Regression Models Project

Nikolas Markou - nikolasmarkou@gmail.com

Introduction

Our assignment is to answer whether or not and how much the Miles Per Gallon (mpg) variable in the mtcars dataset is influenced by the transmission type (am) which may be automatic(0) or manual(1). The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Analysis

Our first test is to establish (before we start modelling) if there is a statistically significant difference between the groups that use automatic and manual transmissions

```
data(mtcars)
t.test(mtcars[mtcars$am == 0,]$mpg, mtcars[mtcars$am == 1,]$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == 0, ]$mpg and mtcars[mtcars$am == 1, ]$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

```
fitTest0 <- lm(mpg~factor(am), data=mtcars)
```

There is statistical significant difference between the non adjusted groups. It shows that manual has greater mpg that automatic by 7.24 miles per gallon. However this difference may not be explained by the transmission alone and it may be coincidence that the cars in the one group were found to have better mileage due to some other factor. To test our hypothesis we fit the data to a model where the difference between the 2 groups is assumed as a constant value. That is there is a constant difference of `mpg` between automatic and manual transmission cars.

```
fitTest1 <- lm(mpg ~ factor(am) + . - am, data=mtcars);
summary(fitTest1);
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + . - am, 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
## factor(am)1  2.52023    2.05665   1.225   0.2340
## 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
## 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
```

We can see that the constant `factor(am)` is not statistically significant. That means that transmission alone may not explain the `mpg` difference well. From the summary we can see that `weight` seems much more significant. So we adjust our model to incorporate that and drop variables of lesser imprortance while keeping the `factor(am)` in. We choose to include the factors `weight`, `horsepower` and `1/4 mile time`.

```
fitTest2 <- lm(mpg ~ factor(am) + wt + hp + qsec, data=mtcars);
summary(fitTest2);
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + hp + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4975 -1.5902 -0.1122  1.1795  4.5404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.44019    9.31887   1.871  0.07215 .
## factor(am)1  2.92550    1.39715   2.094  0.04579 *
## wt          -3.23810    0.88990  -3.639  0.00114 **
## hp          -0.01765    0.01415  -1.247  0.22309
## qsec         0.81060    0.43887   1.847  0.07573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.435 on 27 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8368
## F-statistic: 40.74 on 4 and 27 DF,  p-value: 4.589e-11
```

This model seems to explain the data better (less residual error and `R^2`) while keeping the possibility of overfitting low. It also shows that `weight` and `raw horsepower` are much better predictors of the car's

mileage. We can also deduce from the model that while still not extremely statistically significant the transimission factor is still positive, meaning that manual transmission is better by 2.5 miles per gallon. If however we assume that `hp` and `1/4 mile time` are correlated we could do better at fitting our model with fewer variables. To do so we use the R function `step()` that incrementally fits the best model by removing correlated or unnecessary variables.

```
fitBest <- step(fitTest2, direction='both')
```

```
## Start:  AIC=61.52
## mpg ~ factor(am) + wt + hp + qsec
##
##              Df Sum of Sq    RSS    AIC
## - hp          1     9.219 169.29 61.307
## <none>                    160.07 61.515
## - qsec        1    20.225 180.29 63.323
## - factor(am)  1    25.993 186.06 64.331
## - wt          1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ factor(am) + wt + qsec
##
##              Df Sum of Sq    RSS    AIC
## <none>                    169.29 61.307
## + hp          1     9.219 160.07 61.515
## - factor(am)  1    26.178 195.46 63.908
## - qsec        1   109.034 278.32 75.217
## - wt          1   183.347 352.63 82.790
```

```
summary(fitBest);
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + qsec, 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
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

So finally our model seems to fit the data very nicely, with lower residual error and fewer confounding variables. From the results it is statistically significant that the manual transmission gives you 2.93 more

miles per gallon compared to automatic. This is also confirmed by running ANOVA test on our first model that takes only transmission type as input and our best model that uses `weight` and `1/4 mile time`.

```
anova(fitTest0,fitBest)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also run `confint()` to get the 95% confidence interval of the constant `mpg` improvement due to the transmission type.

```
confint(fitBest)
```

```
##                   2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## factor(am)1  0.04573031  5.825944
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
```

Conclusion

The final analysis suggests that there is statistically significant difference between cars with automatic transmission and manual transmission. From the dataset we conclude with 95% accuracy that **cars with manual transmission have higher mpg** and that change is with 95% confidence between **0.04 and 5 miles per gallon** with the most probable value being **2.9** miles per gallon improvement compared to the automatic transmission.