Motor Trend Magazine

Rich Robinson

Saturday, June 13, 2015

Executive Summary

This analysis explores the relationship between the transmission type of cars and the affect this has on fuel efficiency, measured in Miles per gallon (MPG). In particular the analysis needs to address the following points:

- Is an automatic or manual transmission better for MPG?
- Quantify the difference in MPG between manual and automatic transmissions

To answer the above questions the analysis will first look into the basic statistical properties of the data using mean and standard deviation to gain a general idea of trends. Followering this a more detailed statitical analysis will include the fitting of linear models which can be used to predict future values within a level of confidence. Further, the analysis will look into multivariable linear regression to improve model fit. This will cover selection of regressors and quantifying the outcomes to reach a conclusion.

Exploring the Data Set

We are looking at the *mtcars* dataset provided in `R` , so let’s load it and have a look.

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We can see from the above table there is much information in each record, but for our analysis we are only interested in the `mpg` and `am` columns. The `am` data is a boolean value signifying if the transmission type is manual (1) or automatic (0). For ease of later analysis, we shall convert this in a factor variable.

```
## Source: local data frame [2 x 3]
## 
##       trans     mean       sd
## 1 automatic 17.14737 3.833966
## 2    manual 24.39231 6.166504
```

The plot in Appendix A shows the recorded MPG by transmission type. By observation it can be seen that for cars with automatic transmission the MPG is generally lower than that of manual cars. Also the range of MPG values for automatic cars is much smaller (more concentrated) than that of manual cars. The blue point for each transmission type represents the mean value of the data. The mean MPG for cars with automatic transmission is 17.147 MPG and the mean for manual cars is 24.392 MPG, which confirms the trend of higher MPG for manual cars. To quantify the spread of the data we can look at the standard deviation from the mean. The standard deviation of automatic cars is 3.834 MPG which is lower than that for manual cars of 6.167 MPG, which indicates a range of values closer to the mean.

Hypothesis Testing

A two-sided T-test can quantify if there is a significant difference in the mean values of each group (automatic and manual) and something we should investigate more. Our default and alternative hypotheses are:

$$H_0 : \mu(automatic) = \mu(manual)$$

$$H_a : \mu(automatic) - \mu(manual) \neg 0$$

```
# splitting the data into two subsets by transmission
autoMPG <- dat %>% filter(trans == "automatic")
manMPG <- dat %>% filter (trans == "manual")
t1 <- t.test(autoMPG$mpg, manMPG$mpg, paired = FALSE, var.equal = FALSE)
t1
```

```
##
##  Welch Two Sample t-test
##
## data:  autoMPG$mpg and manMPG$mpg
## 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
```

The results of a two-sided, non-paired t-test with unequal variances shows a p-value of 0.0013736. which is smaller than our significance rate of `0.05` (5%) which means we would reject the null hypothesis. Also the 95% confidence interval does not include zero which again indicates that there is a significant difference in means between the two groups and we should reject the null hypothesis.

Fitting a Linear Regression Model

Let’s see what a basic relationship might show between the outcomes of the groups.

```
fit1 <- lm(mpg ~ am, data = dat)
summary(fit1)
```

```
## 
## Call:
## lm(formula = mpg ~ am, data = dat)
## 
## 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 ***
## am             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 summary information from our regression model shows little more information about our data subset. The coefficients of the model $Y_i = \beta_0 + \beta_1 \times X_1$ are; $\beta_0 =$ **17.147** (intercept) which is the mean MPG of cars with automatic transmission and $\beta_1 =$ **7.245** (slope) that on average cars with manual transmission have a higher MPG by **7.245** mpg. The R^2 value shows us how much of the variation is explained by the regression model. Here, this basic model explains only **35.98%** of the variation in MPG values.

Multivariable Linear Regression

It is logical that more attributes will affect MPG than just transmission type. Now lets look for strong correlations between other attributes and MPG using the `corr()` function.

```
##    attr       corr
## 1   mpg  1.0000000
## 2    wt -0.8676594
## 3   cyl -0.8521620
## 4  disp -0.8475514
## 5    hp -0.7761684
## 6  drat  0.6811719
## 7    vs  0.6640389
## 8    am  0.5998324
## 9  carb -0.5509251
## 10 gear  0.4802848
## 11 qsec  0.4186840
```

The output above shows that many attributes have a stronger affect on MPG than transmission type (am), namely weight, cylinders, displacement, horse power. This would suggest that a multivariate regression model would better predict the possible MPG of a vehicle, with predictors of weight, cylinders, displacement and horse power. Lets look into this further. (See appendix B for the code for the above calculation and a pairs plot for illustration)

However although some of the variables in our dataset correlate well with MPG, they might also correlate well with eachother. We ideally would like a regression model that is parsimonious, which means it has as few confounders as possible which are all orthogonal to one another. To help with this we can use *"Variance Inflation"* to show how a regressor (variable) will affect a model compared to when it is (ideally) orthogonal to other regressors. In `R` we use the `vif()` function in the `car` package.

