Regression Models Course Project

Muralidhar Areti

August 21, 2015

Executive Summary

This report utilises the `mtcars` dataset to explore the the impact of various variables on the quarter mile time of a car. The `mtcars` dataset was extracted from the 1974 *Motor Trend* US magazine providing data based on 10 variables representing properties of car design and performance for 32 cars. Exploratory data analysis is used to understand at a glance the effects of number of cylinders, horsepower and weight have on performance (lower quarter mile times are better) and efficiency (higher miles per gallon are better).

Exploratory Data Analysis

Initial exploratory data analysis was used to ensure we had a good understanding of our dataset. The main packages loaded are the `dplyr` and `ggplot2` to assist us with processing and charting the data.

The first step of the exploratory data analysis is to preview the data present. Normally we would also look for the dimensions and summary, but those details are given from the `mtcars` dataset itself (in `R` you can run the `?mtcars` command for more details on the data set).

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

A summary of the results from the exploratory analysis (see Appendix) is provided below:

Property	Performance (qsec)	Efficiency (mpg)
More cylinders	Good	Bad
More horsepower	Good	Bad
More Weight	-	Bad

Statistical Inference

From the exploratory analysis, we make the following null hypothesis for our statistical inference: Best performance can be obtained by providing more horsepower. We can use a T-test to obtain the p-value which dictate whether or not we reject our null hypothesis.

```
result = t.test(mtcars$hp, mtcars$qsec)
result$p.value
```

```
## [1] 7.244975e-12
```

```
result$estimate
```

```
## mean of x mean of y
## 146.68750  17.84875
```

Regression Analysis

```
full_model = lm(qsec ~ ., data = mtcars)
summary(full_model)
```

The code above gives a residual standard error of 0.7685 on 21 degrees of freedom with an adjusted R-squared of 0.815. This tells us the model can explain for about 82% of variance in the performance.

```
step_model = step(full_model, k=log(nrow(mtcars)))
summary(step_model)
```

The model selected is `qsec ~ disp + wt + vs + carb` with a residual standard error of 0.7588 on 27 degrees of freedom, an adjusted r-squared value of 0.8197 telling us that 82% of the variance of the performance.

```
dispWtVsCarb_model = lm(qsec ~ disp + wt + vs + carb, data=mtcars)
summary(dispWtVsCarb_model)
```

Residual Analysis and Diagnosis

See Appendix for the residual plots. From the plots we can verify the following assumptions:

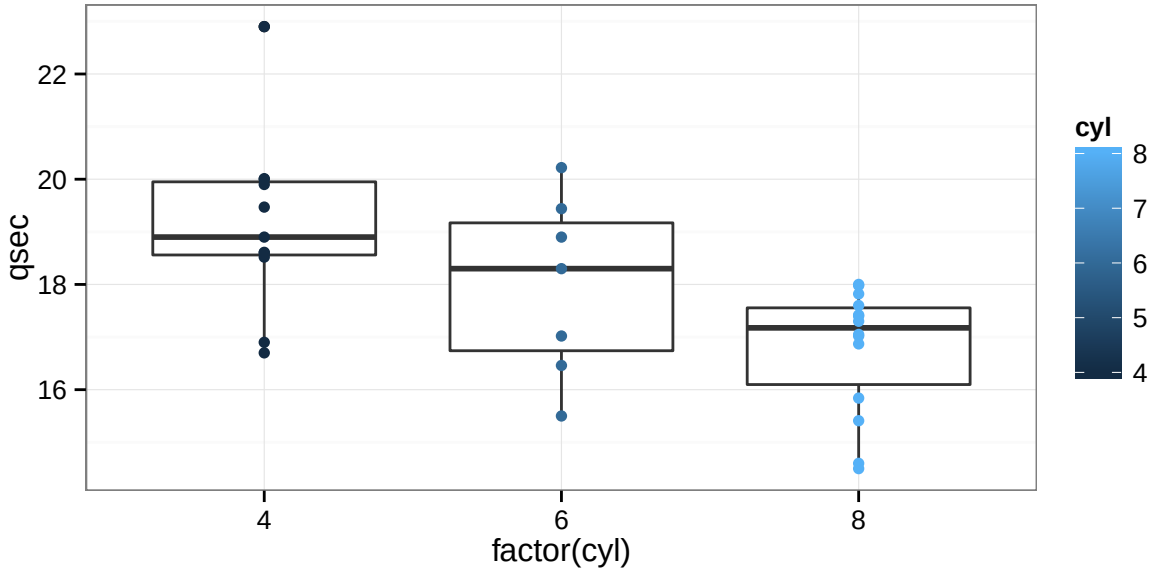
1. The residuals vs fitted plot shows a consistent pattern
2. The normal Q-Q plot show the residuals are normally distributed (this is verified from the points being close to the line)
3. The Scale-Location plot confirms a constant variance assumption (because points are randomly distributed)
4. The residuals vs. leverage shows that we have no outliers as every data point falls within the 0.5 bands.

