Motortrend MPG Analysis

Thomson Kneeland

May 19, 2016

Executive Summary

This paper examines the mtcars dataset, exploring the relationship between mpg (miles per gallon) and other variables. Of particular interest is whether manual or automatic transmission offers a better mpg outcome, and if so, by how much. We examine the data using 1) a base linear model with transmission and mpg only 2) a "best fit" multiple regression model we find with the data, and then 3) compare the two and check for uncertainties.

Although a simple analysis concludes that manual transmissions offer a mean increase of 7.2 mpg over automatic transmissions, a more complex model including weight, cylinder, and horsepower accounted for more mpg variability (83%), with a change in mpg for transmission type of only 1.8 mpg.

Exploratory Analysis

Since we are primarily interested in the relationship between transmission type and mpg, a t.test will ascertain if manual vs automatic significantly affects mpg.

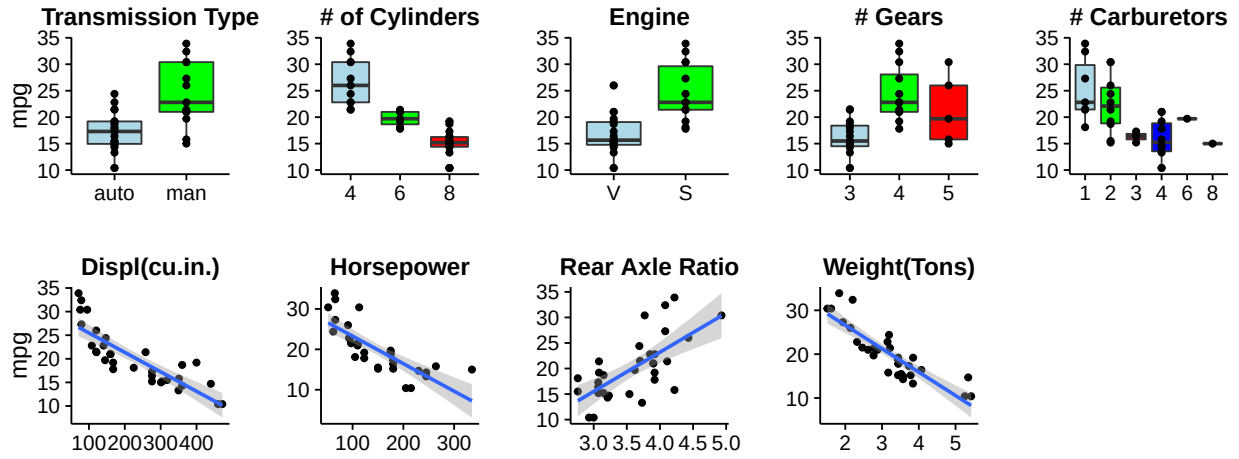
```
t.test(mpg ~ am, data = mtcars)$estimate; t.test(mpg ~ am, data = mtcars)$p.value
```

```
## mean in group auto  mean in group man
##           17.14737           24.39231
```

```
## [1] 0.001373638
```

The t-test confirms that there is a difference in effects by transmission type, as shown by the marked difference in the mean mpg between automatic and manual transmissions (17.1 vs 24.4, a difference of 7.3 mpg). The low p-value of 0.001374 rejects the null hypothesis that transmission has no effect.

Let's visually examine mpg in relation to other potential regressors that may may also impact mpg.



We have a visual confirmation between the mpg results of "transmission type", but there appears to be correlations with other variables as well. A multiple regression model appears be better suited for the data.

Regression Models

Our base linear regression model for comparison will use transmission type as a predictor of mpg.

```
basemodel <- lm(mpg ~ factor(am),data=mtcars)
summary(basemodel)$coef;summary(basemodel)$adj.r.squared
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   17.147368   1.124603 15.247492 1.133983e-15
## factor(am)man  7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] 0.3384589
```

The small p-values are statistically significant, supporting the view that transmission type does affect mpg. However, the model yields an Adjusted R-squared value of 0.34, which we can interpret as meaning 34% of the variability in mpg is explained by the transmission type. This is far from optimal, so we will look for a multiple regression model with a better fit. The R step() function uses "AIC" values to compare and establish a "best fit"" model from all possible regressors. The resulting "best fit" model calculated by the step() function adds the regressors cylinder, horsepower, and weight to transmission type (see Appendix for model calculation code).

```
bestfitmodel <- lm(mpg ~ am + cyl +hp +wt,data=mtcars)
summary(bestfitmodel)$coef; summary(bestfitmodel)$adj.r.squared; sqrt(vif(bestfitmodel))
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## amman        1.80921138 1.39630450  1.295714 2.064597e-01
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
```

