Regression Models Project

Evan Oman

09/24/2015

Executive Summary

The goal of this project is to answer the following questions about the `mtcars` dataset:

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

The result of our analysis of the mtcars dataset seems to indicate that, in general, automatic transmissions are inferior in terms of miles per gallon when compatred with automatic transmissions. When we performed a linear regression model with just transmission type as an explanatory variable, a change from automatic to manual transmission increased the mpg by 7.245 however, transmission type only explained 36% of the variation in mpg. A linear regression model of all significant variables (determined by ANOVA), explained 84% of the variation in mpg. Transmission type was determined to be an insignificant contributory variable to the model.

Report

The Data

We are going to attempt to explain the relationship between miles per gallon (`mpg`) and a set of explanatory variables outlined below:

1. `cyl` : number of cylinders (factor, 4,6,8)
2. `disp` : displacement (cu.in.) (numerical)
3. `hp` : gross horsepower (numerical)
4. `drat` : rear axle ratio (numerical)
5. `wt` : weight (1000 pounds) (numerical)
6. `qsec` : 1/4 mile time (numerical)
7. `vs` : V/S, V-engine or Straight engine (factor, V,S)
8. `am` : transmission type (factor, automatic, manual)
9. `gear` : number of forwards gears (factor, 3,4,5)
10. `carb` : number of carburetors (factor, 1,2,3,4,5,6,7,8)

Data Processing

Here we load the data set and convert the appropriate variables to 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"))
auto = subset(mtcars, am == "Automatic")
man = subset(mtcars, am == "Manual")
```

In order to determine which of the variables above are playing a significant role, we will perform an ANOVA:

```
analysis <- aov(mpg ~ ., data = mtcars)
summary(analysis)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will choose all the variables which have a p value less than `.05` giving us the model `mpg ~ am + cyl + disp + wt .`

Exploratory data analysis

See Appendix A for the plots performed for this section.

Building the Model

We can now turn towards building the regression model for this dataset. We will first use multiple linear regression and the R “step” function, which chooses the optimal model by AIC (Aikake Information Criterion). With mpg as the outcome variable and all other variables, except 1/4 mile, as the explanatory variables we arrive at the model:

```
## (Intercept)        cyl6        cyl8          hp          wt    amManual
##      33.708      -3.031      -2.164      -0.032      -2.497       1.809
```

and the coefficients’ corresponding p-values:

```
## (Intercept)        cyl6        cyl8          hp          wt    amManual
##     0.00000     0.04068     0.35225     0.02693     0.00908     0.20646
```

The model here is: mpg ~ cyl + hp + wt + am. The r^2 for this model is 0.866 which means that this model explains 86.6% of the variation in mpg. The model, arrived at by AIC, includes the variable am, or transmission type. The coefficient for am is 1.809, which we can interepret as, when other variables are held constant, a change from automatic to manual transmission will increase the mpg by 1.809. However, we must note the p-value associated with the transmission type variable. It is 0.206 which is well above our usual 0.05 significant level.

If we remove the transmission type from this model, and refit it we obtain a model mpg ~ cyl + hp + wt. The coefficients and corresponding p-values are:

```
## (Intercept)        cyl6        cyl8          hp          wt
##      35.846      -3.359      -3.186      -0.023      -3.181
```

