regression-analysis.Rmd

Karim Qazi

October 26, 2014

Executive Summary

We are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). This report attempts to answer the following questions as it relates to this high level goal:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions?

We will use the **mtcars** data set to perform this analysis. Both exploratory and inferential analsysis were conducted to create the results below.

Conclusions

The following factors illustrate why **weight** is a much better predictor of **mpg** than **transmission** is:

1. Looking at *Figure 1* in the appendix it's clear that this doesn't fit a linear model well. So I wouldn't use **transmission** to predict **mpg**.
2. Evaluating all the coeffficents for potential predictors of **mpg**, **weight** looks to do a good job of this with ~**3.72** decrease in **mpg**.
3. The low **p-value** when using **weight** to predict **mpg** also supports a good probability for this.
4. Looking at *Figure 2* in the appendix it's clear that **weight** IS a good predictor of **mpg**.
5. The very low residual of **weight** to predict **mpg** provides futher evidence of this. See *Figure 3* in the appendix to see this plotted as well as the actual residual value in the analysis below.

Exploratory Data Analysis

We can see that on average Manual transmission cars get ~ **70.3** better MPG when compared to Automatic transmissions. Which looks pretty compelling, but we haven't done enough analysis to quantify this just yet.

```
##
##
## | Transmission|   MPG|
## |------------:|-----:|
## |            0| 17.15|
## |            1| 24.39|
```

Another logical predictor of **mpg** could be **weight**. Looking at weight we can see it looks to be a potentially better predictor of MPG than transmission type is. Weight is the estimated expected change in **mpg** per 1,000 lb increase in weight.

NOTE: I rounded up the weight in tons so you could get a sense of this more easily, but this does loose valuable precision that must be accounted for in later analysis.

```
##
##
## | Weight (per 1,000 lb)|   MPG|
## |---------------------:|-----:|
## |                     2| 28.09|
## |                     3| 19.73|
## |                     4| 15.78|
## |                     5| 11.83|
```

This analysis will provide some guidance for what's to follow below.

Interpretation of Coefficents

Let's look at all the variables in the data set we can potentially predict **mpg** to all the other potential predictors of **mpg**. Since we originally wanted to see if **transmission** could predict **mpg** we can do a summary of all the coefficents. Seeing that **am** (which indicates transmission) looks to predict ~**2.52** increase in **mpg**, whereas **weight** (per 1,000 lb) predicts a decrease of ~**3.72** leads us to believe that **weight** is a better predictor of **mpg** from this analysis.

```
summary(lm(mpg ~ . , data=mtcars))$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788  0.6573  0.51812
## cyl         -0.11144    1.04502 -0.1066  0.91609
## disp         0.01334    0.01786  0.7468  0.46349
## hp          -0.02148    0.02177 -0.9868  0.33496
## drat         0.78711    1.63537  0.4813  0.63528
## wt          -3.71530    1.89441 -1.9612  0.06325
## qsec         0.82104    0.73084  1.1234  0.27394
## vs           0.31776    2.10451  0.1510  0.88142
## am           2.52023    2.05665  1.2254  0.23399
## gear         0.65541    1.49326  0.4389  0.66521
## carb        -0.19942    0.82875 -0.2406  0.81218
```

Next let's see how we'll this **weight** predictor is performing from a probability standpoint. If we look at the "p-value" when using weight as a predictor of "mpg" we can see that it has a very low p-value of **1.293959e-10** which should indicate that weight has a very good probability of predicting "weight".

```
x <- mtcars$wt
y <- mtcars$mpg
n <- length(y)
fit <- lm(y ~ x)
summary(fit)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   37.285     1.8776  19.858 8.242e-19
## x             -5.344     0.5591  -9.559 1.294e-10
```

All the basic data points are still supporting that **weight** is a good predictor of **mpg** with ~-5.3445 decrease in **mpg**. Let's plot the data and see how it looks in **Figure 2** in the appendix. **weight** looks to predict **MPG** well from a linear regression standpoint.

