Data Analysis of Motor Trend Car Road Tests

Miguel Couto

Executive summary

This report concerns the final project of the Regression Models course of the Data Science Specialization. We studied the data collected by the 1974 Motor Trend US magazine on the fuel consumption and 10 other aspects of 32 cars; a more in-depth description of the data can be found here. This project aims to understand the influence of the type of transmission (automatic or manual) on the number of miles per galon (mpg):

- Is an automatic or manual transmission better for number of miles per galon?

We fit a multivariable linear model to this data, which predicted that on average manual cars have higher fuel efficiency than automatic cars, travelling 2.9 miles more than automatic cars per gallon of fuel.

Exploratory data analysis

We start by plotting the relationship between number of miles per galon of fuel (mpg) and the following variables: number of cylinders in the car's engine (cyl), the car's weight (wt, in 1000 lbs), time in seconds the car took to travel 1/4 mile (qsec), and the type of transmission (am=0 for automatic, am=1 for manual). All plots can be found in the Appendix.

```r
library(datasets); library(ggplot2); library(dplyr)
mtcars$am <- factor(mtcars$am)

ggplot(mtcars, aes(y=mpg, color=am)) + geom_boxplot() + facet_grid(. ~ cyl) +
    theme(axis.text.x = element_blank()) + ylab('Miles per galon') +
    ggtitle('Miles per galon per transmission type and number of cylinders') +
    scale_color_discrete(name = 'Transmission type', labels = c('automatic', 'manual'))
```

```r
ggplot(mtcars, aes(y=mpg, x=wt, color=am)) + geom_point() + xlab('Weight (1000 lbs)') +
    ylab('Miles per galon') + ggtitle('Miles per galon per transmission type and weight')+
    scale_color_discrete(name = 'Transmission type', labels = c('automatic', 'manual'))
```

```r
ggplot(mtcars, aes(y=mpg, x=qsec, color=am)) + geom_point() +
    xlab('Quarter of mile time (s)') + ylab('Miles per galon') +
    ggtitle('Miles per galon per transmission type and 1/4 mile time') +
    scale_color_discrete(name = 'Transmission type', labels = c('automatic', 'manual'))
```

An initial analysis of these plots seems to suggest a few conjectures:

- Manual cars with 4 or 6 cylinders appear to have more miles per galon than automatic cars;
- As the car's weight increases, the number of miles per galon seems to decrease;
- The number of miles per galon seems to increase with the quarter of mile time.

Regression models

We now use a few multivariable linear regression models to ascertain the veracity of these conjectures and answer the question we posed at the beginning.

The data set has 10 variables besides mpg, and there are many linear regression models we could perform. Therefore, we decide on the best linear model according to an R feature that selects the model with the lowest Akaike information criterion (see more here).

```
step(lm(mpg ~ . , data = mtcars))
```

This concluded that the linear regression model that best fits this data is: **mpg** ~ **wt** + **qsec** + **am**. Let's look further into this linear model.

```
bestfit <- lm(mpg ~ wt + qsec + am, data=mtcars)
summary(bestfit)$coef
```

Interpretation. these coefficients' predictions are consistent with the conjectures above:

- the predicted difference in average miles per galon between manual and automatic transmission, holding weight and quarter mile time constant, is 2.935837;
- the average miles per galon will decrease 3.916504 per each 1000 lbs increase in weight, holding quarter mile time and transmission type constant;
- the average miles per galon will increase 1.225886 per 1 second increase in the quarter mile time, holding weight and transmission type constant.

```
t.test(mpg~am,mu=2.9,data=mtcars,paired=FALSE,var.equal=FALSE,alternative='less')$p.value
```

```
## [1] 2.42214e-05
```

This confirms that on average manual cars travel 2.9 miles more than automatic cars per gallon of fuel.

We study the fit of this model to the data by plotting its residuals, dfbeta values (which measure the difference in the model coefficients when including and excluding each data point) and hat values (which measure the potential for influence of each data point). The plots can be found in the Appendix.

```
plot(resid(bestfit), pch=21, col='black', bg='lightblue', frame=FALSE,
     xlab='cars', ylab='residuals', main='Residual plot')
abline(h=0, lwd=2)
```

```
dfbetaMat <- dfbetas(bestfit); title <- c('intercept', 'weight', 'qsec', 'manual')
par(mfrow=c(2,2))
for (i in 1:4){ plot(dfbetaMat[,i], xlab='cars', ylab='dfbeta values',
                     main = paste('dfbeta values for', title[i] , 'term')) }
```

