Linear Regression Assignment

Zhang Jinyan

19/10/2016

Linear Regression Assignment

This is done in submission for Coursera Data Science Specialization Linear Regression Course Assignment. We are supposed to look at the relationship between types of transmission for motor cars, manual or automatic, and its corresponding fuel consumption rate measured in miles per gallon. The data set is obtained from R itself, `mtcars` data set.

First, we need to know which variables to use, and in order to do that, we need to understand the variables.

`?mtcars` brings you the help file for the `mtcars` data set, and we shall split the variables into designed input (e.g. number of cylinders) and performance output (e.g. fuel consumption).

Design Input

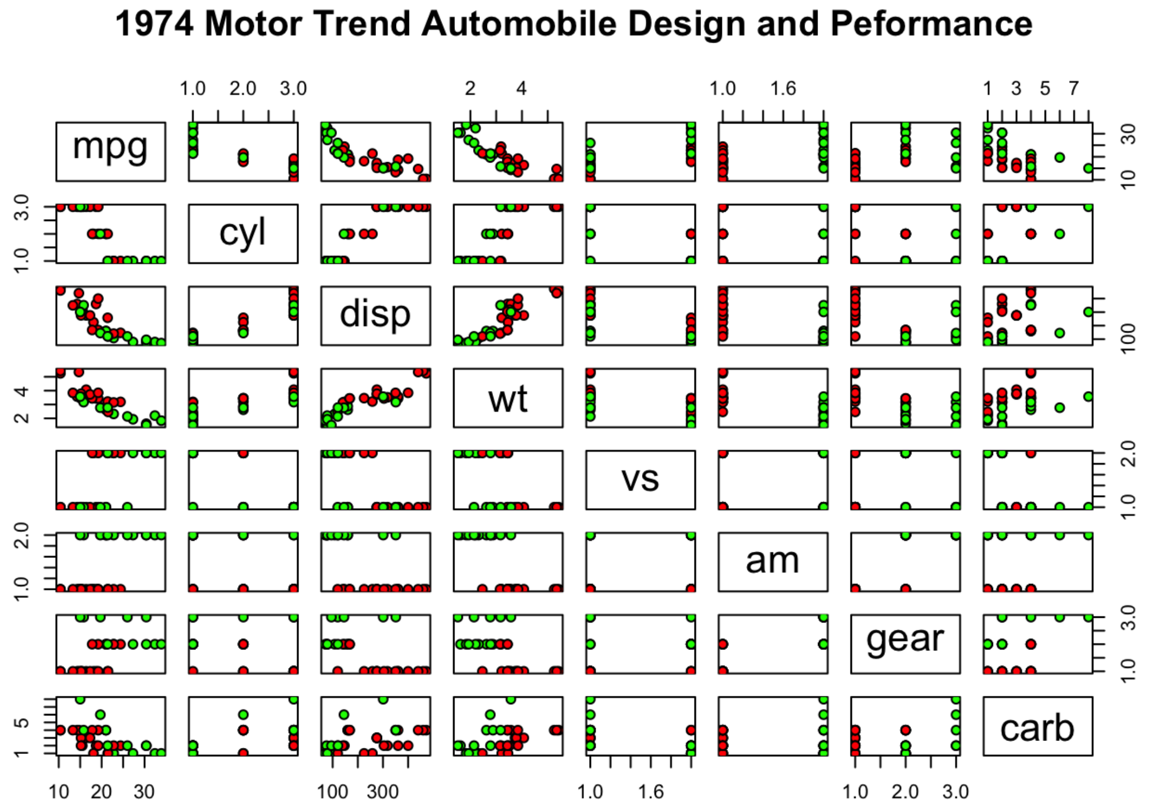
1. `cyl` = number of cylinders (part of the engine)
2. `disp` = swept volume of pistons inside cylinders (cubic inches)
3. `wt` = weight of the car (1000 lbs)
4. `vs` = presence of V engine (0 = V engine, 1 = straight engine)
5. `am` = mode of transmission (0 = automatic, 1 = manual)
6. `gear` = number of forward gears
7. `carb` = number of carburetors (a device that blends air and fuel)

Performance Output

1. `mpg` = miles travelled per gallon of petrol consumed
2. `hp` = horsepower (a measurement of rate at which work is done)
3. `drat` = ratio between driveshaft revolutions per minute
4. `qsec` = time needed to cover 1/4 mile from stop position (acceleration)

So we are essentially looking at how the car design affects the `mpg` variable, the fuel consumption rate. We'll subset out the necessary data and do necessary class changes if applicable, then we'll look at them using some graphs.

```
## 'data.frame':    32 obs. of  8 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```



In the graph above, green represents cars with manual transmission while red represents cars with automatic transmission.