To validate that this is truly a good fit we can look at the residual variation between the actual value and the predictor when taking into account **weight** to predict **mpg**. You can see that the **max** residual is very, very small below. You can see this plotted by looking at **Figure 3** in the appendix.

```
e <- resid(fit)
yhat <- predict(fit)
max(abs(e -(y - yhat)))
```

```
## [1] 7.239e-14
```

Appendix

Figure 1

Illustrates that using transmission type to predict MPG is a poor model fit.

```r
plot(mtcars$am, mtcars$mpg, xlab="Transmission (0=Automatic, 1=Manual)", ylab="MPG")
abline(lm(mtcars$mpg ~ mtcars$am))
```

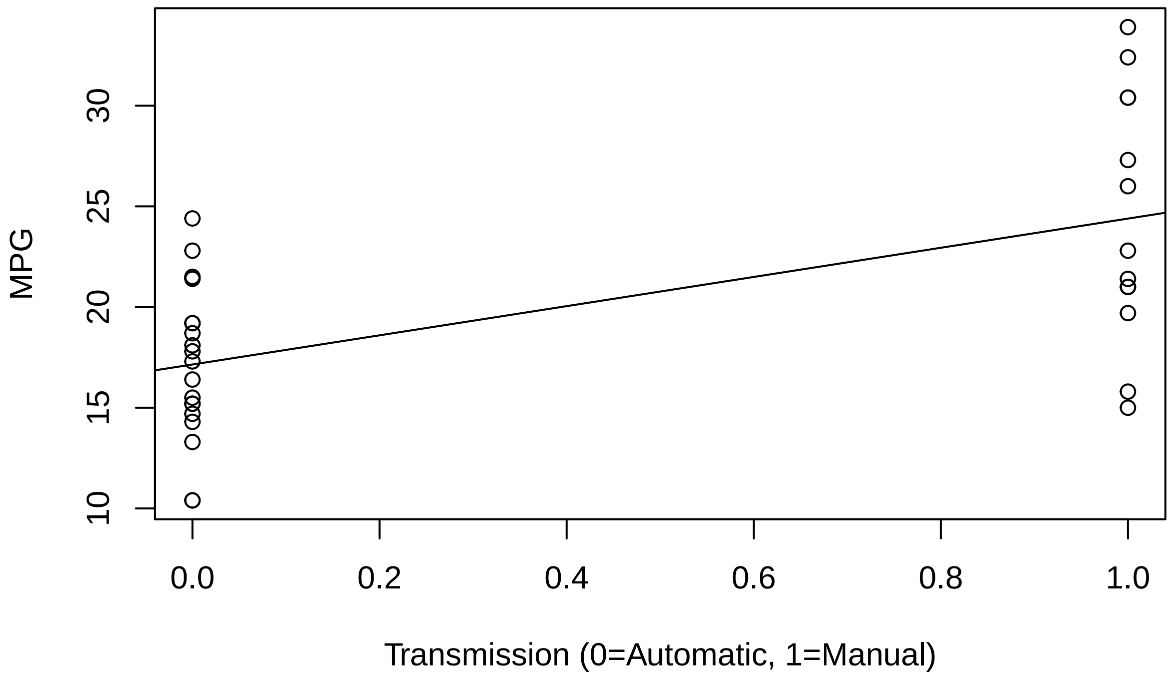


Figure 2

Linear fit of when using car weight to predict MPG.

```r
plot(mtcars$wt, mtcars$mpg, xlab="Weight (per 1,000 lbs)", ylab="MPG")
abline(lm(mtcars$mpg ~ mtcars$wt))
```