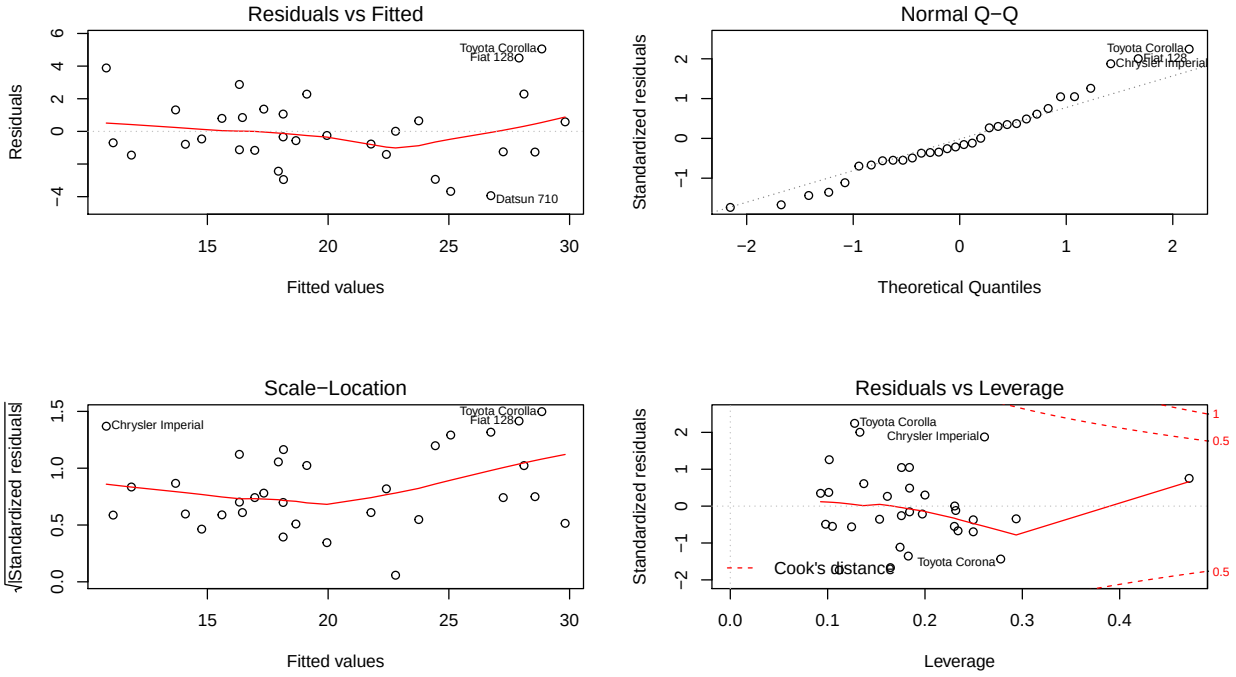
```
## [1] 0.8400875
```

```
##         GVIF       Df GVIF^(1/(2*Df))
## am  1.609589 1.000000        1.268696
## cyl 2.413409 1.414214        1.246401
## hp  2.168784 1.000000        1.472679
## wt  2.001778 1.000000        1.414842
```

This model yields an Adjusted R-Squared value of .84, accounting for 84% of the variability in mpg, a much better fit than our base model. The small p-values support our regressors as being statistically significant. We also find in this model that transmission type now only accounts for an increase of 1.8 mpg. A look at the Variance of Inflation shows that there is some collinearity between regressors.

Diagnostic Plots of Residuals

Let's examine the residual data of this model with diagnostic plots to make sure there are no influential outliers and such.



Residuals appear random with no pattern in the first chart. The QQplot in the upper right indicates the best fit model residuals are normally distributed; however, 3 outliers do have a high degree of leverage, which is concerning considering the small dataset of 32 observations. An ANOVA test of the two models will verify that these models are in fact statistically different enough to merit using the more complex model.

```
anova(basemodel, bestfitmodel)
```

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

The p value is very small, so we reject the null hypothesis that these two models yield similar results. Accordingly, we should use the more complex model that accounts for a greater degree of variability of and correlation to the mpg data.

Conclusion/Uncertainties

We have examined mpg in relation to 1) a simple model of transmission type vs 2) a complex model including regressors transmission type, # of cylinders, horsepower, and weight. After confirming that manual transmission does affect mpg by approximately 7.3 mpg, a basic visual examination of correlations seemed to show that other regressors are also in play. An AIC analysis yielded a better fitting model, confirming our visual assessemnt. The simple model accounted for 34% of the variability in mpg, but a best fit model including additional regressors accounts for 84% of the mpg variability. Having checked residuals for influential outliers, we found 3 data points with a high degree of leverage that could skew the results due to a small data set; a larger data set would be preferable! Nonetheless, an ANOVA test confirmed that our two models are statistically different enough to merit choosing the multiple regression model. Apparently, transmission type is not as significant a factor in determining mpg when other factors like # cylinders, horsepower and weight are taken into account. Finally, we must take note that in adding regressors to our model, the variance of inflation did increase, but is worth the increase in R squared values.

Appendix

Model Selection

The following code displays the AIC method of finding the best fit model from all the data provided. Best fitting model will be that with the lowest AIC score (last model with AIC=61.65)

```
allmodel <- lm(mpg ~ .,data=mtcars)
idealmodel <- step(allmodel, direction = "both")  ##Select best model
```