Effect of transmission type on fuel efficiency

Regression Models Project

Executive Summary

Motor Trend is a magazine about the automobile industry. They are looking at a data set of a collection of cars and they are interested in exploring the relationship between a set of variables and miles per gallon (MPG). They are particularly interested in finding if an automatic or manual transmission is better for MPG and also quantify the MPG difference between automatic and manual transmissions.

The project explores the relationship between miles-per-gallon (MPG) and other variables in the mtcars data set. In particular, the analysis attempts to determine whether an automatic or manual transmission is better for Miles Per Gallon(MPG).

The Analysis section of this document focuses on inference with a simple linear regression model and a multiple regression model. Both models support the conclusion that the cars in this study with manual transmissions have on average significantly higher MPG's than cars with automatic transmissions. The data set includes 32 automobiles of which 13 have manual transmission and 19 have automatic transmission.With respect to the data set it shows that there is a difference in fuel efficency depending on transmission type.

However, transmission type is not the sole factor or responsible or a good predictor of fuel efficiency. We applied the analysis of variance to the dataset, thereby calculating the correlations between the variables, and building a number of models based on that. This analysis showed that the number of cylinders and the weight of the automobile are good predictors of fuel efficiency. With an adjusted R squared of 0.82. To prove our assumption we also added transmission type to this model. The final outcome showed that the difference in fuel effiency for a manual transmission is much smaller, just 0.18 miles per gallon for a vehicle with the same weight and number of cylinders.

Therefore we can conclude that number of cylinders and weight are good predictors of fuel efficiency. Taking transmission type alone is not a good predictor of fuel efficiency. ##Data Transformation## We will first load the mtcars data set and then tranform certain variable into factors.

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Exploratory Data Analysis

The next step is to analyse relationships between different variables of interest. To begin with we will plot the relationship between all the variables of the mtcars dataset. This plot deffinitely shows that there is a strong correlation between the variables cyl, disp, hp, drat, wt, vs and am with mpg (Appendix - Figure 1).

The next analysis is the effect of car transmission type on mpg (Appendix - Figure 2). This is achieved by plotting a box plot for the distribution of mpg for each level of am (Automatic or Manual). This plot clearly shows that manual transmission tend to have higher mpg. A linear model analysis is a further extension of this which is discussed in regression analysis section. ##Regression Model Analysis## Here we build linear regression models using different variables in order to find the best fit and compare it with the base model. After selecting a model we also perform analysis of residuals. ###Model building and selection### For our initial model we included all variables as predictors of mpg. After stepwise model selection to select significant predictors for the final model to get a best model.This is achieved by using step function which will perform this selection by calling lm repeatedly to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods. This ensures that we have included useful variables while omitting ones that do not contribute significantly to predicting mpg.

```
initial_model <- lm(mpg ~ ., data = mtcars)
best_fit_model <- step(initial_model, direction = "both")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
## + carb  5   13.5989 120.40 76.403
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## + gear  2    5.0215 134.00 69.828
## - wt    1   31.1896 170.21 71.482
```

```
## + carb  5   14.6475 124.38 73.442
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## + drat  1    0.9672 139.02 67.005
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## + gear  2    4.8033 135.19 68.110
## - wt    1   30.7151 170.71 69.574
## + carb  5   13.0509 126.94 72.095
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## + disp  1     1.247 139.99 65.227
## + drat  1     0.666 140.57 65.359
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## + gear  2     4.684 136.55 66.431
## - wt    1    39.645 180.88 69.428
## + carb  5     2.331 138.91 72.978
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
## - cyl   2    25.284 168.96 63.246
## + qsec  1     2.442 141.24 63.511
## - am    1    16.443 160.12 63.527
## + disp  1     0.589 143.09 63.928
## + drat  1     0.330 143.35 63.986
## + gear  2     3.437 140.24 65.284
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
## + carb  5     3.480 140.20 71.275
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
```

```
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## + vs    1     7.346 143.68 62.059
## + qsec  1     7.044 143.98 62.126
## - cyl   2    29.265 180.29 63.323
## + disp  1     0.617 150.41 63.524
## + drat  1     0.220 150.81 63.608
## + gear  2     1.361 149.66 65.365
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
## + carb  5     5.633 145.39 70.438
```