```
##    Regressor       vif
## 1       disp 21.620241
## 2        cyl 15.373833
## 3         wt 15.164887
## 4         hp  9.832037
## 5       carb  7.908747
## 6       qsec  7.527958
## 7       gear  5.357452
## 8         vs  4.965873
## 9         am  4.648487
## 10      drat  3.374620
```

The above results show that the most influential regressor are displacement, cylinders, weight and horse power. Thus, lets try a linear model with these regressors plus our cofounder of interest, transmission. We can then use the `anova()` function to compare our two regression models.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + disp + cyl + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 163.12  4    557.78 22.226 4.507e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

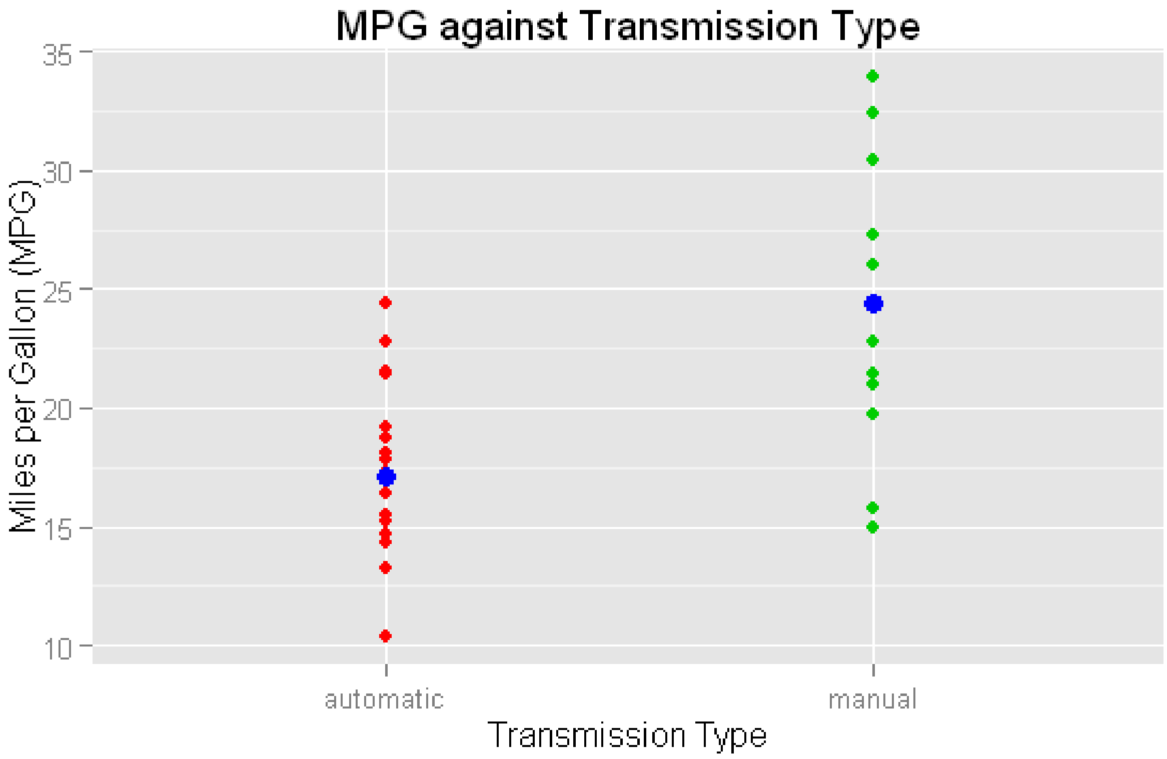
The results show a very small p-value which indicates the additional regressors in the second model are of significance in producing a btter regression model than out original. Recall that the original regresion model explained 35.98% of the variation. The new model explains **85.51%**. In comparison, a regression model using the remaining cofounders to model MPG has an R^2 value of 78.83%, so the new model better explains the variation in MPG than the cofounders we chose to omit.