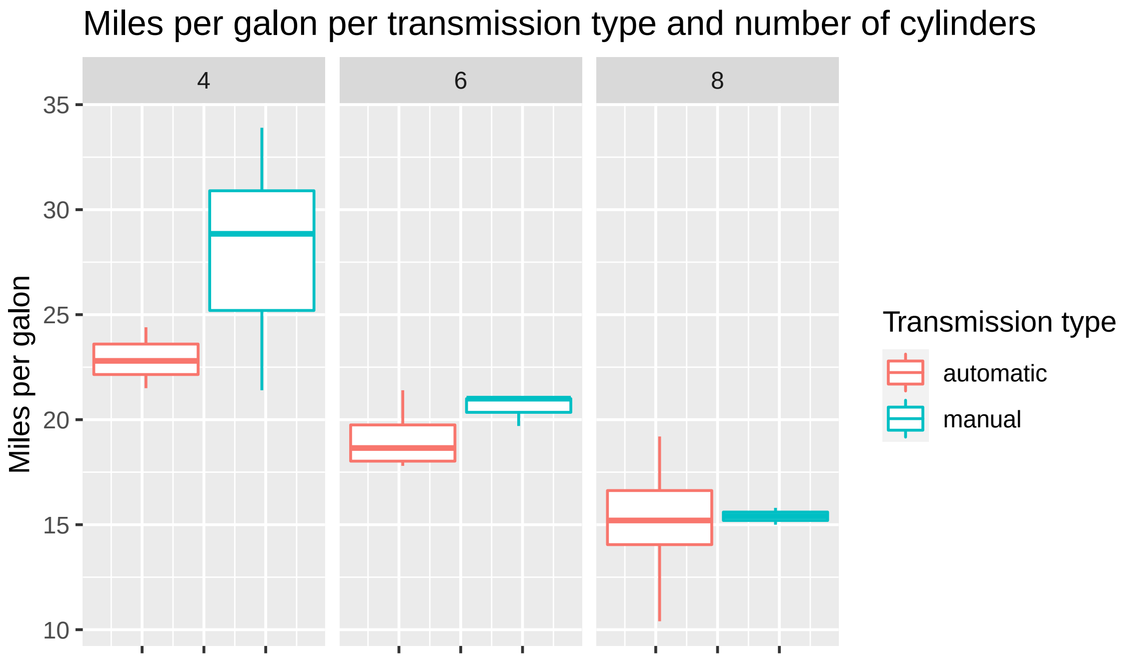
```
par(mfrow=c(1,1))
plot(hatvalues(bestfit), pch=21, col='black', bg='lightblue', frame=FALSE,
     xlab='cars', ylab='hat values', main='hat values plot')
```

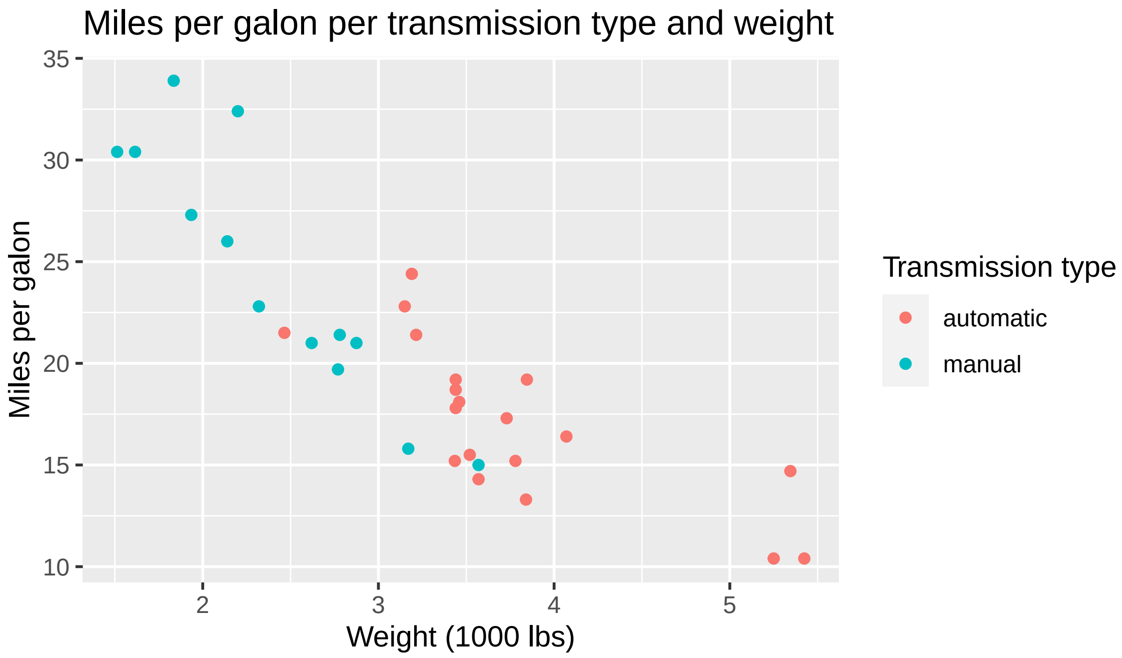
The residual plot shows no discernable pattern, as we hoped it would, and the dfbeta and hat values are all quite low, hence these show no problem with the model.

