Regression Models Project: Motor Trend MPG Analysis

Telvis Calhoun

February 26, 2016

Overview

In this project, we explore the `Motor Trend Car Road Tests` (mtcars) dataset. We'll analyze this dataset to answer the following questions.

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

Exploratory Analysis

First, lets load libraries and datasets used in the analysis.

```r
library(datasets)
library(ggplot2)
library(dplyr)
data("mtcars")
```

A quick summary of the data shows `mtcars` dataset 11 variables. For this analysis, we will investigate the Miles/(US) gallon `mpg` as a function of the Transmission type `am`.

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Let's create a factor variable called `am_factor` that will show the strings 'automatic' where `am == 1` and 'manual' where `am == 0`.

```
mtcars <- mutate(mtcars, am=factor(ifelse(am==1, 'automatic', 'manual')))
table(mtcars$am)
```

```
##
## automatic    manual
##        13        19
```

Model Selection

Before we can characterize the relationship between `mpg` and the transmission type (`am`), we must first search for other variables that can distort, or confound the relationship between the variables. First we will "get our hands dirty" and quantify the percentage change in the `am` coefficient when we adjust for all other variables.

```
# baseline
a <- summary(lm(mpg ~ am, data=mtcars))$coef[2]

# calc percent change in the baseline for each covariate
rbind(
  c('baseline', 'cyl', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'gear', 'carb'),
  round(c((a-a)/a,
    100 * (a - summary(lm(mpg ~ am + cyl, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + disp, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + hp, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + drat, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + wt, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + qsec, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + vs, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + gear, data=mtcars))$coef[2])/a,
    100 * (a - summary(lm(mpg ~ am + carb, data=mtcars))$coef[2])/a
  ), digits=2)
)
```

```
##      [,1]       [,2]    [,3]    [,4]    [,5]    [,6]     [,7]     [,8]
## [1,] "baseline" "cyl"   "disp"  "hp"    "drat"  "wt"     "qsec"   "vs"
## [2,] "0"        "64.57" "74.69" "27.16" "61.25" "100.33" "-22.52" "16.26"
##      [,9]   [,10]
## [1,] "gear" "carb"
## [2,] "1.43" "-5.63"
```