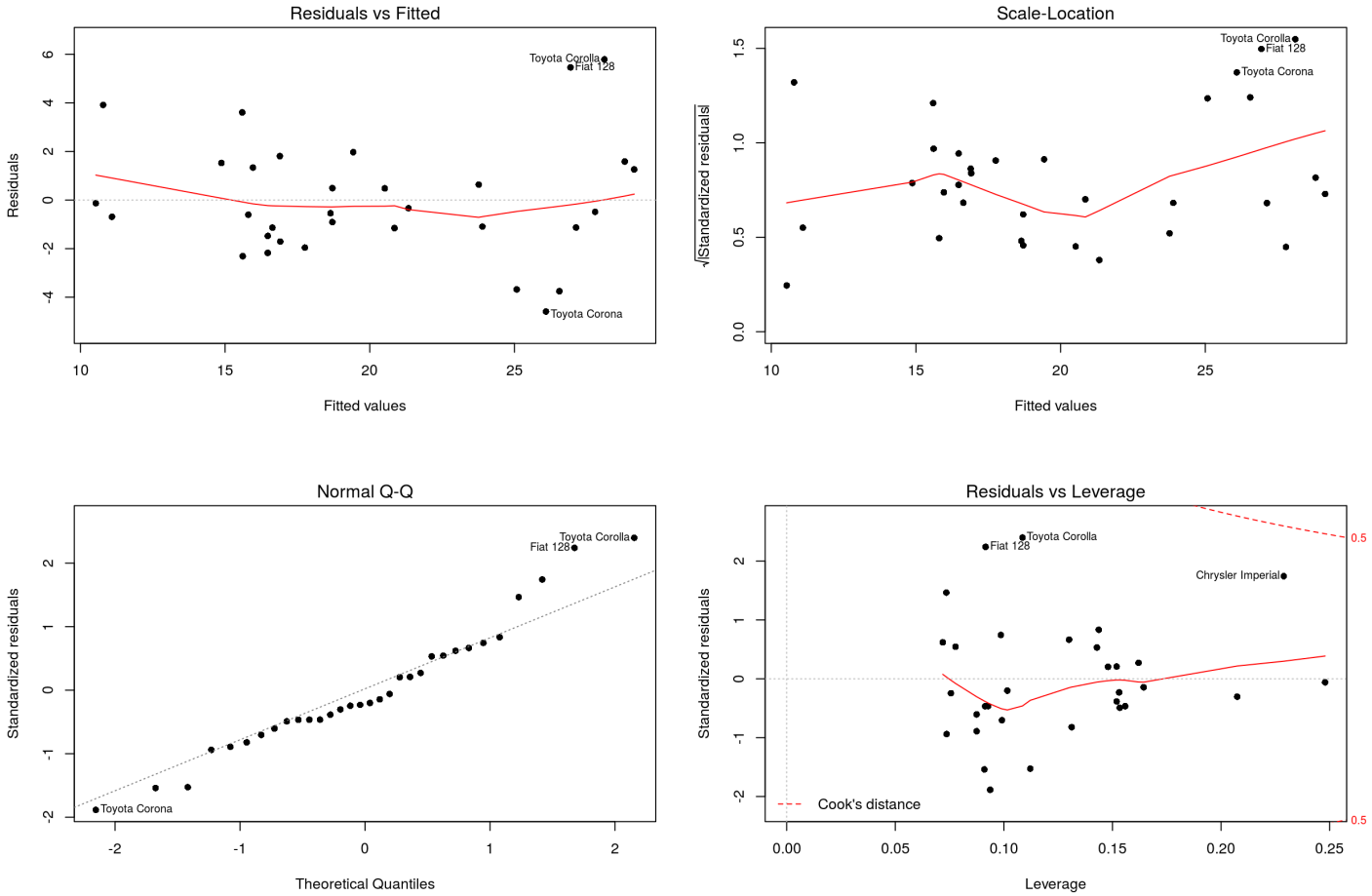
```
## (Intercept)        cyl6        cyl8          hp          wt
##     0.00000     0.02375     0.15370     0.06361     0.00014
```

An ANOVA test between this model and this previous results in a p-value of 0.206 which means that we should choose the simpler model. However, we can further simplify the model by removing the hp or horsepower variable since the p-value for its coefficient is above 0.05. This results in the model mpg ~ cyl + wt. An ANOVA test between this model and previous two models resulted in p-values of 0.064 and 0.082, respectively. Both are above 0.05 so we can choose the simpler model.
The final model is:

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.9908     1.8878  18.006  < 2e-16 ***
## cyl6         -4.2556     1.3861  -3.070 0.004718 **
## cyl8         -6.0709     1.6523  -3.674 0.000999 ***
## wt           -3.2056     0.7539  -4.252 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

Here, all of the coefficients are highly significant. This model has an r^2 of 0.837 which means that it explains 83.7% of the variation in mpg. This is very close to our original model with 2 more variables. Note that transmission type is not included in the final model.

The plots below are the diagnostic plots for the model. There do not appear to be any problems with these plots; the residuals appear randomly, the standardized residuals appear normally distributed, and there are not any highly influential outliers.



Model with only Transmission Type

The project asks us about two specific points:

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

We have seen in the explaoratory data analysis section that manual transmissions are generally better for mpg. To quantify the difference, a linear model was fit using only mpg as an explanatory variable. This model produced an 0.36 which means that it explains 36% of the variation in mpg, much less than the final model arrived at above. The coefficient is significant, and at 7.245, which we can interepret as, automatic to manual transmission will increase the mpg by 7.245. Sounds like transmission type does have an impact on mpg, however when the previous models are considered, other variables are much more impartant than transmission type in explaining the variation in mpg.

Conclusion

We have determined that there is a difference in mpg in relation to transmission type and have quantified that difference. However, transmission type does not appear to be a very good explanatory variable for mpg; weight, horsepower, and number of cylinders are all more significant variables.

Appendix A Here we perform some basic exploratory data analysis which seems to indicate that manual transmissions have higher MPG:

```
boxplot(auto$mpg,man$mpg,varwidth=TRUE,xlab="transmission type",ylab="mpg",main="Comparison of MPG of Automatic vs. Manual Transmissi
on",names=c("automatic","manual"))
```