```
## [1] "Correlation between disp and wt"
```

```
## [1] 0.8879799
```

```
## [1] "Number of carburetors over number of cars"
```

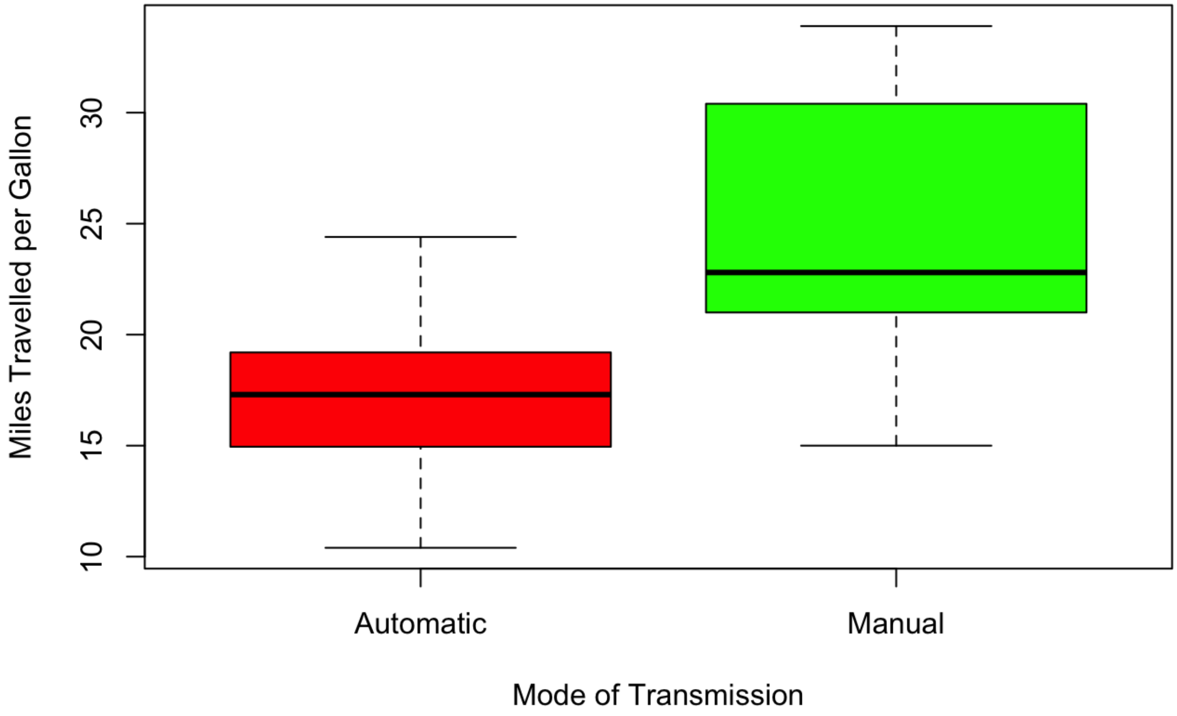
```
##
##  1  2  3  4  6  8
##  7 10  3 10  1  1
```

We will also remove the variable `disp` because it is highly correlated with `wt` at significance level of 0 (< 0.05). And lastly, we'll keep `carb` as a numeric input as there are only one data point for both 6 and 8 carburetors.

Difference in MPG by Mode of Transmission

To simply understand if there is a difference in the `mpg` values based on the mode of transmission, manual or automatic, we can do a `t.test` and a simple plot to show the difference.

```
##
##  Welch Two Sample t-test
##
## data:  manual$mpg and auto$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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```



From the boxplot, we can see a difference in median of `mpg` between the mode of transmission. From the `t.test`, the p-value 0.001 < 0.05, thus we would reject the null hypothesis that there is not a difference between the `mpg` of different mode of transmission.

Model Selection

We also wish to understand the `mpg` variable based on the rest of the variables. We will proceed to identify the design variables that affect the fuel consumption rate by backward elimination, i.e., removing the variable with the highest p-value until all p-values lie below a certain threshold. The baseline that was used for comparison is a car with automatic transmission and 4 cylinders.

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 30.9723364  3.9268509  7.8873217 5.466275e-08
## cyl6        -2.3734509  1.9004023 -1.2489203 2.242568e-01
## cyl8        -3.0519437  3.0133646 -1.0128027 3.216911e-01
## wt          -2.5698823  1.0494197 -2.4488604 2.237144e-02
## vs1          0.4737314  1.9964586  0.2372859 8.145359e-01
## amManual     1.2720877  2.0093720  0.6330773 5.329280e-01
## gear4        1.7578553  2.3217072  0.7571391 4.566535e-01
## gear5        1.1299165  2.9820578  0.3789050 7.082338e-01
## carb        -0.8245789  0.6105265 -1.3506030 1.899625e-01
```