The output shows that `cyl`, `disp`, `hp`, `drat`, `wt` all change the `mpg` coeffient by greater than +/- 25%. Now let's generate nested models for these 5 variables using and evaluate them using a nested likelihood ratio tests.

```
fit1 <- lm(mpg ~ am, data=mtcars)
fit2 <- lm(mpg ~ am + cyl, data=mtcars)
fit3 <- lm(mpg ~ am + cyl + disp, data=mtcars)
fit4 <- lm(mpg ~ am + cyl + disp + hp, data=mtcars)
fit5 <- lm(mpg ~ am + cyl + disp + hp + drat, data=mtcars)
fit6 <- lm(mpg ~ am + cyl + disp + hp + drat + wt, data=mtcars)
anova(fit1, fit2, fit3, fit4, fit5, fit6)
```

```
## 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
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 69.1874 1.146e-08 ***
## 3     28 252.08  1     19.28  2.9675  0.097302 .
## 4     27 216.37  1     35.71  5.4967  0.027298 *
## 5     26 214.50  1      1.87  0.2879  0.596326
## 6     25 162.43  1     52.06  8.0130  0.009033 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that models 2, 4 and 6 have the greates significance. This suggests that `disp` and `drat` should be excluded from the model. The comparison below that good model lowest p-value with the variables: `am`, `cyl`, `hp` and `wt`.

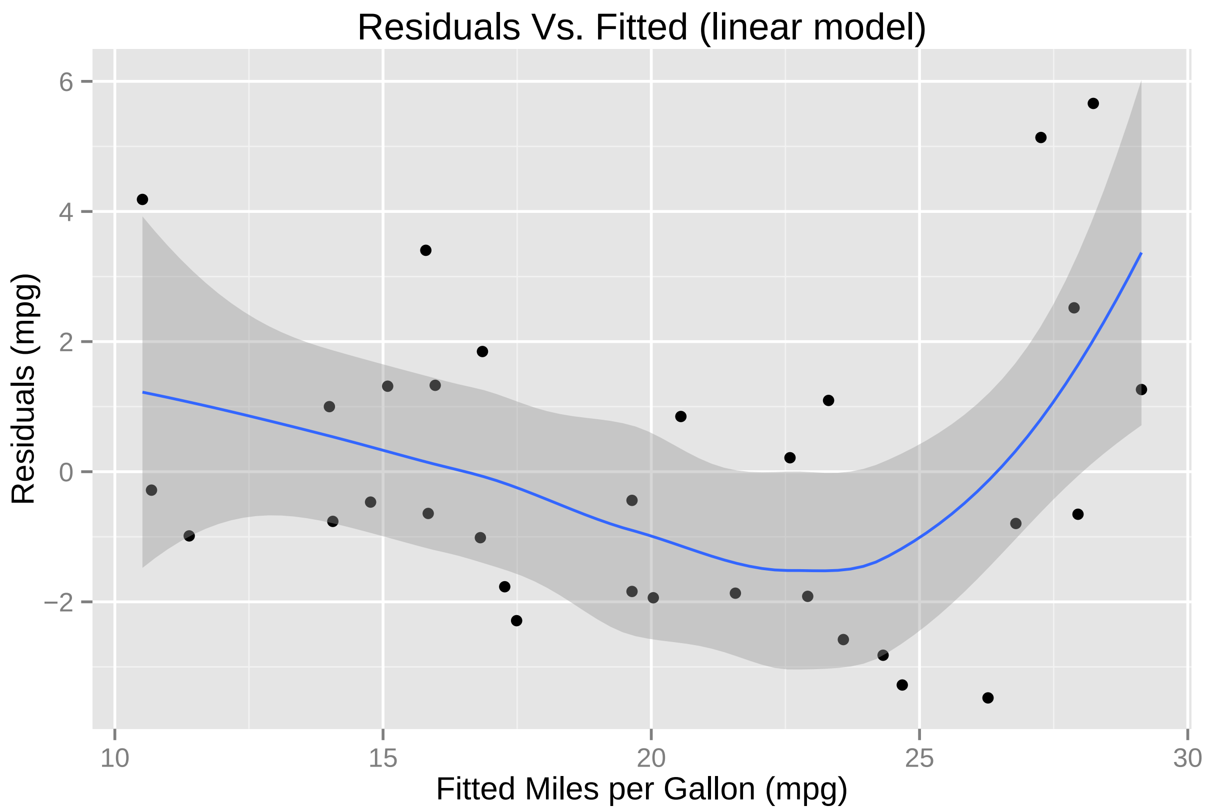
```
fit1 <- lm(mpg ~ am, data=mtcars)
fit7 <- lm(mpg ~ am + cyl + hp + wt, data=mtcars)
```

The `r.squared` value for this model indicates that 83% of the total variablity is explained by the linear relationship between the `mpg` and `am`, `cyl`, `hp` and `wt`.

```
summary(fit7)$adj.r.squared
```

```
## [1] 0.8266657
```

Finally, we plot the residuals to search for a pattern in the residuals vs the fitted (yhat) values. The plot shows a slight curve in the values - but nothing too bad.



Automatic Vs. Manual Comparison

Now that we have a model containing the necessary covariates, let's calculate the change in `mpg` for transmission type after adjusting for Number of Cylinders (`cyl`), Gross Horsepower (`hp`) and weight (`wt`). The results show that the `mpg` for a manual transmission decreases by `1.48 mpg` holding other variables constant.

```
summary(fit_lm)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.62458346 2.09640689 17.947176 1.556106e-16
## ammanual    -1.47804771 1.44114927 -1.025603 3.141799e-01
## cyl         -0.74515702 0.58278741 -1.278609 2.119166e-01
## hp          -0.02495106 0.01364614 -1.828433 7.855337e-02
## wt          -2.60648071 0.91983749 -2.833632 8.603218e-03
```

However, the confidence interval for `am` ranges from `-4.4` to `1.47`. Because it includes 0, we cannot reject the null hypothesis after adjusting for Number of Cylinders (`cyl`), Gross Horsepower (`hp`) and weight (`wt`).

```
confint(fit_lm)
```

```
##                   2.5 %       97.5 %
## (Intercept) 33.32311183 41.926055080
## ammanual    -4.43504176  1.478946352
## cyl         -1.94093802  0.450623969
## hp          -0.05295064  0.003048517
## wt          -4.49383134 -0.719130075
```

Conclusion

We show that we most accurately model the relationship between `MPG` and `transmission type` using a `poisson` generalized linear model. This model has lower residual error than `linear` model. `TODO: FIXME` : The results show that an automatic transmission has `0.0%` greater fuel efficiency than manual transmission. However, the results show that the difference in fuel efficiency decreases by `0.0%` when we adjust for the number of cylinders.