Conclusion

In conclusion we can now say that even given the additional cofounders of displacement, cylinders, weight and horse power that car with manual transmission have better MPG than those with automatic transmission. From the summary information in Appendix C we can see that manual cars on average are **1.56**mpg higher. However the associated p-value is relatively high, so maybe a more sophistocated model is required for a high significance level. The residual plots in Appendix D show that the residuals are fairly evenly distributed around zero meaning that the model is more *homoskedastic* than *heteroskedastic* (non constant variance), which indicates the model is a *'good'* fit. The plots also highlight a few high influence points (Chrysler Imperial & Toyota Corolla). Further analysis could look specifically at these records and using *influence measures* assess the affect these have had on the model and would a more accurate model be possible if we chose to omit them.

Appendix A - Code: Plot: MPG vs Transmission

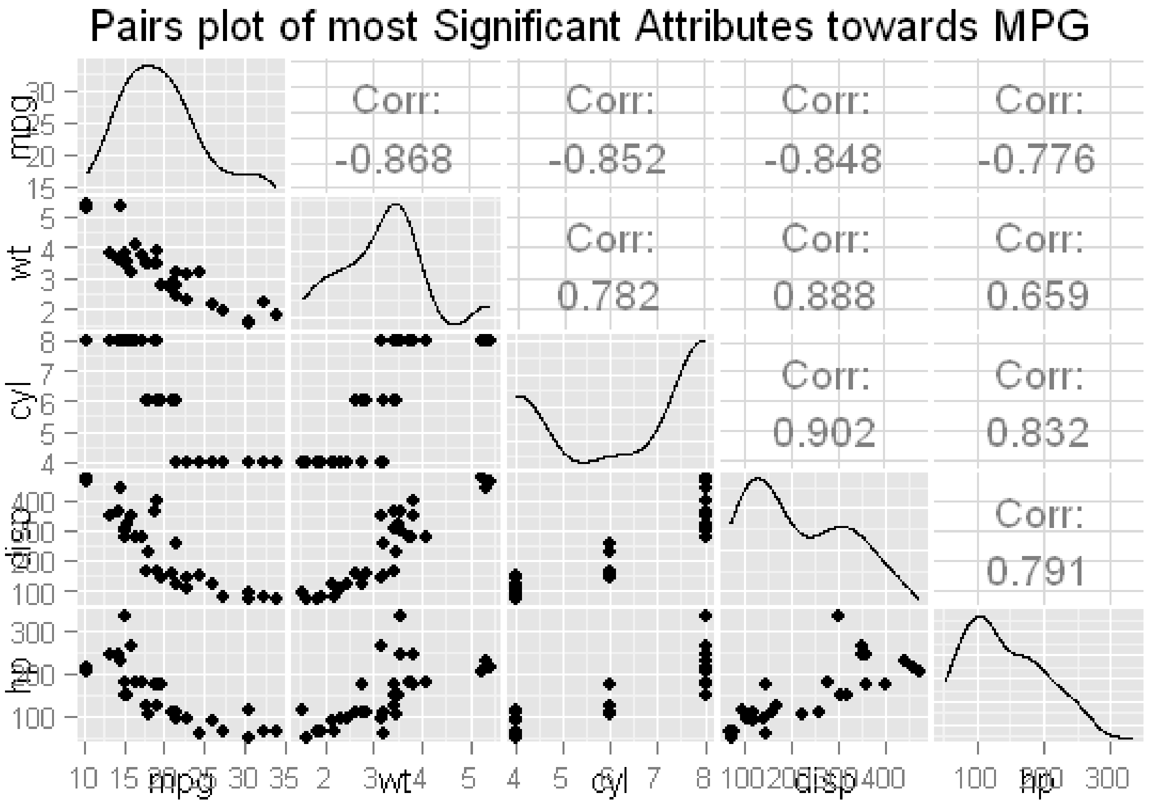
```
dat <- mtcars %>% select(mpg, am)

