Regession Models - Course Project

pdwessel

March 18, 2017

Executive Summary

This report looks at the relationship between transmission type and fuel efficiency, using the *mtcars* dataset. We start by evaluating the correlation between all the potential variables to identify predictors to include in the model. A number of linear multivariable models are built and compared via an ANOVA table. A model is selected with transmission type and number of carburetors as the predictors. The residual is examined for patterns that would suggest an ill fit. Finally the coefficient for transmission type tested for signifance versus the null hypothesis. The report find that a manual transmission increased fuel efficiency by ~7.09 miles/gallon over an automatic transmission.

Introduction

The objective of this report is to identify generally is an automatic or manual transmission is better for fuel economy and then quatify the difference between the two. For this report the *mtcars* dataset is used.

Exploratory Analysis

First we will look to see what variables are in the dataset and their respective classes.

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4       21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag   21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

From *str(mtcars)* we can see that all the variables are type numeric, and from *?mtcars* we can see that 0 and 1 represent automatic and manual transmissions respectively.

Analysis

Step 1 - Correlation

As we are interested in explaining mpg with different transmisions, I am going to look at the correlation between transmission type (am) and the others provided to determine which variables are candidates for exlusions when fitting the model.

From the correlation coefficients, we see that the least correlated variables are 'qsec', 'am', 'gear', and 'carb'. My strategy for buidling linear multivariable models will be to start with 'am' and then add in these three variables according to how they are correlated to 'am'. Thus I will build 4 models, and look at the anova table to determine my 'best fit'. Note that I will explictly make carb and gear factor variables. The models will be:

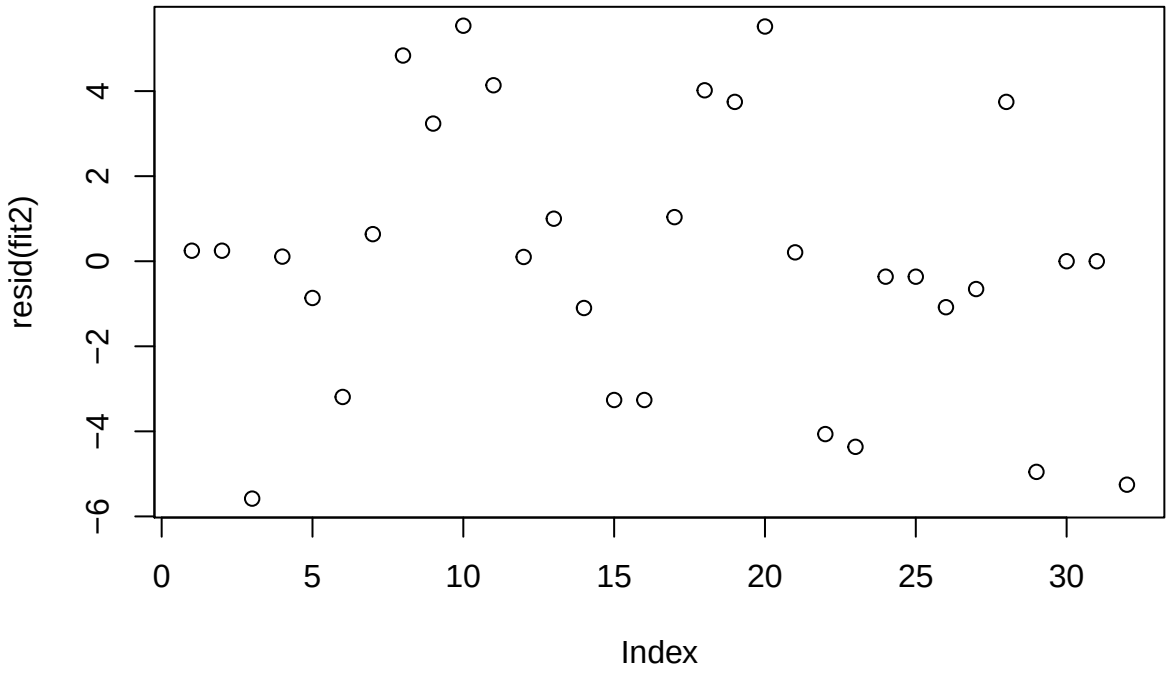
- mpg ~ am
- mpg ~ am + factor(carb)
- mpg ~ am + factor(carb) + qsec
- mpg ~ am + factor(carb) + qsec + factor(gear)

From the anova table in annex a, I will discard model 4 as the probably of significance code from the p value is greater than 0.1. I took a look at the coefficients for models 2 and model 3.

What we see is that the intercept in model 3 is -2.07 and has a **huge** standard deviation of 9.34 and based on the test statistic we cannot infer with any confidence that this is the acutal intercept for this model. This is really due to qsec variable which is the 1/4 mile time for the car. This variable will never be zero, but given the SD I am going to throw out this model as well.

That leaves me with model 2 as mu chosen model which contains tranmission type (am) and the number of carburetors (carb) to explain mpg.

The last step is to take a quick look at the residual plot to ensure that there is no discernable pattern which would suggest that I missed some predictor variable.



The residual values seem to be distributed fairly evenly around zero with no discernable pattern. Thus the model becomes:

$$mpg = 21.29 + 7.09am - 1.73carb_2 - 4.99carb_3 - 7.62carb_4 - 8.68carb_6 - 13.38carb_8$$

What this means for our mpg~am relationship is that going from an automatic transmission (am = 0) to a manual transmission (am = 1) will increase the mpg by roughly 7.09 miles/gallon.

The final evaluation will the against the coefficient itself to verify that it is statistically signifance vs the null (the coefficent is zero). Looking at the coefficients for the model, we see that the test statistic for the hypothesis that the increase is 7.09 vs no increase is 4.99, which gives a p-value of 3.79e5 which mean we can reject the null hypothesis and the coefficient is valid.

Summary

In this report, we looked at numerous models to look at the relationship between transmisson type and fuel efficiency (mpg). We determined that in general a manual transmision will increase fuel efficiency, and we quantifed that increase to be roughly 7.09 miles/gal.

Annex A - Code

```
knitr::opts_chunk$set(echo = FALSE)
data("mtcars")
head(mtcars,2)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

```
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
```

```
fit1 <- lm(mpg~am, data = mtcars)
fit2 <- lm(mpg~am+factor(carb), data = mtcars)
fit3 <- lm(mpg~am+factor(carb)+qsec, data = mtcars)
fit4 <- lm(mpg~am+factor(carb)+qsec+factor(gear), data = mtcars)
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + factor(carb)
## Model 3: mpg ~ am + factor(carb) + qsec
## Model 4: mpg ~ am + factor(carb) + qsec + factor(gear)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     25 313.12  5    407.78 8.8408 0.000102 ***
## 3     24 247.16  1     65.96 7.1502 0.013864 *
## 4     22 202.95  2     44.21 2.3961 0.114439
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit2)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    21.291494   1.564399 13.610019 4.618193e-13
## am              7.089885   1.419669  4.994042 3.785004e-05
## factor(carb)2  -1.727448   1.760945 -0.980978 3.360023e-01
## factor(carb)3  -4.991494   2.573368 -1.939674 6.378189e-02
## factor(carb)4  -7.628460   1.786108 -4.270995 2.464016e-04
## factor(carb)6  -8.681379   3.831978 -2.265508 3.239872e-02
## factor(carb)8 -13.381379   3.831978 -3.492029 1.801134e-03
```

```
summary(fit3)$coef
```

```
##                   Estimate Std. Error      t value     Pr(>|t|)
## (Intercept)   -2.073115949  9.3403808 -0.221951973 8.262291e-01
## am             8.230478010  1.3639238  6.034411879 3.131037e-06
## factor(carb)2  0.006334381  1.7375267  0.003645631 9.971213e-01
## factor(carb)3 -2.196796608  2.5815509 -0.850960016 4.032023e-01
## factor(carb)4 -4.358965182  2.0717151 -2.104037000 4.603593e-02
## factor(carb)6 -4.504549493  3.8467486 -1.171001783 2.530937e-01
## factor(carb)8 -8.156648287  4.0417350 -2.018105641 5.489162e-02
## qsec           1.164334673  0.4600623  2.530819665 1.834791e-02
```

```
plot(resid(fit2))
```