Conclusion

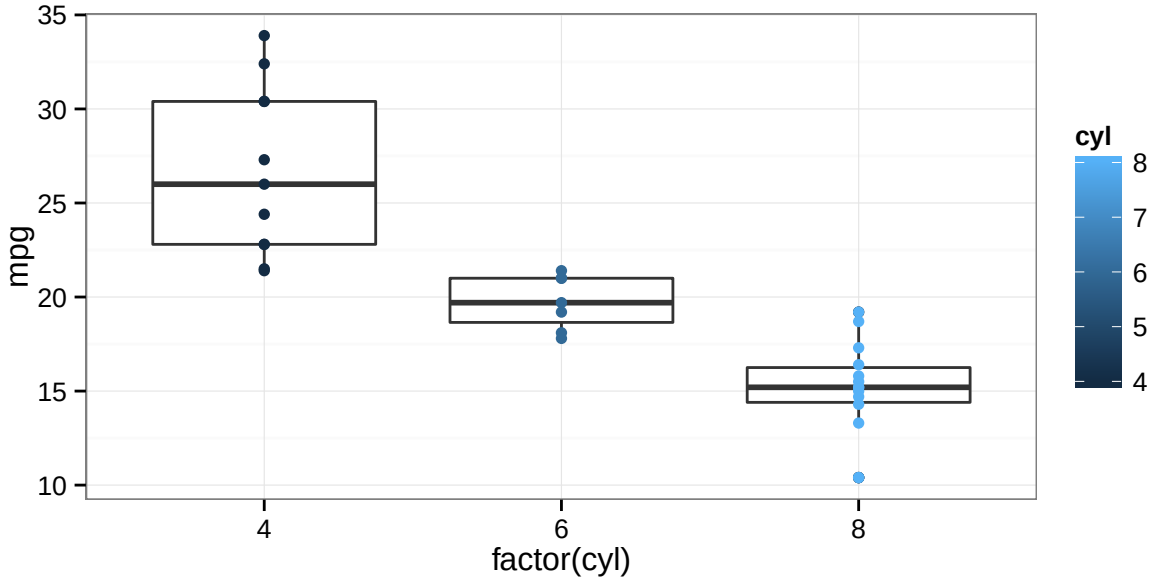
From the above analyses we can conclude that our basic assumptions of linear regression have been met.

Appendix

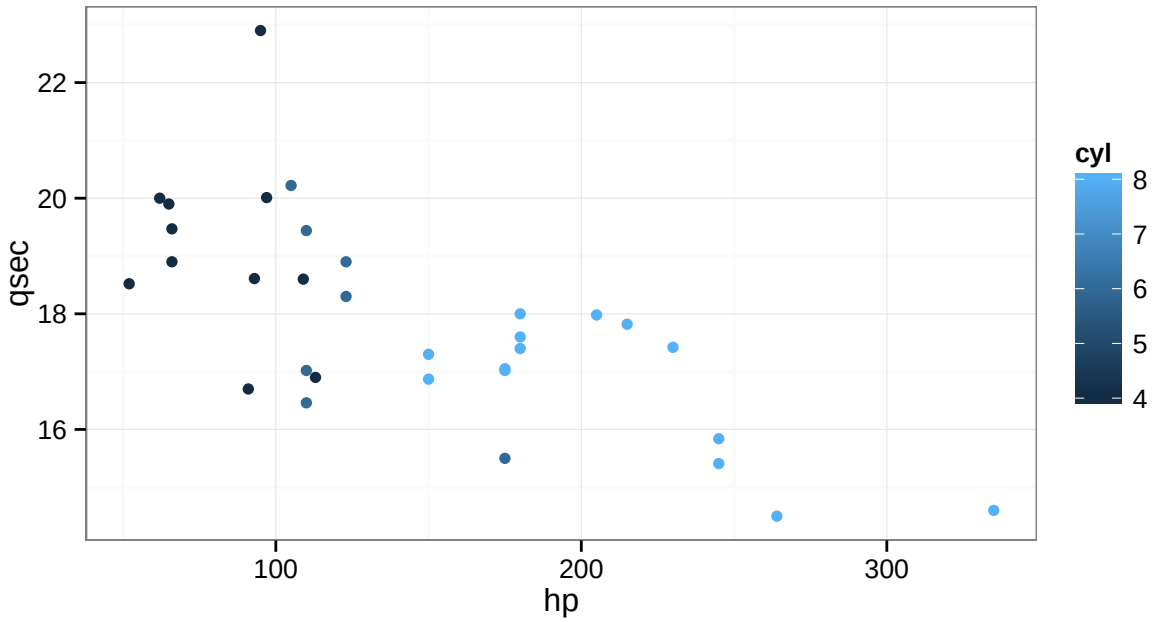
```
# figure
ggplot(mtcars, aes(factor(cyl), qsec)) +
    geom_boxplot() +
    geom_point(aes(colour=cyl)) +
    theme_bw()
```