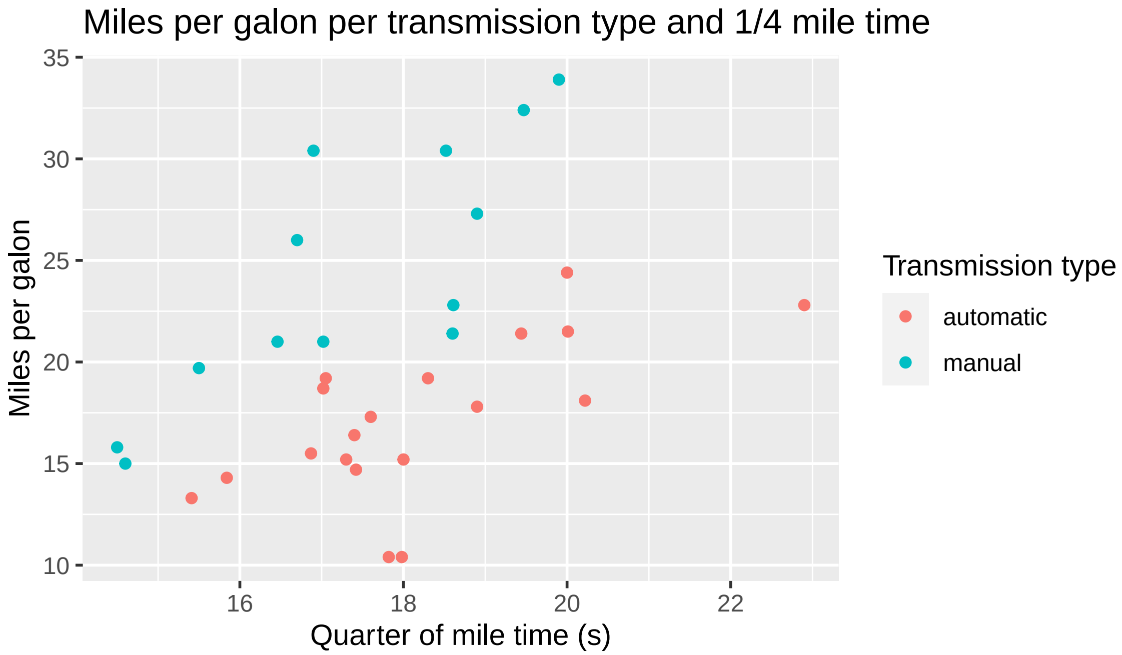
Conclusion: Our model predicts that manual cars have higher fuel efficiency than automatic cars, travelling 2.9 miles more per gallon of fuel (for a fixed weight and quarter mile time).

Appendix

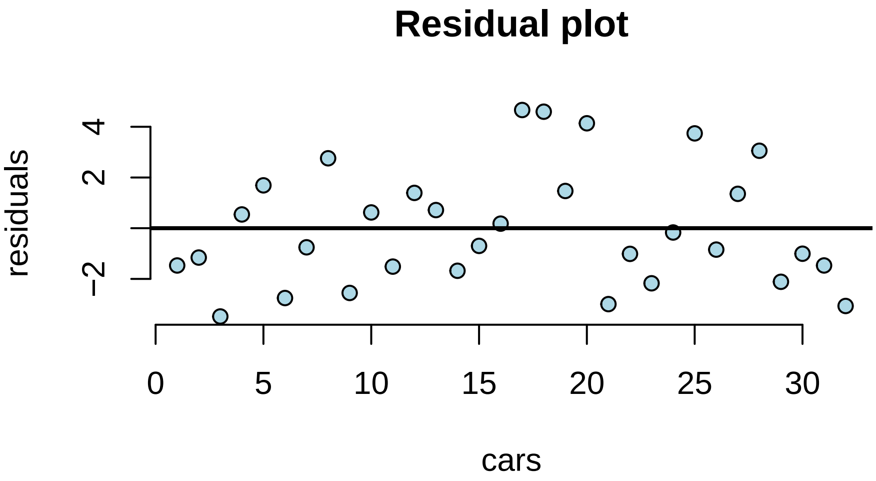
Exploratory data analyses plots:


Miles per galon per transmission type and number of cylinders

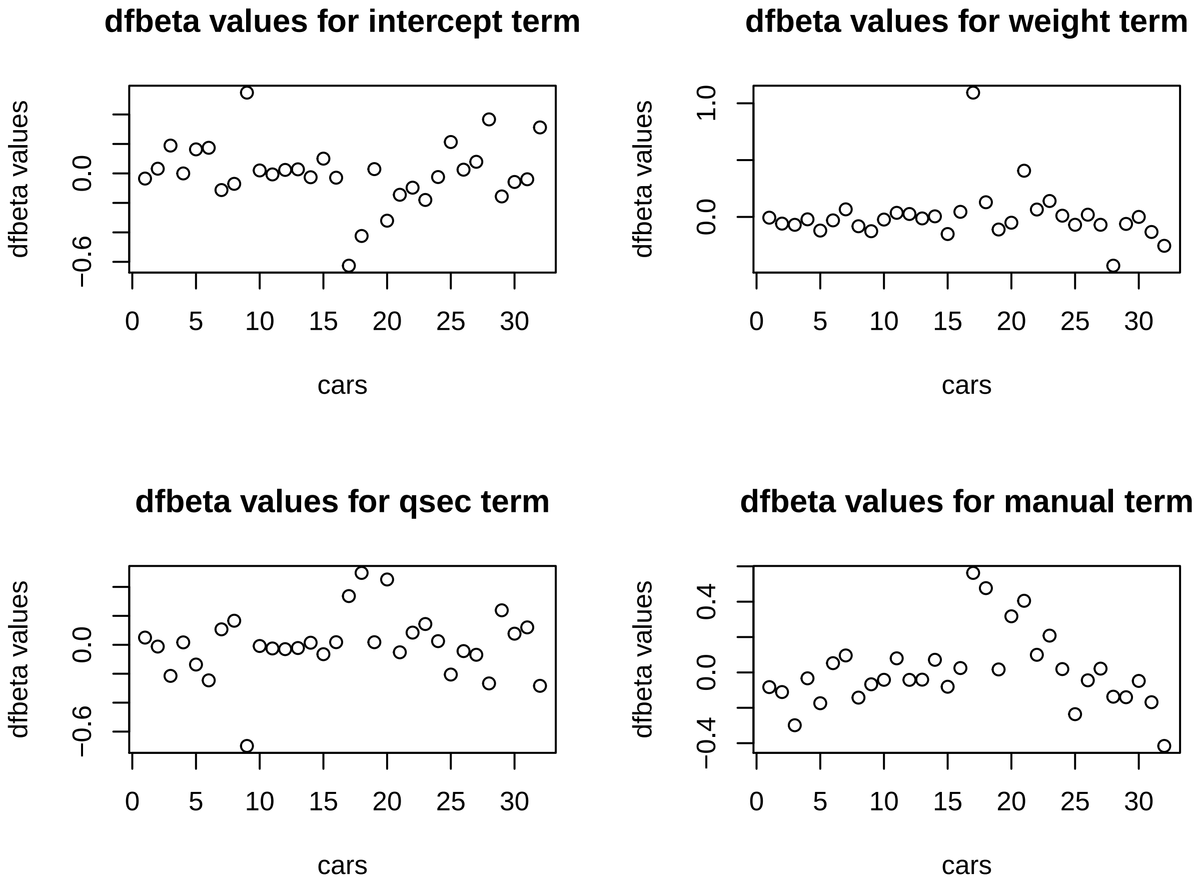

Miles per galon per transmission type and weight



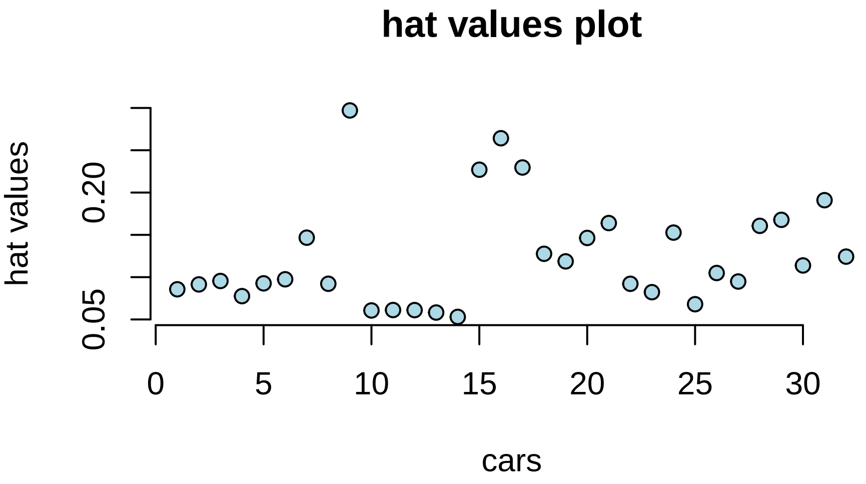
Residual plot:



dfbeta values plot:



hat values plot:


hat values plot
hat values
cars