## adding a new column with am variable as a factor
dat <- dat %>% mutate(trans = factor(am, labels = c("automatic", "manual")))

g1 <- ggplot(dat, aes(x = trans, y = mpg)) + geom_point(color = dat$am+2)
g1 <- g1 + labs(title = "MPG against Transmission Type") + labs(x = "Transmission Typ
e") + labs(y = "Miles per Gallon (MPG)")

## calculating some basic stats
stats <- dat %>% select(mpg, trans) %>% group_by(trans) %>% summarise_each(funs(mean, s
d))
stats

## adding the stats as reference points on plot
g1 <- g1 + geom_point(data = stats, aes(x = trans, y = mean), color = "blue", size = 3)
```



Appendix B - Code: Plot: Correlation between Attributes

```
## calculating correlations of the data
foo <- cor(mtcars, mtcars$mpg)
corrs <- data.frame(attr = rownames(foo), corr = foo[,1])
## arranging into order of significance by ABS value to preserve relationship (positive
or negative)
corrs <- arrange(corrs, desc(abs(corr)))
corrs
```

```
## selecting column numbers
col_nums <- as.numeric(order(desc(abs(foo))))
## creating pairs plot
ggpairs(select(mtcars, col_nums[1:5]), title = "Pairs plot of most Significant Attribut
es towards MPG")
```



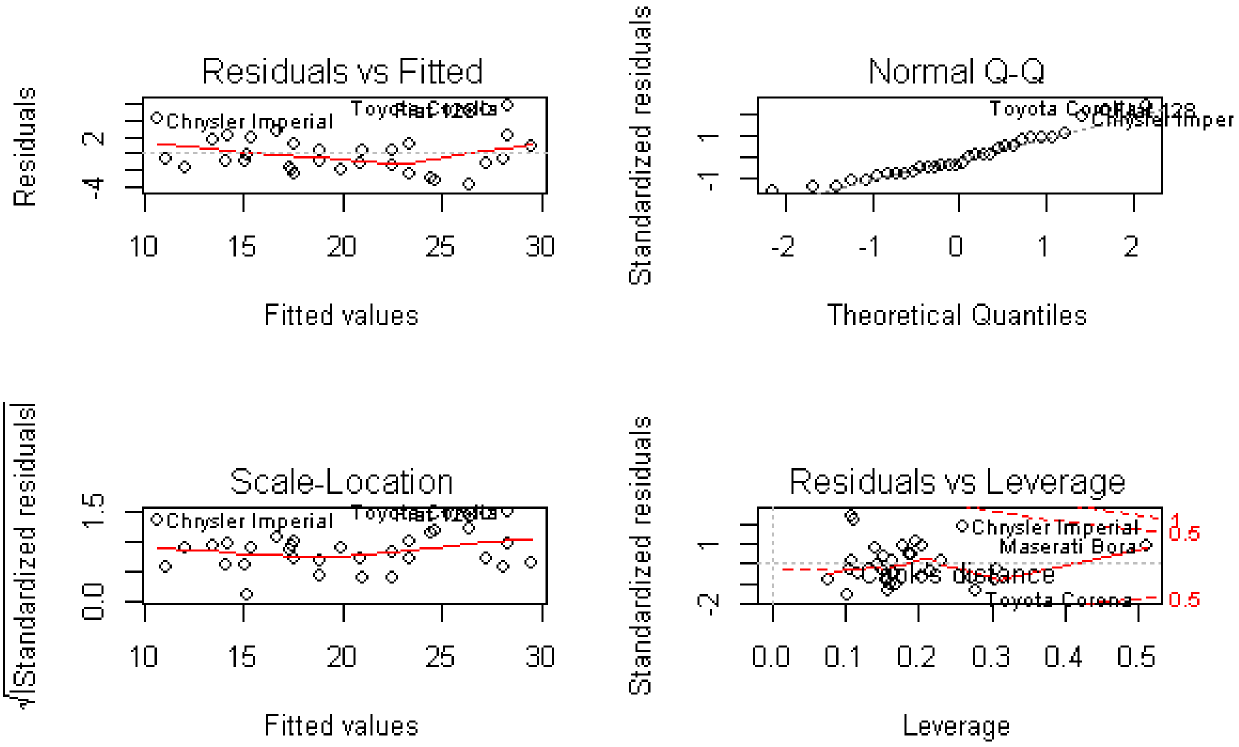
Appendix C - Code: VIF & Regressor Selection

```
fit <- lm(mpg ~ ., data = mtcars)
bar <- vif(fit)
vifs_long <- data.frame(Regressor = names(bar), vif = bar)
vifs_long <- arrange(vifs_long, desc(vif)) # arranging
vifs_wide <- dcast(vifs_long, .~ Regressor, value.var="vif")[, 2:(dim(vifs_long)[1]+1)]
# for better presentation
vifs_long
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + disp + cyl + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am           1.55649    1.44054   1.080  0.28984
## disp         0.01226    0.01171   1.047  0.30472
## cyl         -1.10638    0.67636  -1.636  0.11393
## wt          -3.30262    1.13364  -2.913  0.00726 **
## hp          -0.02796    0.01392  -2.008  0.05510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

Appendix D - Plot: Residual Plots


Residuals vs Fitted
Residuals
-4
2
Chrysler Imperial
Toyota Corolla
Fiat 128
10
15
20
25
30
Fitted values
Normal Q-Q
Standardized residuals
-1
1
Toyota Corolla
Fiat 128
Chrysler Imper
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
0.0
1.5
Chrysler Imperial
Toyota Corolla
Fiat 128
10
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
-2
1
Chrysler Imperial
Maserati Bora
Cook's distance
Toyota Corona
1
0.5
0.5
0.0
0.1
0.2
0.3
0.4
0.5
Leverage