Regression Models Assignment

Diane Clow

September 27, 2015

Executive Summary This report is using the mtcars data provided within R. Using this data set I am studding the relationships specifically between transmission (manual and automatic) to MPG (miles per gallon) to answer the following questions

- Is a manual or automatic transmission better for MPG?
- Can you quantify the MPG difference between automatic and manual transmissions?

In performing the following analysis it is clear that for this data a manual transmission will get much better gas mileage than the automatic transmission. The difference in MPG can best be examined with the linear model that compares mpg to transmission, horsepower, weight of the car, and the number of cylinders.

Inputing and Visual Examination of the Data To start this we need to load a define the values provided in the mtcars data set. Below is the first six rows of the data to give a general feel as to what the data looks like.

```
##                   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
```

```
##
## Attaching package: 'ggplot2'
##
## The following object is masked from 'mtcars':
##
##     mpg
```

Inference and Is There A Difference The first step to answering the above questions is proving if there is a significant difference in the mpg depending on the transmission type. Looking at the box-plot shown in Appendix A.1, visually there is strong evidence that having an automatic transmission will give better gas mileage.

To prove that there is a difference depending on transmission I will perform a t-test. My null hypothesis is that the two means are equal, and my alternative hypothesis is that they are different.

```
diff_test <- t.test(mpg ~ am, data=mtcars)
diff_test
```

As shown in the test above the p-value is 0.001374. Because this is such an extreme p-value we reject the null hypothesis and conclude that there is a significant difference in the mpg values based on what type of transmission the car has.

Regression Analysis To start with a model using all the available parameters is built. This initial model has a Residual standard error of 2.833 on 15 degrees of freedom. The Adjusted R-squared value is .779. The results for the below code are hidden to conserve space.

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

Now I am going to use the step function in R to see if I get make my model simpler. The step function in R uses the AIC method to run a series of test to create the model with the minimum AIC score. AIC rewards the goodness of the fit, but it increases for each variable used. The results for the below code are hidden to conserve space.

```
Model <- step(FullModel)
summary(Model)
```

Running the test below, the best AIC value is 61.65 when the model calculated mpg using cly, hp, wt, and am. The residual standard error on this new model is 2.41 with 26 degrees of freedom and an adjusted R squared value of .8401. This means that a better model is produced using only these four values instead of all ten.

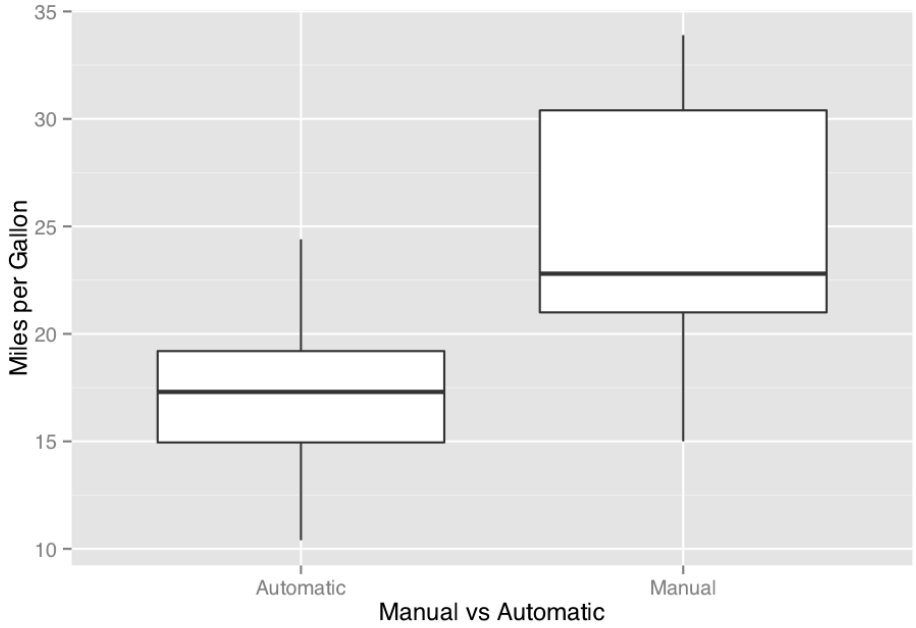
Residuals The last step before concluding that this is an accurate model is checking the residuals to make sure that there isn't some large factor that we should account for. A quick summary looking at four plots will help: Residuals vs Fitted, Q-Q plot, Scale-Location, and Residuals vs Leverage. See Appendix A.2 for the charts.

- The Residuals vs Fitted chart shows no pattern in the residuals, showing independence between the variables used.
- Because the Normal Q-Q plot mostly follows a straight line, the residuals are normally distributed
- The Scale-Location plot shows that the residuals are homeostatic, showing that the model is a good fit
- The Residuals vs Leverage plot shows no outliers, showing that there isn't a major problem that hasn't been accounted for

Appendex A.1

BOXPLOT for Inference and Is There A Difference

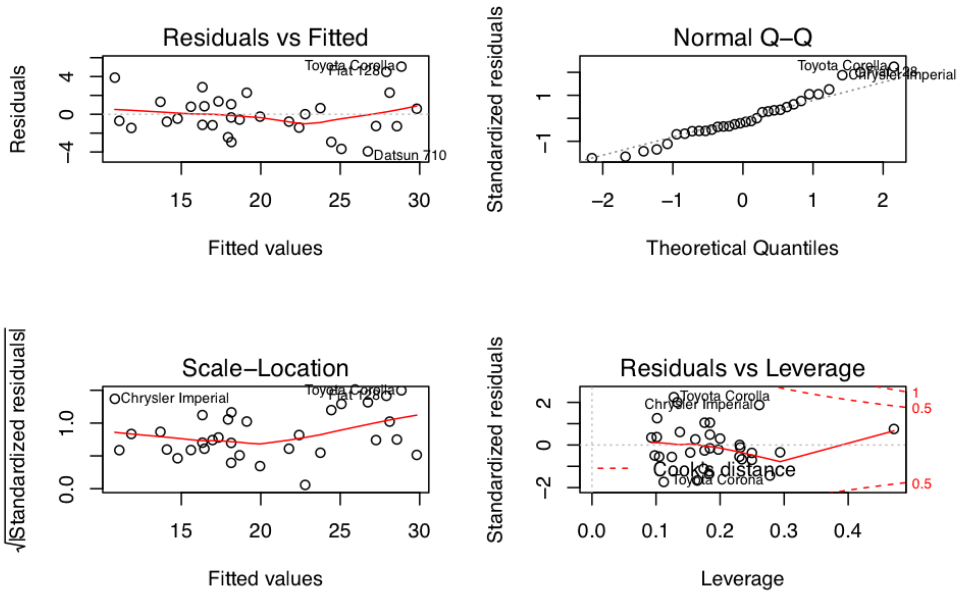
```
ggplot(mtcars, aes(x=am, y=mpg)) + geom_boxplot() +
    xlab("Manual vs Automatic") + ylab("Miles per Gallon") +
    scale_x_discrete(breaks=c("0", "1"),
                     labels=c("Automatic", "Manual"))
```



A.2

Residuals vs Fitted, Q-Q plot, Scale-Location, and Residuals vs Leverage for Residuals

```
FinalModel <- lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
par(mfrow = c(2, 2))
plot(FinalModel)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale–Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona