Car Data Analysis

Jacob Townson

July 21, 2015

Overview

In this report, I will analyze the data of the mtcars data set. The goal of this analysis is to use linear modeling with R to answer whether manual transmission is better for gas mileage or if automatic is, and also to try to quantify the difference between the two types of transmission. Before I begin, I must load in the mtcars data.

```
data("mtcars")
```

Making the Model

In order to answer the questions given, we must find the best model. To start with the modeling process, let's look at the model with all of the given variables in the mtcars data set. As seen in the summary given in **Figure 1**, the model with all the variables is not very precise when it comes to P values. In **Figure 1** you can also see the residuals for the model. In the residuals, the Q-Q plot looks to be close to the line, however the line of the residuals vs fitted does not seem to be very good. The line in this graph should be horizontal. So there need to be some changes made to this current model. To find out what, we should see if there are any correlations between the variables.

In **Figure 2** I present the code to find if there are any correlations in the data. If this code is run in R (and drastically zoomed in), one can see that the need for some quadratic variables my be required in order to make the model work the way we want it to. For example, if one were to look at the comparison between gear and mpg, one might notice that the correlation almost looks quadratic. This happens to a few of our variables in the model, so we will add their square terms into the mix.

```
dispsq <- (mtcars$disp)^2
hpsq <- (mtcars$hp)^2
wtsq <- (mtcars$wt)^2
gearsq <- (mtcars$gear)^2
carsDat <- cbind(mtcars, dispsq, hpsq, wtsq, gearsq)
```

And now with this new data that includes the squares of variables that appear to have a quadratic relationship, we should be able to find an accurate model.

To help narrow down our options in variables, I will use the stepAIC function in the MASS package of R. This function attempts to remove unnecessary variables from the model.

```
cars.mod <- lm(mpg~am+., data = carsDat)
stepAIC(cars.mod, scope = list(lower = ~am))
```

This code presents us with these values as the coefficients with the following P values:

This table presents us with the best model for dealing with correlations with other variables. To test this we will run residual tests on the model. These are presented in **Figure 3**.

As one can see in these residuals, the model looks much better than the original with all of the variables and no quadratics added in. Thus we can conclude that this model works well for the given situation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.9865	7.0077	4.99	0.0001
am	-1.5554	1.8280	-0.85	0.4040
cyl	1.5858	1.0012	1.58	0.1275
disp	-0.0966	0.0394	-2.45	0.0225
hp	-0.0940	0.0432	-2.18	0.0403
wt	-2.3698	1.3184	-1.80	0.0860
gear	2.0449	1.2988	1.57	0.1297
carb	-1.0919	0.6525	-1.67	0.1084
dispsq	0.0001	0.0001	2.74	0.0119
hpsq	0.0002	0.0001	2.16	0.0420

Conclusion

Recall our reason for doing all of these tests, our goal was to find out if automatic or manual transmission is better for gas mileage, and by how much. Well, if one refers to the help file, the am value in the model equals 0 if it is automatic and 1 if it is manual. So we can see that if $am = 1$ (if the car is manual) the value of the miles per gallon goes down approximately 1.55 in our model. This would lead us to believe that automatic cars get better gas mileage than manual cars. This is extremely interesting because in our original model (refer to **Figure 1**) the am coefficient in the model was positive which would have lead us to believe otherwise. So alternating the model as we needed actually helped to completely correct our answer. So our answer becomes that automatic transmission is better, and it gets about 1.55 miles per gallon better.

There is just one problem with this conclusion. If one were to look at the P value given in the table presented with our coefficients, one could see that the P value for the variable am isn't very low. This leads us to believe that in actuality, transmission doesn't seem to be an applicable factor in the model for miles per gallon.

Thus our final conclusion must be that transmission doesn't actually seem to make a difference because of the high P value. This gives us evidence that the model doesn't need transmission in order to be accurate. However, if transmission was needed, an automatic transmission would be better by approximately 1.55 miles per gallon.

