Regression analyses of MPG data

Jon

19 January 2016

Executive Summary

This report uses the mtcars dataset to evaluate the follow questions.

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

To answer these questions I first evaluated several linear regression models to select the one that best described the data without overfitting. The model selected was `mpg ~ wt + qsec + am`. Using the coefficients of this model I determined the effect size for switching from automatic to manual while the other variables are kept the same. The conclusion based on this investigation is that manual is better for MPG, and has a estimated effect of 2.9 MPG increase for switching from automatic to manual.

Exploratory Data Analyses

The mtcars dataset contains 11 variables, output of `str(mtcars)` and `summary(mtcars)` can be seen in figure 1 of the appendix. From this output we can see that transmission type is coded as a numeric, but really should be a factor. This is corrected using `mtcars$am <- as.factor(mtcars$am); levels(mtcars$am) <- c("auto", "man")`. Before correcting this a correlation matrices `cor(mtcars)` can be generated to analyse variable correlations, See figure 2 of the appendix for output.

Method for model selection

To answer the questions about transmission type on MPG we need to generate a model that will allow the transmission type to be evaluated while holding the other factor equal, i.e. simply comparing the mean MPG of the two transmission types is likely invalided as transmission type is confounded with other variables that influence MPG. Simply creating a model that uses all variables is also not the best model as with the small dataset it's likely to overfit the data.

Using the correlation table focussing on the mpg column the wt (weight) has the highest correlation to MPG. The next 3 values cyl, disp and hp are also high, but when checking these variables for correlation to wt it can be seen they are also highly correlated and inclusion will increase the model bias. To select further variables the step function is used to automatically search for the best model.

```
mod <- lm(mpg ~ ., data = mtcars)
stepOutput <- step(mod,trace=FALSE)
stepOutput$call
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

The output of steps is wt, qsec and am. To confirm the output of steps, we can compare the models `mpg ~ am` vs `mpg ~ am + wt` vs `mpg ~ am + wt + qsec` and `mpg ~ am` vs `mpg ~ am + qsec` vs `mpg ~ am + qsec + wt` using anova see figure 3 for outputs.

In both cases we see that adding both wt and qsec had a significant effect ($P < 0.05$). Therefore the model that will be used to answer the MPG questions is `mpg ~ wt + qsec + am`.

Model Residual plot and Diagnostics

Reviewing the residuals plots from figure 4 the only things looks a bit concerning is the Residuals vs fitted plot which may be showing a bit of heteroskedasticity, although it doesn't appear to be too bad and may be just an artefact of the low sample numbers.

Conclusion

Summary of selected model.

```
##
## 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 ***
## am            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
```

With 95% confidence, we estimate that switching from automatic to manual while keeping weight and 1/4 second mile time the same will results in a 0.05 to 5.83 increase in MPG. Therefore, given a car switching from automatic to manual, which is unlikely to affect the weight or 1/4 second mile time is better for MPG, although this difference may be very small.

Appendix

figure 1

```
data("mtcars")
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  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

figure 2

```
cor(mtcars)
```

```
##             mpg        cyl       disp         hp        drat         wt
```

```
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("auto", "man")
```

figure 3

```
fit1 <- lm(mpg ~ am, data = mtcars)
fit2 <- lm(mpg ~ wt + am , data = mtcars)
fit3 <-lm(mpg ~ wt + qsec + am, data = mtcars)
anova(fit1, fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 73.203 2.673e-09 ***
## 3     28 169.29  1    109.03 18.034 0.0002162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit1 <- lm(mpg ~ am, data = mtcars)
fit2 <- lm(mpg ~ qsec + am , data = mtcars)
fit3 <-lm(mpg ~ wt + qsec + am, data = mtcars)
anova(fit1, fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ qsec + am
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 352.63  1    368.26 60.911 1.679e-08 ***
## 3     28 169.29  1    183.35 30.326 6.953e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

figure 4