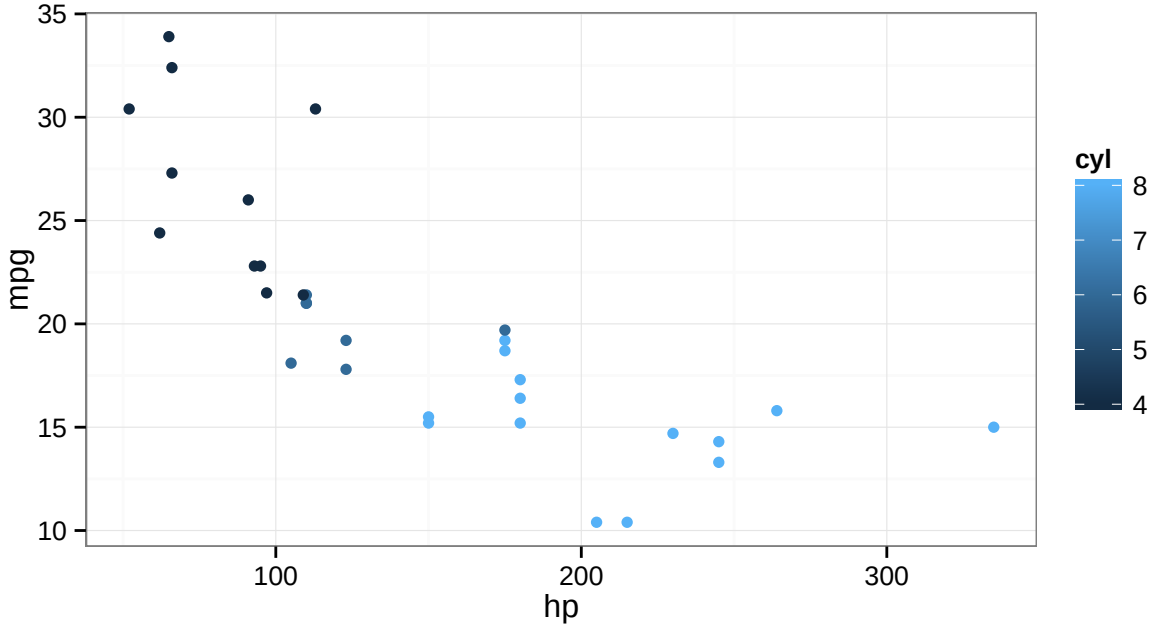

```
# figure
ggplot(mtcars, aes(factor(cyl), mpg)) +
    geom_boxplot() +
    geom_point(aes(colour=cyl)) +
    theme_bw()
```