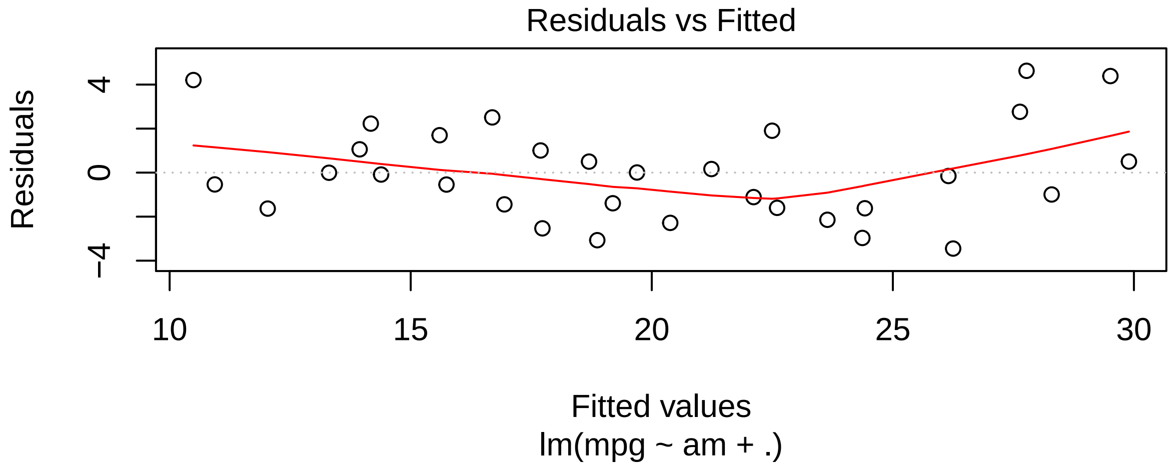
Appendix

Figure 1

```
all.cars <- lm(mpg~am+., data = mtcars)
xtable(summary(all.cars)$coef)
```

```
plot(all.cars, which=1:2, labels.id = '')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.30	18.72	0.66	0.52
am	2.52	2.06	1.23	0.23
cyl	-0.11	1.05	-0.11	0.92
disp	0.01	0.02	0.75	0.46
hp	-0.02	0.02	-0.99	0.33
drat	0.79	1.64	0.48	0.64
wt	-3.72	1.89	-1.96	0.06
qsec	0.82	0.73	1.12	0.27
vs	0.32	2.10	0.15	0.88