The best fit model obtained from the above computations shows that variables cyl, wt and hp as confounders and am as the independent variable. This is shown by using the summary command.

```r
summary(best_fit_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, 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 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## 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 adjusted R-squared value of 0.84 is the maximum obtained using all combinations of variables. We can conclude that more than 84% of the variability is explained by the above model.

Now we compare the base model with only am as the predictor variable and the best model containing confounder variables also.

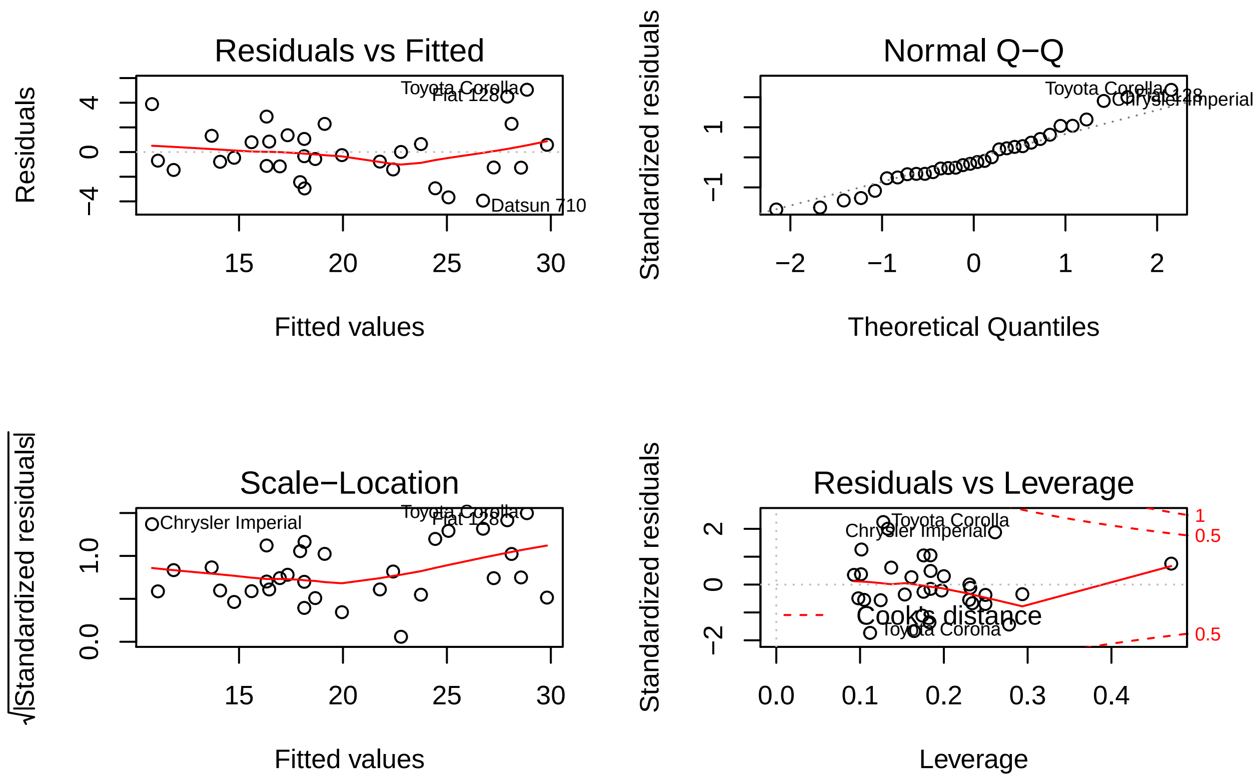
```r
base_fit_model <- lm(mpg ~ am, data = mtcars)
anova(base_fit_model, best_fit_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the above results, the p-value obtained is highly significant and so we can reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model. ##Model Residuals and Diagnostics## In this section, we have the residual plots of our regression model along with computation of regression diagnostics for our liner model. This excercise helped us in examining the residuals and finding leverage points to find any potential problems with the model.

```r
par(mfrow=c(2, 2))
plot(best_fit_model)
```



Following observations are inferred from the above plots:-

- The points in the Residuals vs. Fitted plot are randomly scattered which verifies the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots that may indicate values of increased leverage of outliers.

In the following section, we show computation of some regression diagnostics of our model to find out these leverage points. We compute top three points in each case of influence measures.

```
leverage_point <- hatvalues(best_fit_model)
tail(sort(leverage_point),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
influential_point <- dfbetas(best_fit_model)
tail(sort(influential_point[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

Based on above results, we can infer that our analysis was correct. As there are the same cars models mentioned in the residual plots. ##Statistical Inference## We finally perform a t-test on the two subsets of mpg data: manual and automatic transmission. Assuming that the transmission data has a normal distribution and tests the null hypothesis that they come from the same distribution.

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Appendix

Figure 1 - Pairs plot for the "mtcars" dataset