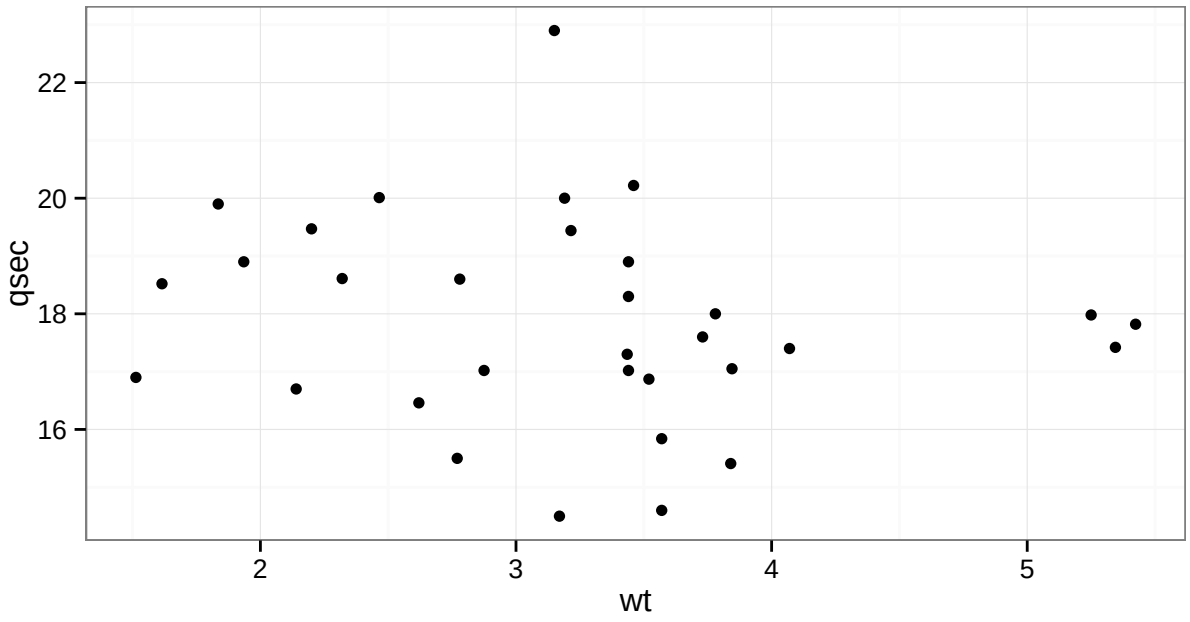
```
# figure
ggplot(mtcars, aes(hp, qsec)) +
    geom_point(aes(colour=cyl)) +
    theme_bw()
```



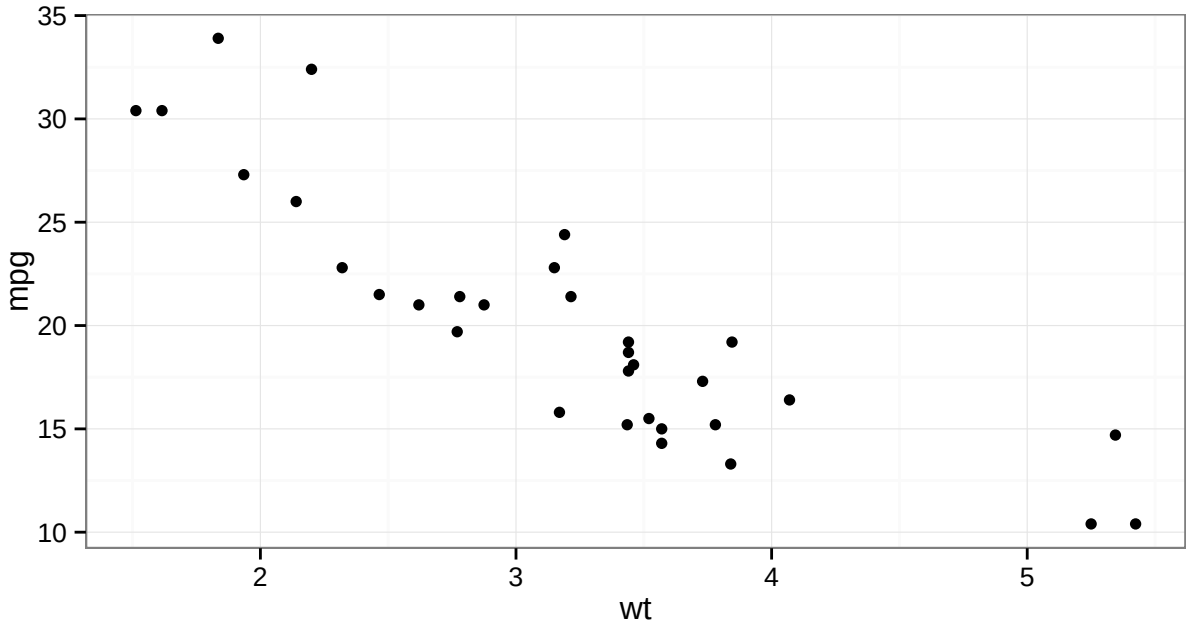

```
# figure
ggplot(mtcars, aes(hp, mpg)) +
    geom_point(aes(colour=cyl)) +
    theme_bw()
```



