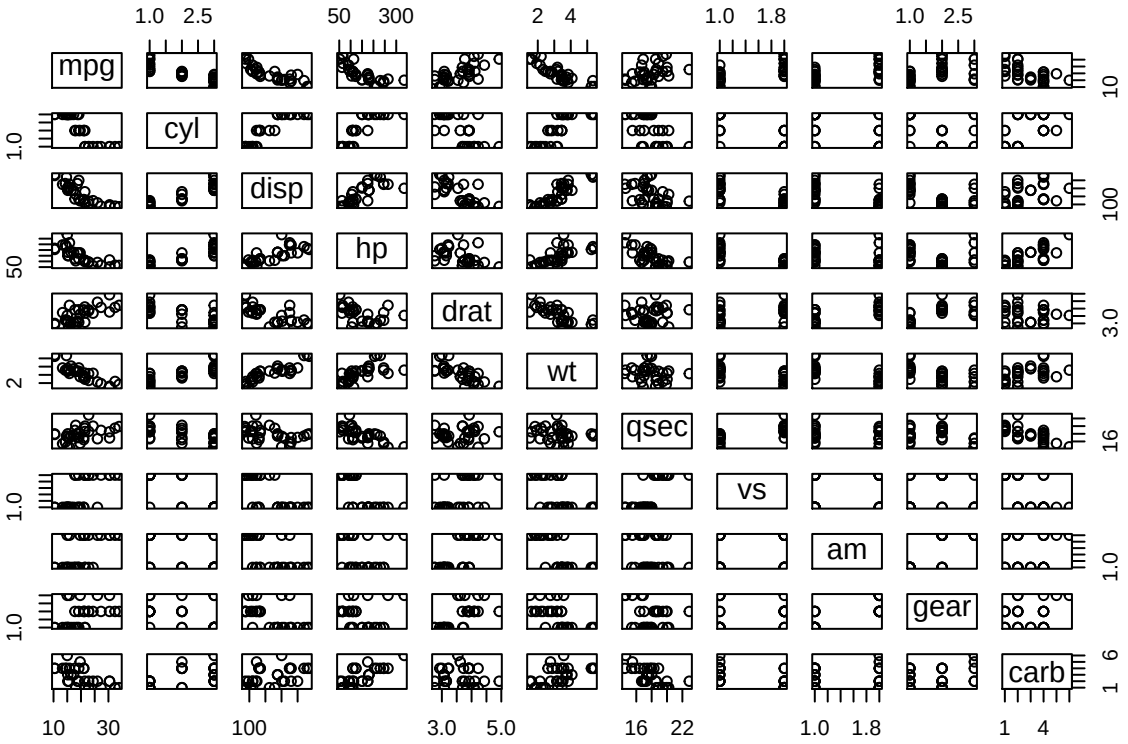
- Null hypothesis: The difference in the mean MPG between transmission types is not significantly different from zero.
- Alternate hypothesis: The difference in the mean MPG between transmission types is significantly different from zero.

The variances in MPG for both groups are assumed to be equal.

```
##
##  Two Sample t-test
##
## data:  mpg by am
## t = -4.1061, df = 30, p-value = 0.000285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.84837  -3.64151
## sample estimates:
## mean in group automatic    mean in group manual
##                17.14737                24.39231
```

A.3: mtcars dataset pairplot

The following plot visualizes correlations between all variables in the mtcars dataset:



A.4: Nested linear models and ANOVA test results

A set of ten nested linear models is defined. The simplest model only includes transmission type as a predictor. The other predictors are added one by one to the other models. To determine the relative impact of each predictor on model variance, an ANOVA test is run:

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + cyl
## Model  3: mpg ~ am + cyl + disp
## Model  4: mpg ~ am + cyl + disp + hp
## Model  5: mpg ~ am + cyl + disp + hp + drat
## Model  6: mpg ~ am + cyl + disp + hp + drat + wt
## Model  7: mpg ~ am + cyl + disp + hp + drat + wt + qsec
## Model  8: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs
## Model  9: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 10: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
##    Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1      30 720.90
## 2      28 264.50  2    456.40 28.4297 7.89e-06 ***
## 3      27 230.46  1     34.04  4.2402  0.05728 .
## 4      26 183.04  1     47.42  5.9078  0.02809 *
## 5      25 182.38  1      0.66  0.0820  0.77855
## 6      24 150.10  1     32.28  4.0216  0.06331 .
## 7      23 141.21  1      8.89  1.1081  0.30916
## 8      22 139.02  1      2.18  0.2719  0.60964
## 9      20 134.00  2      5.02  0.3128  0.73606
## 10     15 120.40  5     13.60  0.3388  0.88144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.5: Model selection

Based on the ANOVA test results of the previous section, the final linear model is defined using the predictors "am" (transmission type), "cyl" (number of cylinders) and "hp" (gross horsepower). Due to high correlation with other predictors and in the interest of avoiding variance inflation, the variables "disp" (displacement

in cubic inches) and "wt" (weight, in 1000 lbs.) are left out of the final model. The summary of the final model is provided below:

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.231 -1.535 -0.141  1.408  5.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.29590    1.42394  19.169  < 2e-16 ***
## ammanual     4.15786    1.25655   3.309  0.00266 **
## cyl6        -3.92458    1.53751  -2.553  0.01666 *
## cyl8        -3.53341    2.50279  -1.412  0.16943
## hp          -0.04424    0.01458  -3.035  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 27 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7989
## F-statistic: 31.79 on 4 and 27 DF,  p-value: 7.401e-10
```

A.6: Residual plots