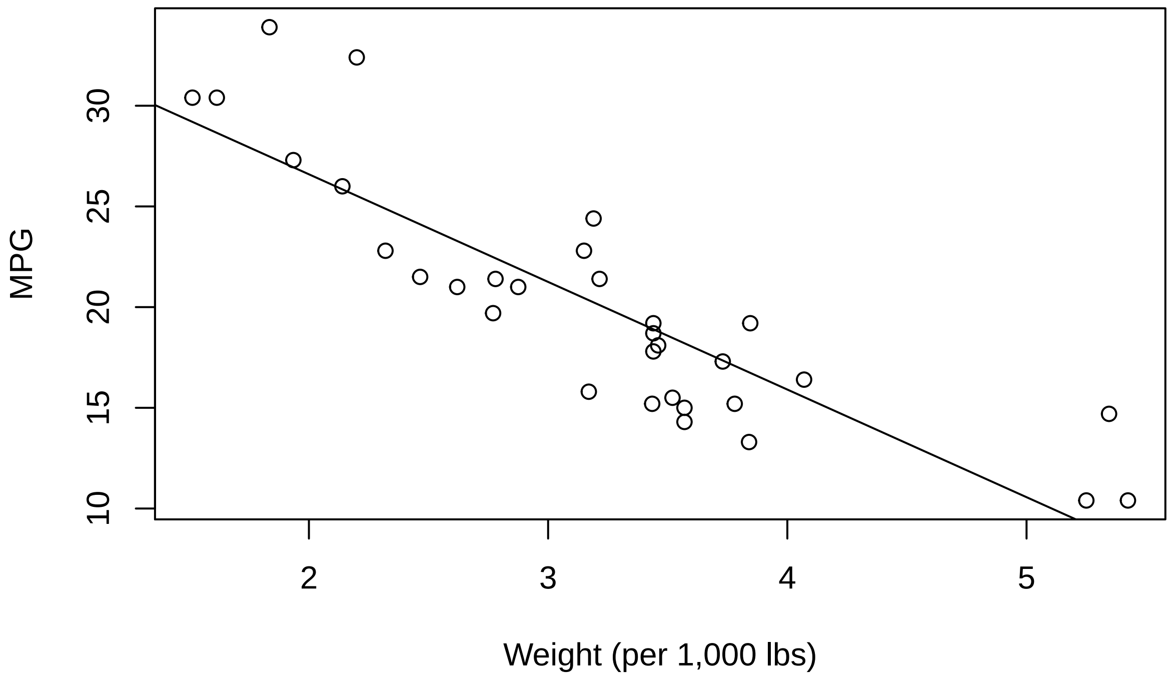
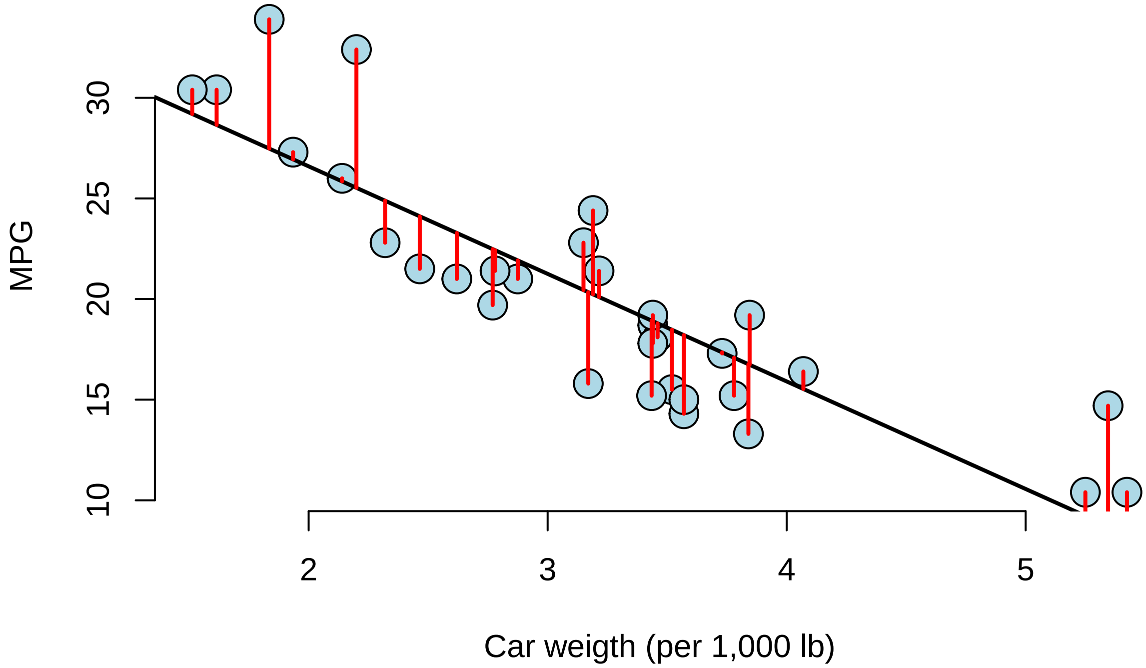


Figure 3

Residual variation for linear model using weight to predict MPG.

```
plot(x, y,
     xlab = "Car weigth (per 1,000 lb)",
     ylab = "MPG",
     bg = "lightblue",
     col = "black", cex = 2, pch = 21,frame = FALSE)
abline(fit, lwd = 2)
for (i in 1 : n)
  lines(c(x[i], x[i]), c(y[i], yhat[i]), col = "red" , lwd = 2)
```


MPG
10
15
20
25
30
2
3
4
5
Car weigth (per 1,000 lb)