Regression Models Course Project

Jonathan Chang

March 9, 2017

Executive Summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

Exploratory Data Analysis

First, we load the mtcars dataset, and take a quick look at variables. We take the *am* variable we are interested in, and change it to a factor variable with "Automatic" and "Manual" levels. (0 corresponds to automatic and 1 corresponds to manual.)

```
data(mtcars)
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
```

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

We then do a quick boxplot (in the appendix) of MPG vs. Transmission type to take an initial look at the effect of Automatic vs. Manual transmissions on MPG performance. We see that the plot (including the median) for Manual transmission is higher than that of the Automatic transmission, suggesting better performance for cars with Manual transmissions.

```
boxplot(mpg ~ am, data = mtcars, main="MPG vs. Transmission Type", ylab="MPG")
```

We then perform a t-test (in appendix), assuming that the transmission data has a normal distribution. The null hypothesis is that the type of transmission (automatic or manual) does not effect a change in MPG.

```
t.test(mpg ~ am, data=mtcars,paired=FALSE,var.equal=FALSE)
```

Our p-value $= 0.001374 < 0.05$ which means that we reject the null hypothesis that there is no difference in MPG, and stick with our observation that manual trasmission is better for MPG.

Regression Model

We begin by fitting a basic linear regression model of am (predictor) on mpg (outcome).

```
fit_basic <- lm(mpg ~ am, data = mtcars)
summary(fit_basic)
```

This model shows us that an automatic transmission car has 17.147 mpg, while with manual transmission, mpg increases by 7.245. We have a residual standard error of 4.902 on 30 degrees of freedom which means that we have almost 5mpg unexplained by our model. We have an Adjusted R-squared value of 0.3385, which means that we can only explain about 34% of the total variability of our model with am as the sole predictor.

Next, we try to find a better model. We begin with a full model with all the variables as predictors.

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

Here, we see that we have a Residual standard error of 2.65 on 21 degrees of freedom and an Adjusted R-squared of 0.8066, which is significantly better than our basic model. The problem is that none of the coefficients are significant at the 0.05 significance level.

Finally, we use the step method to perform variable selection in both directions based on AIC, a means of measuring the relative quality of a statistical model for a given set of data.

```
fit_step <- step(fit_full, direction="both")
summary(fit_step)
```

It turns out that our best model used the following variables: wt, qsec, and am. We have a Residual standard error of 2.459 on 28 degrees of freedom and an Adjusted R-squared of 0.8336. So, we can say that our model can explain about 83% of the variability of the MPG variable.

Conclusions

```
summary(fit_step)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

```
confint(fit_step)
```

```
##                   2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
## amManual     0.04573031  5.825944
```

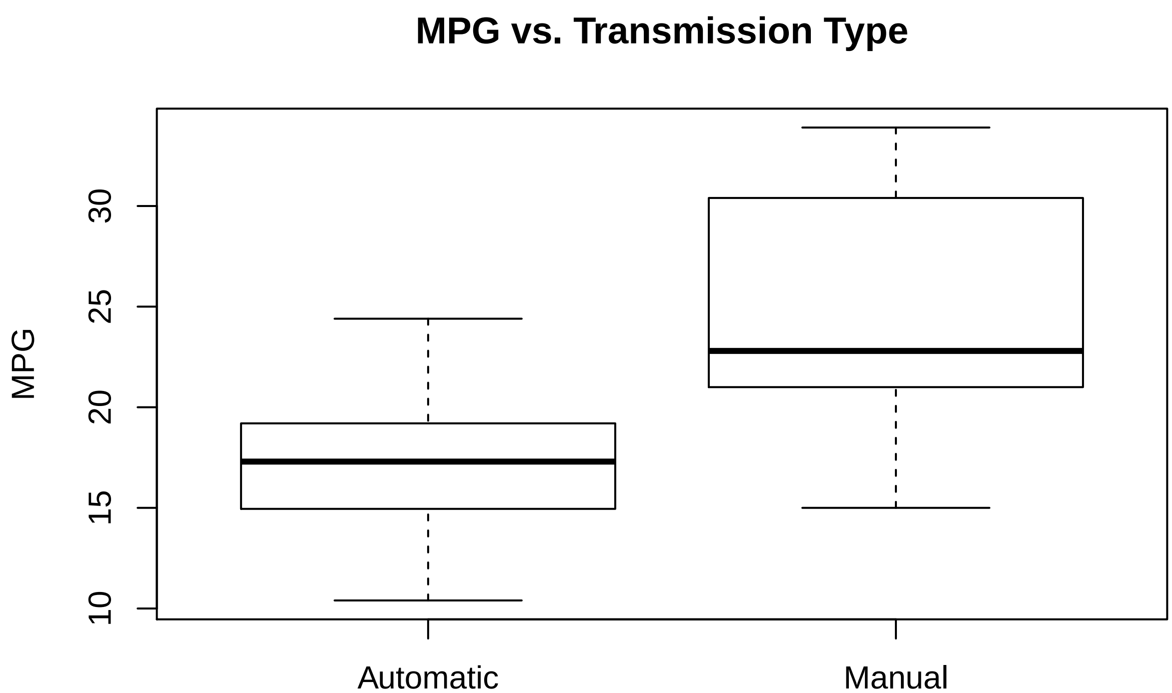
Using anova (in the appendix), we see that the p-value for our fit_step model is highly significant, and we can reject the null hypothesis that wt, qsec, and am do not contribute to the change in MPG. So we select the model: mpg ~ wt + qsec + am. In our summary, we see that our coefficients have p-values < 0.05 and they make sense. For instance, every increase in 1,000lb decreases MPG by roughly 2.5 which is intuitive. Meanwhile, what we are really interested in quantifying, is that with a manual transmission, MPG is increased by 2.94. Taking the confint of our fit_step model, we are 95% confident that this value is between 0.05 and 5.83.

Appendix

```r
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```r
boxplot(mpg ~ am, data = mtcars, main="MPG vs. Transmission Type", ylab="MPG")
```




```r
t.test(mpg ~ am, data=mtcars,paired=FALSE,var.equal=FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual 
##                17.14737                24.39231
```

```r
summary(fit_basic)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```r
summary(fit_full)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
summary(fit_step)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```r
anova(fit_basic, fit_step, fit_full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```