```
##              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 31.488042  3.2056347  9.8227171 6.962839e-10
## cyl6        -2.515345  1.7680472 -1.4226686 1.677009e-01
## cyl8        -3.436577  2.4897977 -1.3802635 1.802277e-01
## wt          -2.577991  1.0280354 -2.5076867 1.932053e-02
## amManual     1.086270  1.8137503  0.5989084 5.548455e-01
## gear4        1.884244  2.2149090  0.8507093 4.033388e-01
## gear5        1.201925  2.9076678  0.4133639 6.830103e-01
## carb        -0.848118  0.5904508 -1.4363908 1.637994e-01
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 32.3753388  2.8764732 11.255220 1.710687e-11
## cyl6        -3.0493364  1.5743292 -1.936911 6.369383e-02
## cyl8        -4.9313482  1.7946645 -2.747783 1.075615e-02
## wt          -2.5550551  0.9616660 -2.656905 1.330032e-02
## amManual     1.6066292  1.5687011  1.024178 3.151864e-01
## carb        -0.6729417  0.4278906 -1.572696 1.278807e-01
```

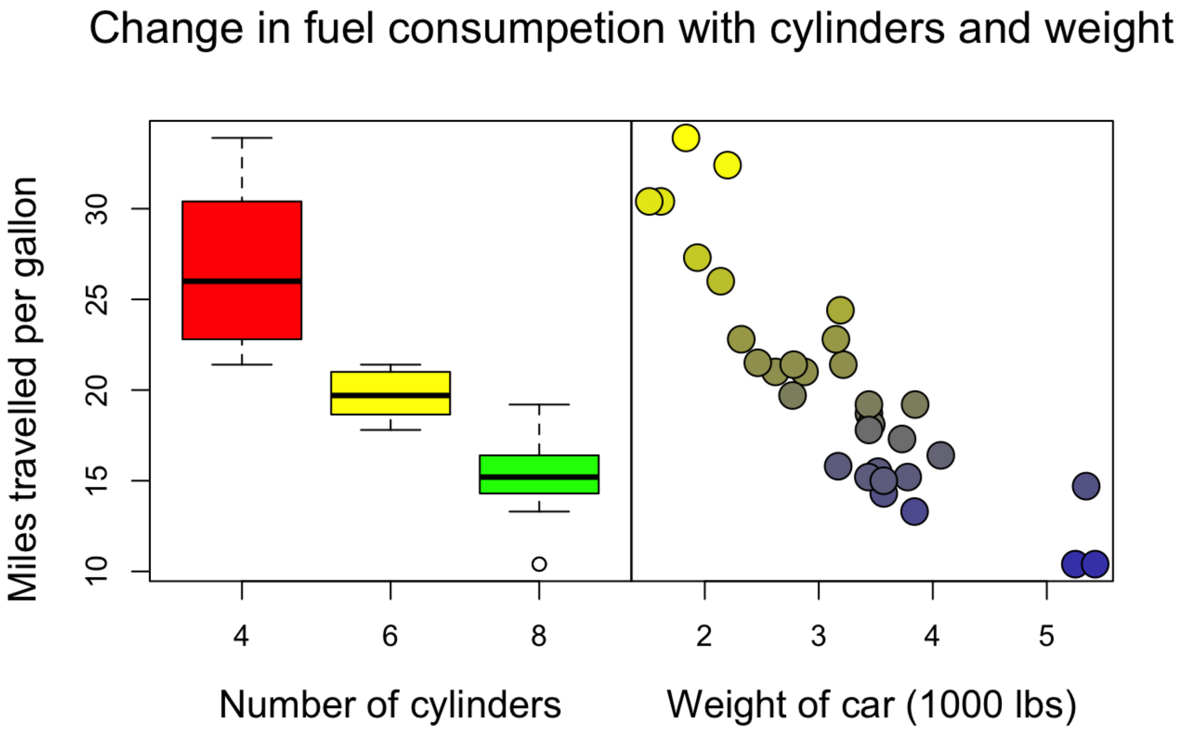
At this point, the p-value for `am` variable is the highest at 0.315, and it is more than the threshold at `0.05`. This means that setting all remaining variables constant, there is no difference between the `mpg` of car with automatic or manual transmission.

This is in contrast with the results we got from the `t.test`. One reason for this is that there might be strong correlations between the different variables with the mode of transmission. The correlations will affect the coefficients of the variables, thus affecting the significance of that variable.

However, if we were to continue with the model selection, we will end up with only two variables, `cyl` and `wt` as the variables affecting the `mpg`.

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.5591960  1.9323926 17.884148 1.699028e-16
## cyl6        -3.5016653  1.5124802 -2.315181 2.844232e-02
## cyl8        -5.3150437  1.7567160 -3.025557 5.396269e-03
## wt          -3.1742236  0.7485537 -4.240475 2.336223e-04
## carb        -0.4142148  0.3456735 -1.198283 2.412195e-01
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.990794  1.8877934 18.005569 6.257246e-17
## cyl6        -4.255582  1.3860728 -3.070244 4.717834e-03
## cyl8        -6.070860  1.6522878 -3.674214 9.991893e-04
## wt          -3.205613  0.7538957 -4.252065 2.130435e-04
```



Interpreting the Coefficients

The coefficient for `cyl6` means the average `mpg` value will change by -4.256 when the number of cylinders change from 4 to 6 with a 95% confidence interval of -7.1, -1.412.

Whereas, the `mpg` value will change by -6.071 when the number of cylinders change from 4 to 8 with a 95% confidence interval of -9.461, -2.681.

The coefficient for `wt` means the average `mpg` value will change by -3.206 when the weight of car decreases by 1000 lbs with a confidence interval of -4.752, -1.659.

Also note that since all the values in the confidence intervals are negative, it corresponds to the p-value and that there is a true decrease in `mpg` values when the performance variable changes.