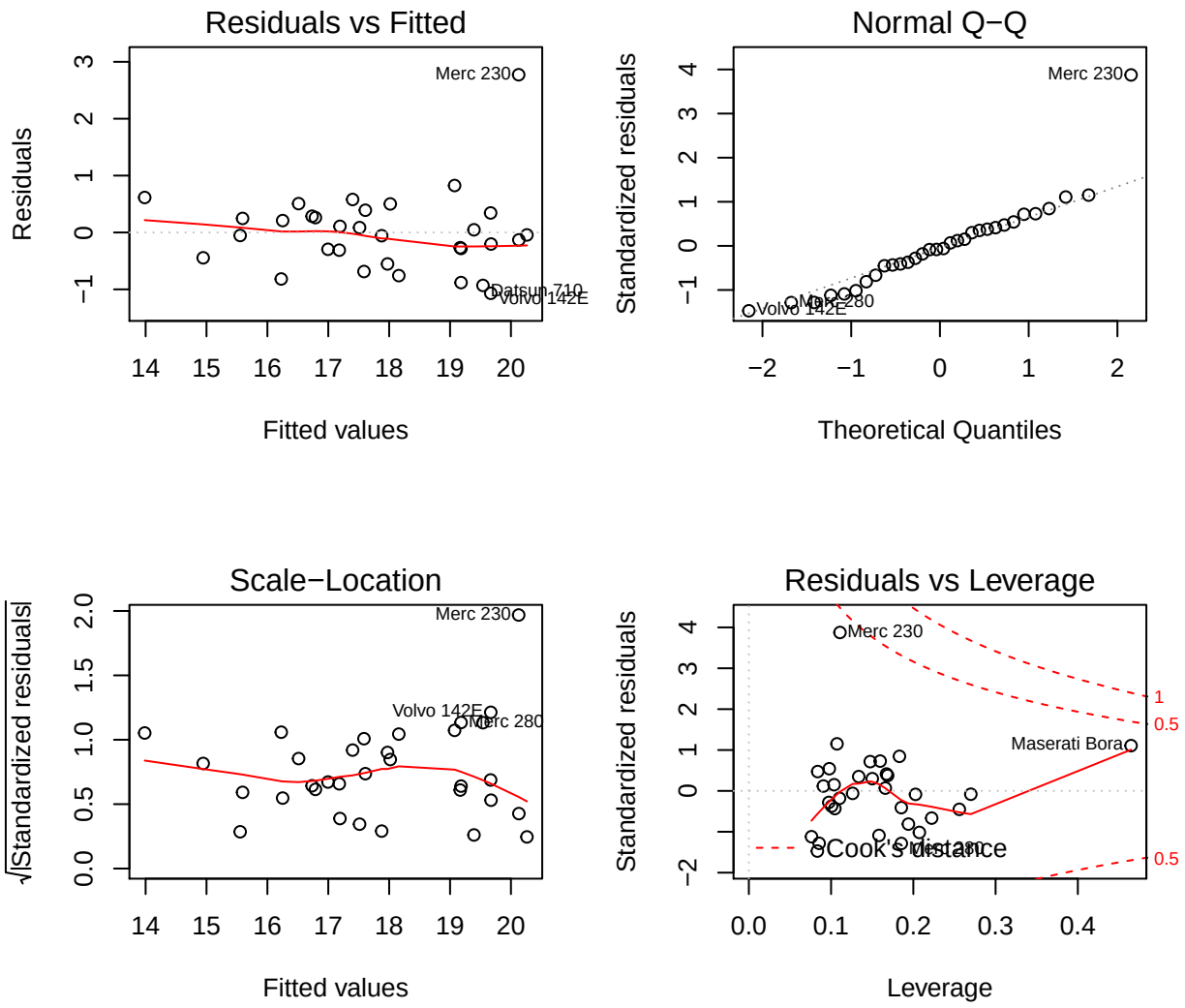
```
# figure
ggplot(mtcars, aes(wt, qsec)) +
    geom_point() +
    theme_bw()
```




```
# figure
ggplot(mtcars, aes(wt, mpg)) +
    geom_point() +
    theme_bw()
```



```
par(mfrow=c(2,2))
plot(dispWtVsCarb_model)
```



```
ggpairs(mtcars)
```