Mtcars dataset analysis

Ola

Friday, June 19, 2015

This is an analysis for the Regression Models course by the Johns Hopkins University on Coursera.

The dataset mtcars 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). The data frame has 32 observations on 11 variables. We load the data set below, together with the ggplot2 library.

```
data(mtcars)
library(ggplot2)
```

Is an automatic or manual transmission better for MPG?

Our dataset is divided into automatic (am=0) or manual (am=1) transmission cars. We perform a t-test to see if a difference between mean MPG values for this subsets of data (am=0 and am=1) is statistically significant.

```
test_result <- t.test(mpg ~ am, data = mtcars)
test_result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

```
test_result$p.value
```

```
## [1] 0.001373638
```

The p-value is 0.0014 and we can reject the null hypothesis. We can conclude that the mpg of am=1 (manual) group is significantly larger than am=0 (automatic) group. This is illustrated in Fig. 1 in the Appendix.

Quantify the MPG difference between automatic and manual transmissions"

We fit the model to the available data. The simplest fit is based on am as the only regressor.

```
fit1 <- lm(mpg ~ am, data=mtcars)
summary(fit1)$adj.r.squared
```

```
## [1] 0.3384589
```

This simple model has adjusted R-squared: 0.34. A more complex model would take into account also other parameters. To see how much other parameters available in mtcars data influence mpg values, we use the aov function.

```
full_param_set <- aov(mpg ~ ., data = mtcars)
summary(full_param_set)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *
## hp           1    9.4     9.4   1.334  0.26103
## drat         1   16.5    16.5   2.345  0.14064
## wt           1   77.5    77.5  11.031  0.00324 **
## qsec         1    3.9     3.9   0.562  0.46166
## vs           1    0.1     0.1   0.018  0.89317
## am           1   14.5    14.5   2.061  0.16586
## gear         1    1.0     1.0   0.138  0.71365
## carb         1    0.4     0.4   0.058  0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variables with low p-value (below 0.05) are most influencial. We choose cyl, disp, wt and am for the fit to a linear model. You can see an illustration of that in Fig. 2 in the appendix.

```
fit2 <- lm(mpg ~ cyl + disp + wt + am, data=mtcars)
summary(fit2)$coefficients
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 8.677574e-12
## cyl         -1.784173258 0.61819218 -2.8861142 7.581533e-03
## disp         0.007403833 0.01208067  0.6128661 5.450930e-01
## wt          -3.583425472 1.18650433 -3.0201537 5.468412e-03
## am           0.129065571 1.32151163  0.0976651 9.229196e-01
```

We can remove disp from the fit, as it is not significant. (We prefer to leave am, as it is assumed in the given problem).

```
fit3 <- lm(mpg ~ cyl + wt + am, data=mtcars)
summary(fit3)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
## am           0.1764932  1.3044515  0.1353007 8.933421e-01
```

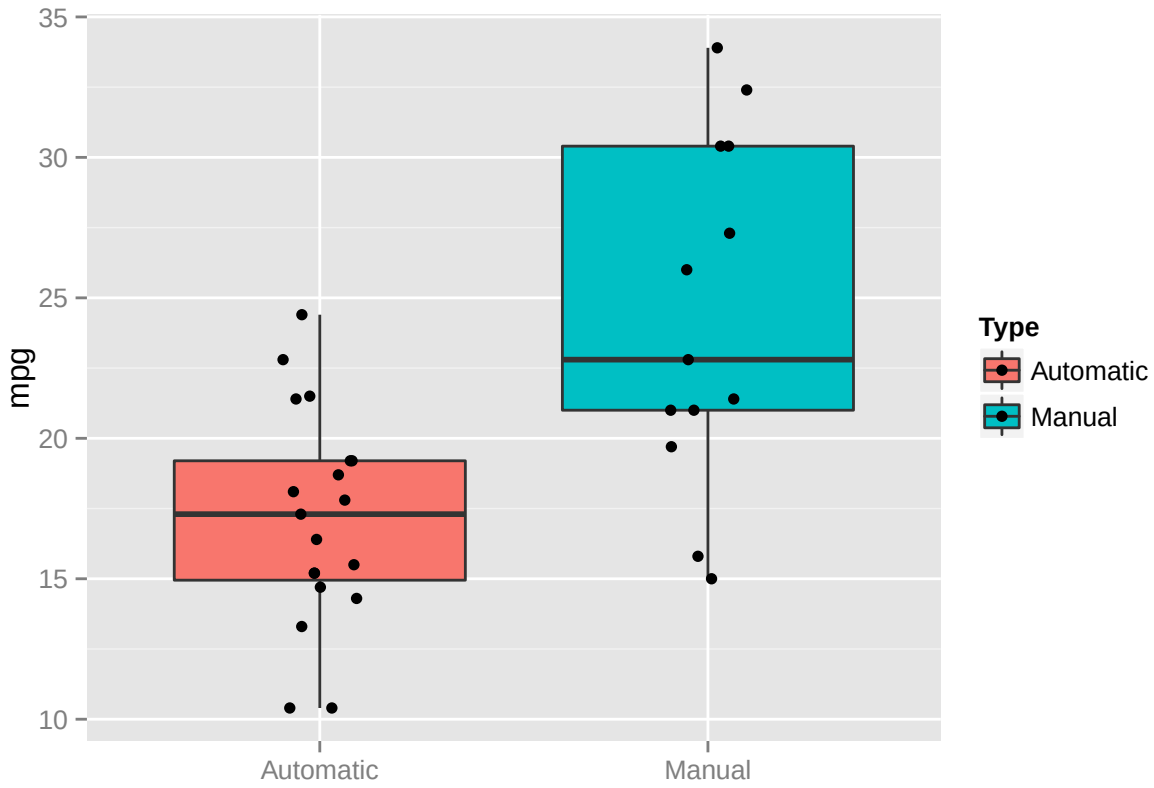
This model seems to be our best - its adjusted r-squared is 0.81 and this is our final model. However, we cannot reject the null hypothesis for the variable am.

Our first impression that the am is very significant in predicting the value of mpg occured to be flase, when other variables has been included. Other variables: cyl and wt are more important to determination of mpg.

Appendix A - supporting figures

MPG boxplot for am=0 (automatic) and am=1 (manual).

```
#Fig. 1
g = ggplot(mtcars, aes(factor(am), mpg, fill=factor(am)))
g = g + geom_boxplot()
g = g + geom_jitter(position=position_jitter(width=.1, height=0))
g = g + scale_colour_discrete(name = "Type")
g = g + scale_fill_discrete(name="Type", breaks=c("0", "1"),
            labels=c("Automatic", "Manual"))
g = g + scale_x_discrete(breaks=c("0", "1"), labels=c("Automatic", "Manual"))
g = g + xlab("")
g
```



A matrix of scatterplots for mtcars data is produced.

```
#Fig. 2
pairs(mtcars, panel=panel.smooth, main="MTcars pair graphs")
```

MTcars pair graphs