```
pairs(mpg ~ ., data = mtcars)
```

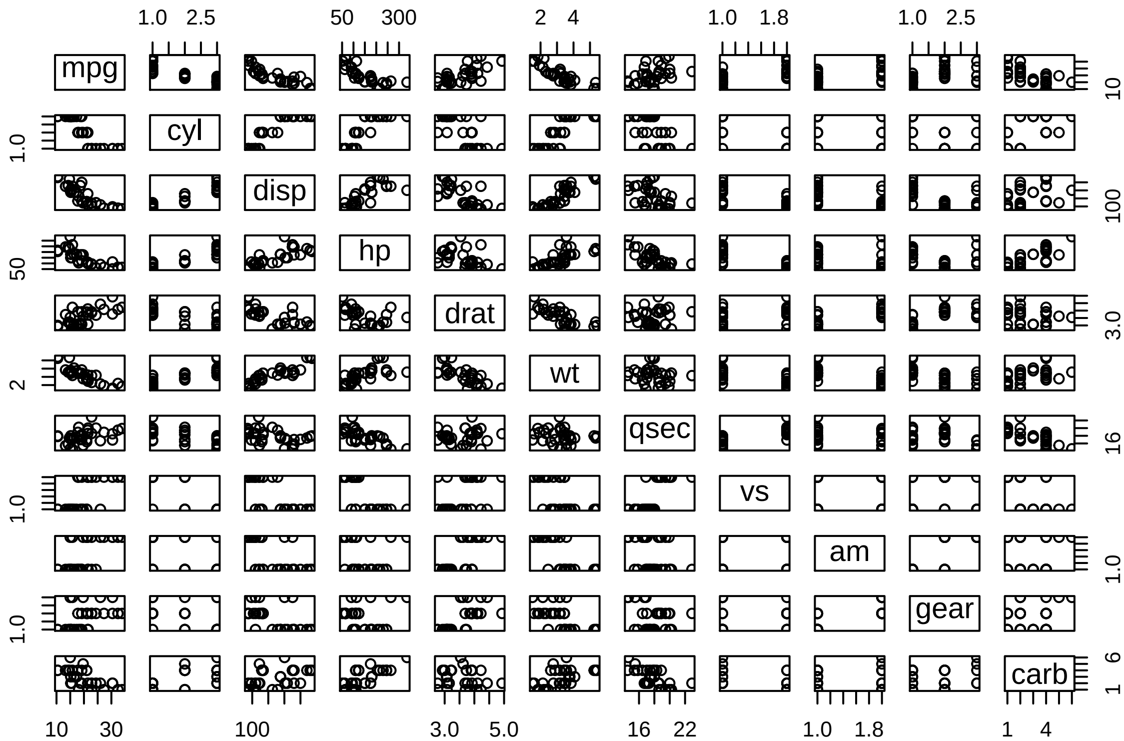
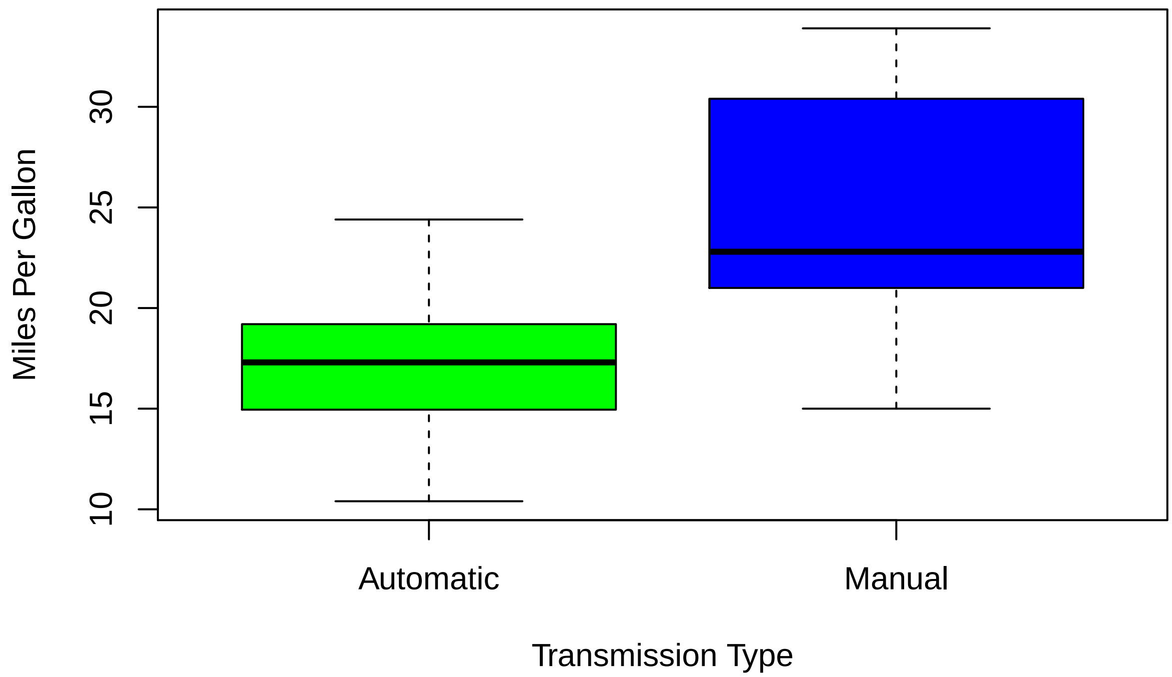


Figure 2 - Boxplot of miles per gallon by transmission type

```
boxplot(mpg ~ am, data = mtcars, col = (c("green","blue")), ylab = "Miles Per Gallon", xlab = "Transmis
```



Conclusions

As per the analysis done in this project, we can conclude that:

- Cars with Manual transmission get 1.8 more miles per gallon compared to cars with Automatic transmission.
- mpg will decrease by 2.5 for every 1000 lb increase in wt.
- With the increase in number of cylinders from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively.