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.

```
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`.

```
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
```

Automatic Vs. Manual Comparison

We can calculate the mean `mpg` for both 'automatic' and 'manual' transmissions by fitting a linear model with a "dummy variable" `am_factor` and dropping the intercept by including a `- 1` in the formula. The `Estimate` column shows the group mean is `24.4 mpg` for 'automatic' transmissions and `17.14 mpg` for 'manual' transmissions. The p-value shows the significance compared to the `0 estimate`.

```
summary(lm(mpg ~ am - 1, mtcars))$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## amautomatic 24.39231   1.359578 17.94109 1.376283e-17
## ammanual    17.14737   1.124603 15.24749 1.133983e-15
```

To sanity check, let's use `dplyr` to calculate the group means to verify our `lm` output. The output shows the means are identical to the `lm` output.

```
summarise(group_by(mtcars, am), mn=mean(mpg))
```

```
## Source: local data frame [2 x 2]
##
##          am       mn
##      (fctr)    (dbl)
## 1 automatic 24.39231
## 2    manual 17.14737
```

We can use a `lm` to calculate the statistical significance of the difference between the group mean mpg for 'automatic' and 'manual' transmission. The model uses the 'automatic' mpg as the intercept - where the intercept is the estimated mean for the reference level. The estimate for 'manual' transmission mpg is `-7.24 mpg` less than the reference level. The 'manual' p-value shows difference in the group mean from the reference is statistically significant ($p < 0.05$).

```
summary(lm(mpg ~ am, mtcars))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 24.392308   1.359578 17.941085 1.376283e-17
## ammanual    -7.244939   1.764422 -4.106127 2.850207e-04
```

The confidence interval is entirely below 0. Therefore we are confident that the automatic transmission reduces the `mpg`.

```
confint(lm(mpg ~ am, mtcars))
```

```
##                 2.5 %   97.5 %
## (Intercept)  21.61568 27.16894
## ammanual    -10.84837 -3.64151
```

Model Selection

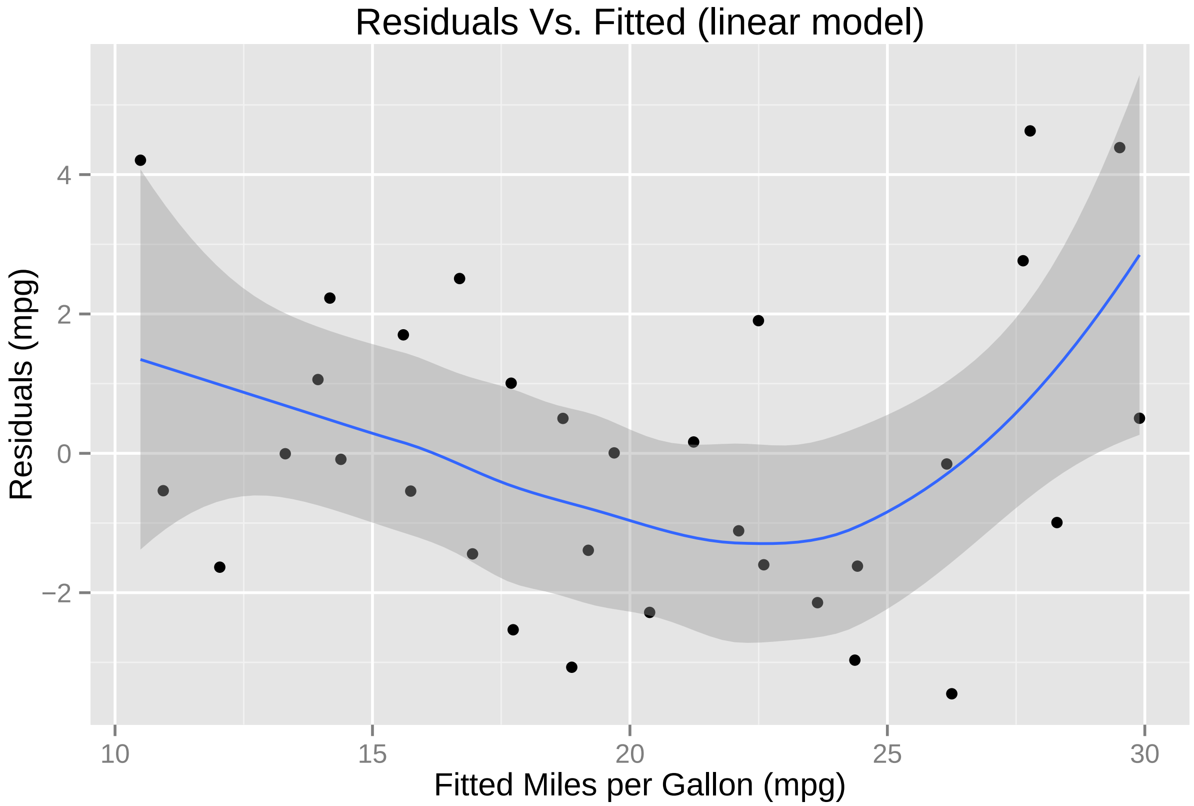
In the previous section, we use a linear model to perform statistical inference. However, a linear model may not be best for modeling the `mtcars` data because the outcome `mpg` is (1) always greater than 0 and (2) potentially unbounded. A poisson model may be better suited for this data. Let's explore this by plotting the residuals for a linear and poisson model.

Linear Regression

Let's fit a linear regression model for mpg ~ all other variables in `mtcars`, calc the residuals and the fitted (yhat) values.

