Project

Thursday, December 18, 2014

Let's take a look and investigate the dataset mtcars. According to the R description, 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).

This is a data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (lb/1000)

[, 7] qsec 1/4 mile time

[, 8] vs V/S

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

Source of data: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391-411.

We have to analyze this data in the following context: "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)."

The following questions must be answered from this data: -Is an automatic or manual transmission better for MPG? -Quantify the MPG difference between automatic and manual transmissions?

Let's take a look at the data, grouping it by transmission and looking at mpg variable:

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.2
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
summarise(group_by(mtcars, am), mean(mpg), sd(mpg), qty=n())
```

```
## Source: local data frame [2 x 4]
##
##   am mean(mpg)  sd(mpg) qty
## 1  0  17.14737 3.833966  19
## 2  1  24.39231 6.166504  13
```

At first sight, the mpg mean for automatics are lower than those that are manual. So automatics seem to be more inefficient but we cannot conclude that yet.

```
lm(data=mtcars, formula=mpg ~ 0 + factor(am))
```

```
##
## Call:
## lm(formula = mpg ~ 0 + factor(am), data = mtcars)
##
## Coefficients:
## factor(am)0  factor(am)1
##       17.15        24.39
```

If we try to create a linear regression model, we can see above that the coefficients are the means of both automatic and manual measured mpg. So we did not get anything new here.

Let's try a regression to the mean:

```
lm(data=mtcars, formula=I(mpg - mean(mpg)) ~ 0 + factor(am))
```

```
##
## Call:
## lm(formula = I(mpg - mean(mpg)) ~ 0 + factor(am), data = mtcars)
##
## Coefficients:
## factor(am)0  factor(am)1
##      -2.943        4.302
```

So, in other words, automatic cars mpg are in average 4.302 above the mean, while manuals are in average 2.943 below the mean. Now we begin to see the difference between automatics and manuals.

Let's investigate further. If we try to predict the 95% confidence interval of this linear regression, we would have the following:

```
reg <-lm(data=mtcars, formula=mpg ~ 0 + factor(am))
predict(reg, data.frame(am=0), interval="confidence") ## automatic
```

```
##        fit      lwr      upr
## 1 17.14737 14.85062 19.44411
```

```
predict(reg, data.frame(am=1), interval="confidence") ## manual
```

```
##        fit      lwr      upr
## 1 24.39231 21.61568 27.16894
```

In 95% of the cases, we will have an automatic car mpg between [17.147, 19.444] and a manual mpg between [24.392, 27.168], which is quite a difference, since in the worst case scenario we will have a difference of about 24.392 - 19.444 = 4.948, which is about 20% of upper bound manual cars interval.