Regression Analysis of mtcars Dataset

Nicholas Nappy

January 2015

Executive Summary

This is a basic regression analysis completed for the Johns Hopkins University Regression Models course project on Coursera (https://www.coursera.org/course/regmods) in January 2015.

The purpose of the analysis was to address two questions using the mtcars dataset that comes with base R:

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

mtcars Dataset

The mtcars dataset contains 32 observations with measurements on 11 variables extracted from the 1974 Motor Trend US magazine.

```
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
```

For more information on the dataset please visit http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html

Analysis and Model Selection

A comparison of the mean mpg by transmission type indicates there is a substantial difference in means between the two groups (0 = Automatic, 1 = Manual)

```
mtcars2 <- mtcars
mtcars2$am <- factor(mtcars$am)
mtcars2$cyl <- factor(mtcars2$cyl)
aggregate(mpg ~ am, mtcars2, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

Please see Figure 1.1 in the appendix for boxplots of the data.

A simple linear regression of mpg on transmission type confirms the relationship:

```
fit1 <- lm(mpg ~ am, mtcars2)
fit1$coef
```

```
## (Intercept)         am1
##   17.147368    7.244939
```

Though the relationship is statistically significant (not shown in R output because of space limiations), it is not clear whether transmission is simply acting as a 'surrogate' for one or more of the other variables collected.

Each variable was then added to the model and it's impact on the adjusted R^2 statistic evaluated. The variable which had the largest impact on this statistic was then added to the model until adding additional variables had little to no effect on the adjusted R^2 statistic.

The two models selected using this approach follow*:

```
fit2 <- lm(mpg ~ am + hp + wt, mtcars2)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am1          2.083710   1.376420   1.514 0.141268
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## wt          -2.878575   0.904971  -3.181 0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

```
fit3 <- lm(mpg ~ am + hp + wt + cyl, mtcars2)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt + cyl, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## am1          1.80921    1.39630   1.296  0.20646
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

When covariates are included in the model, the effect of the transmission type on mpg decreases and is not statistically significant in either of the two models.

```r
anova(fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + hp + wt
## Model 2: mpg ~ am + hp + wt + cyl
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 180.29
## 2     26 151.03  2    29.265 2.5191    0.1 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA analysis indicates that the model including the 'cyl' variable is not statistically significant and therefore does not offer an improvement on the model without it. In order to select the "best" model, comparative residual plots (see figures 1.2 and 1.3 in the appendix) and diagnostics were performed.

```r
library(car)
vif(fit2)
```

```
##       am       hp       wt
## 2.271082 2.088124 3.774838
```

```r
vif(fit3)
```

```
##         GVIF Df GVIF^(1/(2*Df))
## am  2.590777  1        1.609589
## hp  4.703625  1        2.168784
## wt  4.007113  1        2.001778
## cyl 5.824545  2        1.553515
```

fit3 has a higher adjusted R^2 statistic, relatively low VIF and the residuals more closely approximate a normal distribution than the residuals of fit2. Therefore, it seems fit3 models the data more accurately than fit2.

Conclusion

The coefficient for transmission type in the fit3 model is 1.81 which has the following interpretation: when hp (horsepower), wt (weight) and cyl (# of cylinders) of the vehicles are taken into account, the mpg will increase by 1.81 if it is a manual transmission. However, this value is not statistically significant and therefore it is reasonable to conclude that transmission type has no effect on mpg when adjusting for the effects of horsepower, weight and number of cylinders in the model.

This conclusion holds true for the fit2 model as well.