```
fit_lm <- lm(mpg ~ ., data=mtcars)
```

Now let's plot the residuals vs. fitted values.

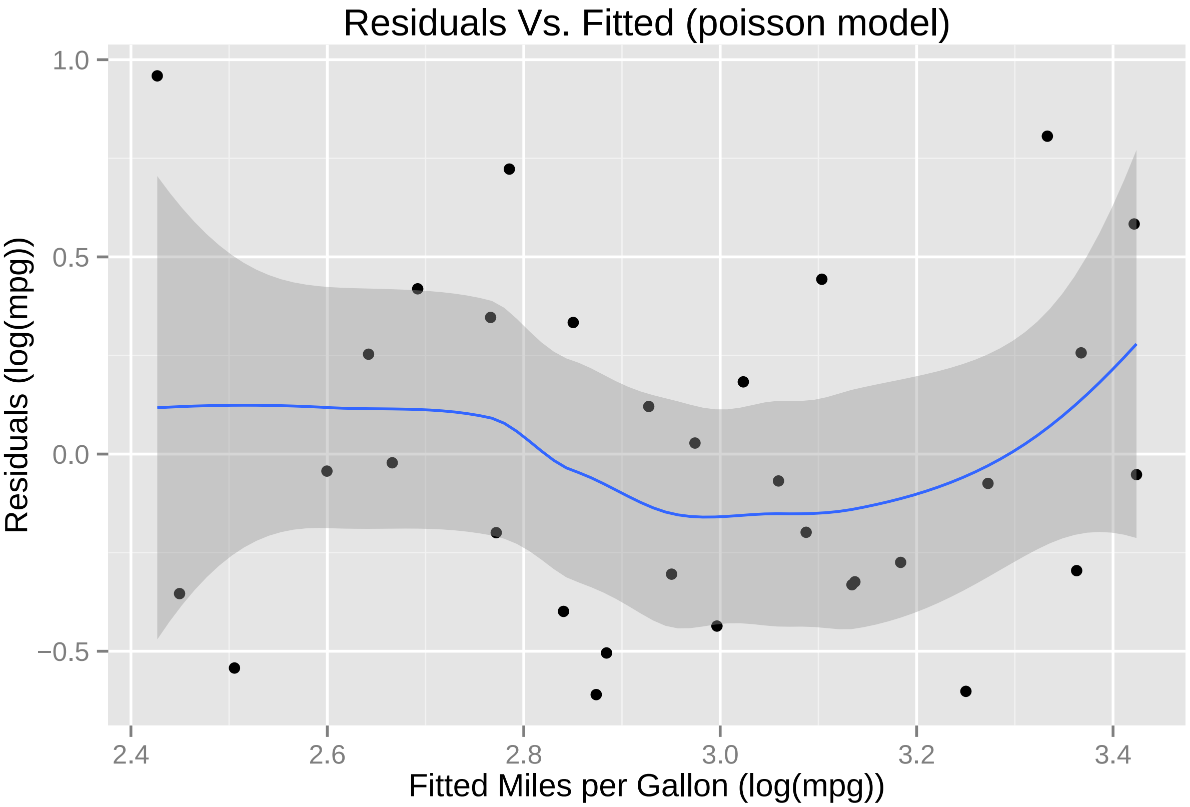


The plot shows a curve in the residuals. The residuals become more negative until they reach ~22.5 mpg then they become more positive. This suggests that the linear model is a poor fit for the `mtcars` data.

```
fit_pois <- glm(formula= mpg ~ ., data=mtcars, family=poisson)
```

Poisson Model

Let's fit a poisson regression model for mpg ~ all other variables in `mtcars`, calc the residuals and the fitted (yhat) values.



The residual plot for the poisson model looks smoother than for the linear model. This is because the output (Y) is now the log(E[Y]) and residuals are in the same log units. Using the log "squashes" the variance in the residuals.

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.