gear	0.66	1.49	0.44	0.67
carb	-0.20	0.83	-0.24	0.81



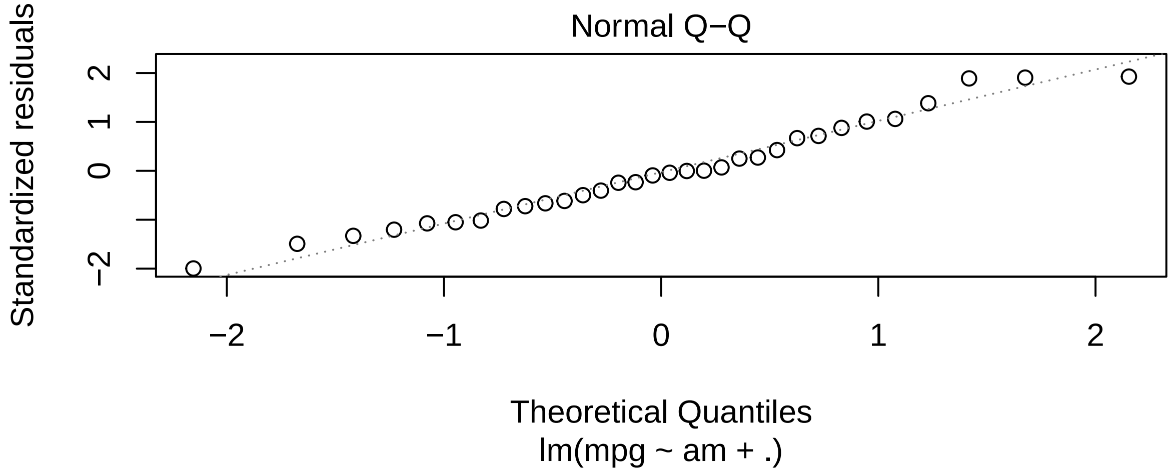
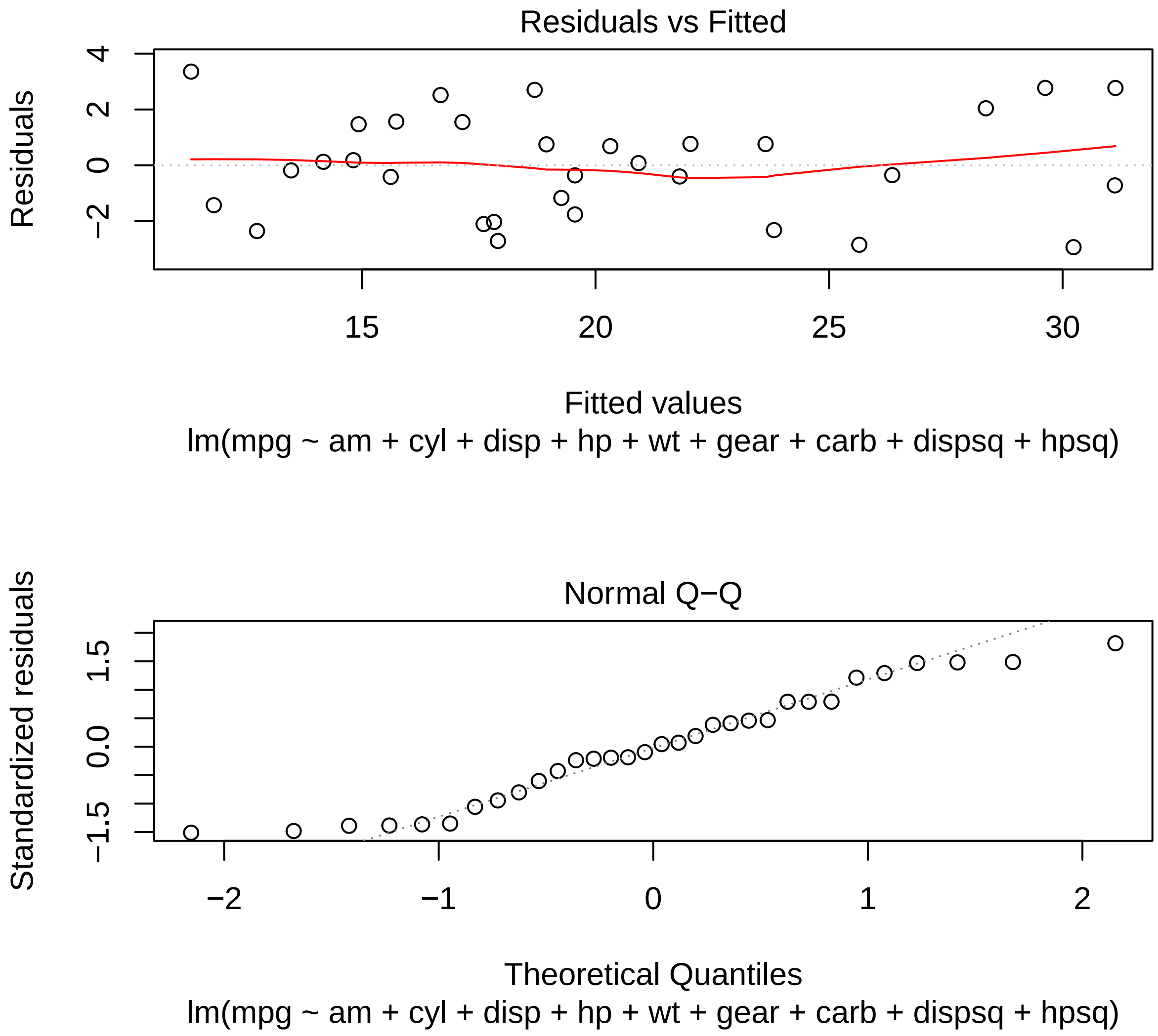


Figure 2

```r
pairs(mtcars)
```

Figure 3

```r
plot(cars.mod, which=1:2, labels.id = '')
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ am + cyl + disp + hp + wt + gear + carb + dispsq + hpsq)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ am + cyl + disp + hp + wt + gear + carb + dispsq + hpsq)