Appendix

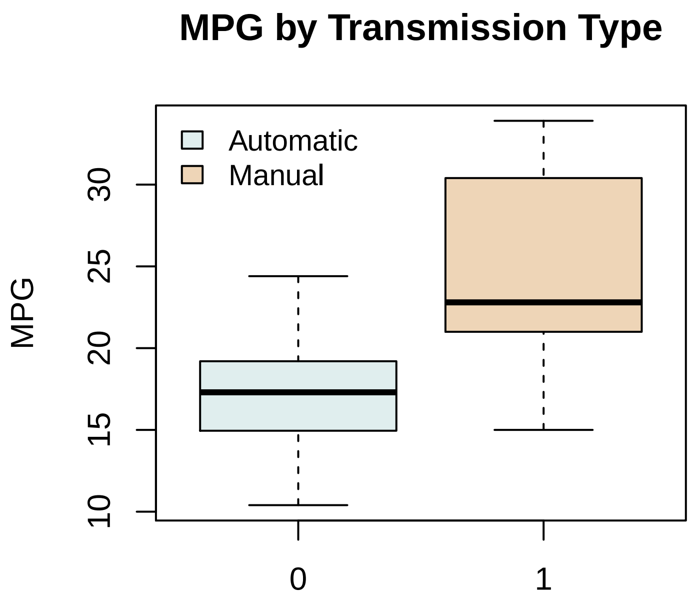

Figure 1.1

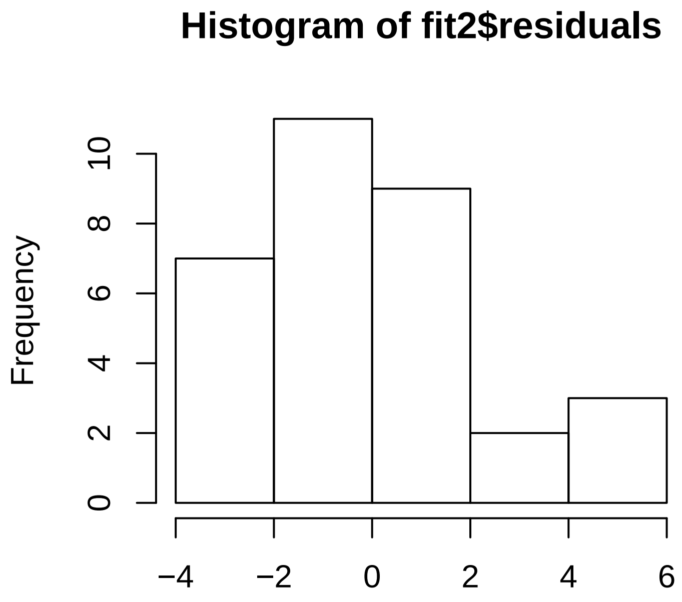

Figure 1.2

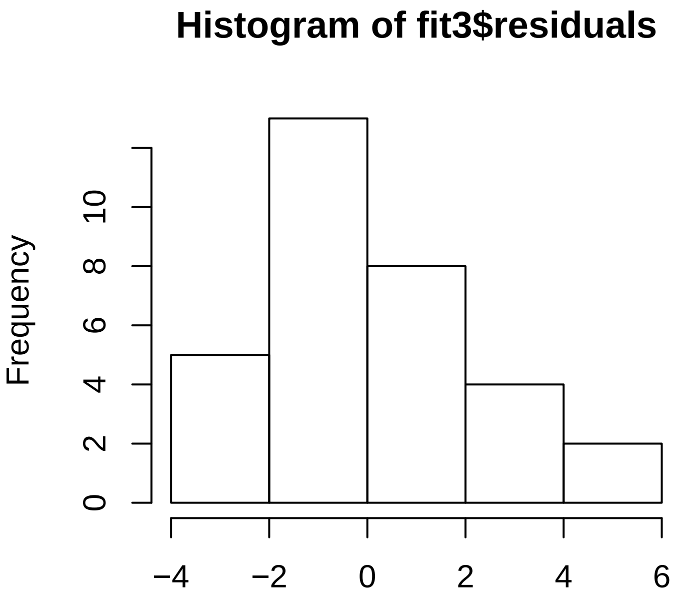

Figure 1.3

- A model with interaction terms between hp and wt ‘lm(mpg ~ am + wt:hp)’ also produced good results though it was not selected because of extremely high variance inflation factors and space limitations of this report.