Regression-Models-Assignment

MengNan

Thursday, July 24, 2014

This is an assignment for Regression Models Class. This work is for Motor Trend, a magazine about the automobile industry who are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome), and mainly answer the following two questions:

- "Is an automatic or manual transmission better for MPG"?
- "Quantify the MPG difference between automatic and manual transmissions"

To answer these two questions, this paper first extract "useful" features(features may have influence on mpg) from the dataset mtcars. By calculating the empirical covariance between mpg and other features, and order the result, empirical covariances of all features which are larger than 0.68 have been picked. They are: cyl,disp,hp,drat and wt. Then, model mpg with all the other features respectively by linear model and get five residuals. I order the residuals to decide which has the best linear relation with mpg. Next, choose one of the remaining features(features have the least residuals) and include it in the linear model. If the new model have less residuals and standard errors, then accept it, otherwise drop it and try another one. Finally, try to find high dimensional relationship among features. I eventually choose the linear model lm(mpg ~ wt + cyl + hp + disp + hp*disp) to predict the mpg of both automatic cars and manual cars and find that: manual car have a higher average level of mpg but also a larger variance(or standard deviation). So under my preliminary analysis, I provide some suggestions:

- If you are a manager or a decision maker of a car factory, you should probably choose manual transmission of cars, because you will get more profits(for manual cars higher empirical mean).
- If you are a home keeper or a housewife, you should chooes to buy an automatic car, because it can have higher chance helping you save money(for automatic cars have less variance)

Univariate analysis

In this part, I use linear model to model the feature mpg and all the other features respectively and get the residual sum of squares and covariation

```
##            sumresid_cyl sumresid_disp sumresid_hp sumresid_drat
## sumresid       308.3342      317.1587    447.6743      603.5667
## covariance      -0.8522       -0.8476     -0.7762        0.6812
##            sumresid_wt sumresid_qsec sumresid_vs sumresid_am sumresid_gear
## sumresid      278.3219      928.6553     629.519    720.8966      866.2980
## covariance     -0.8677        0.4187       0.664      0.5998        0.4803
##            sumresid_carb
## sumresid        784.2711
## covariance       -0.5509
```

As you can see the five feature have both least residual and largest absolute value of covariance. So I choose the feature "wt" as the original linear model feature. Here is the fitted model's information:

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   37.285     1.8776  19.858 8.242e-19
## mtcars$wt     -5.344     0.5591  -9.559 1.294e-10
```

So, the intercept is 37.285 and the coefficient of "wt" in the fitted model is -5.344. So we can see a negative relation between mpg and weight("wt" stands for weight).

The confidence interval of weight with 95% confidence is

```
## [1] 33.45 41.12
```

So with 95% confidence, we estimate that a 0.001 bl decrease in weight results in a 33.45 to 41.12 increase in mpg(mile/gallon).

```
##   sumresid_wt  sumresid_cyl sumresid_disp   sumresid_hp sumresid_drat
##         278.3         308.3         317.2         447.7         603.6
##   sumresid_vs   sumresid_am sumresid_carb sumresid_gear sumresid_qsec
##         629.5         720.9         784.3         866.3         928.7

##   rho_wt  rho_cyl rho_disp   rho_hp rho_drat
##  -0.8677  -0.8522  -0.8476  -0.7762   0.6812
```

Multivariate analysis

As you can see, a feature has already been chosen to the linear model, but did the other features have no linear relationship with mpg? Probably not! So I ordered the residuals and chose another four least residuals corresponding to features :"cyl" ,"disp", "hp" and "drat". They were included in this linear model one by one in the increasing order of residuals. The first included feature is "cyl", because including it results in largest amount of decrese of standard deviation. Feature "hp" is next, then feature "disp" and the multiply of "disp" and "hp". The standard deviations of different number of features included are present in Table1. Here is the **coefficients(interception)** of my final model:

```
##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 44.4642648  2.796e+00 15.903463 6.507e-15
## wt          -3.4325272  9.218e-01 -3.723533 9.574e-04
## cyl          0.0044802  7.511e-01  0.005965 9.953e-01
## hp          -0.0958736  2.928e-02 -3.274348 2.994e-03
## disp        -0.0422844  2.210e-02 -1.913476 6.676e-02
## hp:disp      0.0002592  9.355e-05  2.771161 1.018e-02
```

So we can see that the intercept is 44.46 and coefficients of every features are: wt(-3.43), cyl(0.0045), hp(-0.095), disp(-0.042), hp:disp(0.00026). So the linear model is: $mpg = 44.64 - 3.43wt + 0.0045cyl - 0.096hp - 0.042disp + 0.00026hp * disp$.

As the last result show, this is the best model we have made so far from now. And also let's do some confidence analysis. The confidence interval of all the feature in the Table2.

The confi_intvals2 to confi_intvals5 represent confidence intervals of different feature used to model the mpg confi_intval1 represent confidence interval of intercept. From the residuals figure(Figure 3) we can see that the distribution of residuals is like Gaussian distribution which means that our model is good. So I use the model to predict the mpg both of automatic and manual cars and get the results:

```
##            Mean Variance
## Manual    24.32    29.53
## Automatic 17.20    13.79
```

So, it is clear that manual cars have a higher average level of mpg but also a larger variance. If you need a large number of cars, you may choose manual ones but if you just need a singal one or a little number of car, you are suggested to choose automatic ones...

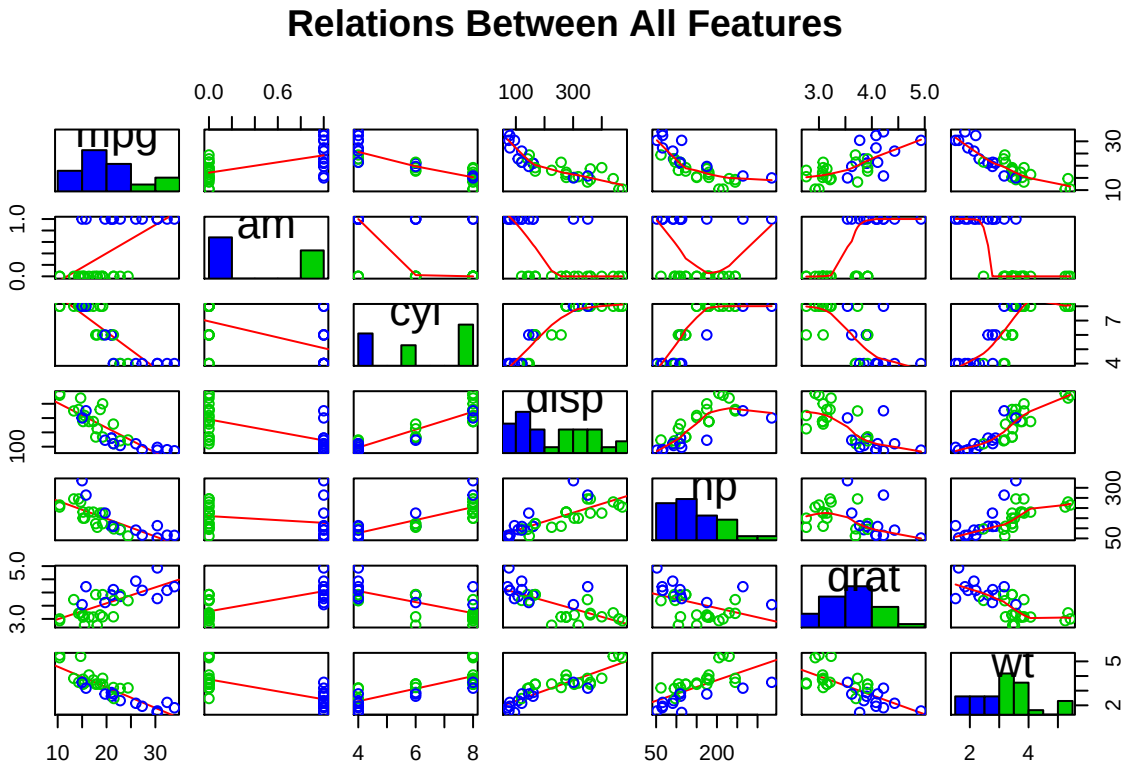
Appendix

* Table1 The standard deviations of different features

```
##                     [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## sumresid_result 191.172 246.683 195.048 269.241 188.492 176.621 170.444
## sd_result         2.483   2.821   2.508   2.947   2.466   2.387   2.345
##                     [,8]   [,9]
## sumresid_result 167.426 131.58
## sd_result         2.324   2.06
```

* Table2 Confidence Intervals of model lm(mpg ~ wt + cyl + hp + disp + hp:disp)

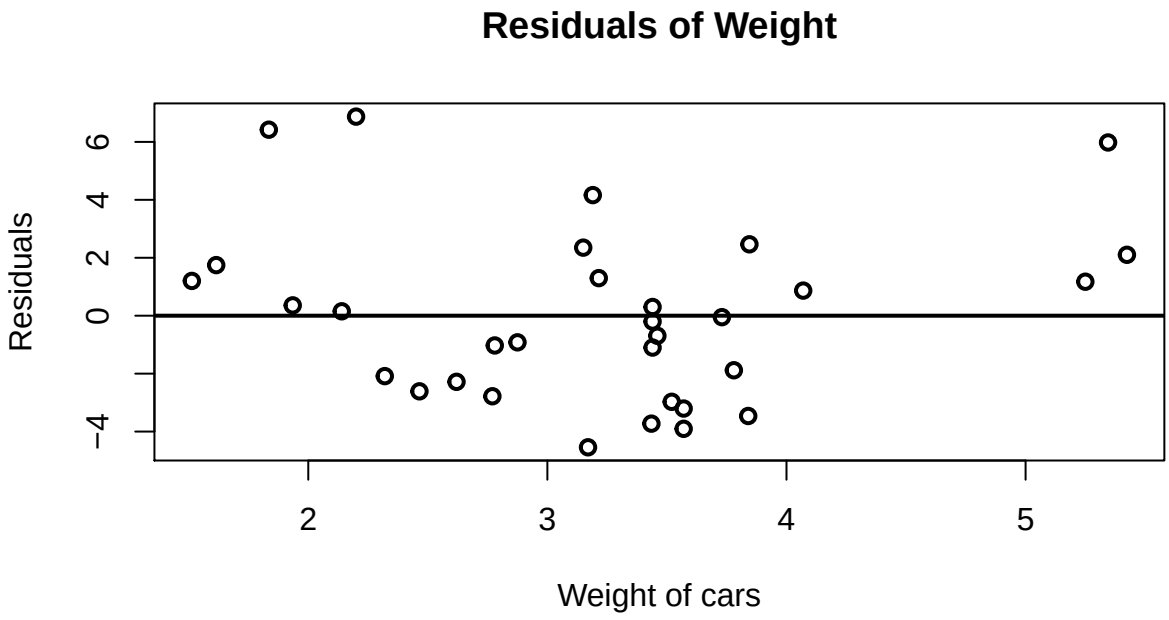
```
##   confi_intvals1 confi_intvals2 confi_intvals3 confi_intvals4
## 1          38.72         -5.327         -1.539       -0.15606
## 2          50.21         -1.538          1.548       -0.03569
##   confi_intvals5 confi_intvals6
## 1      -0.087708      6.694e-05
## 2       0.003139      4.515e-04
```



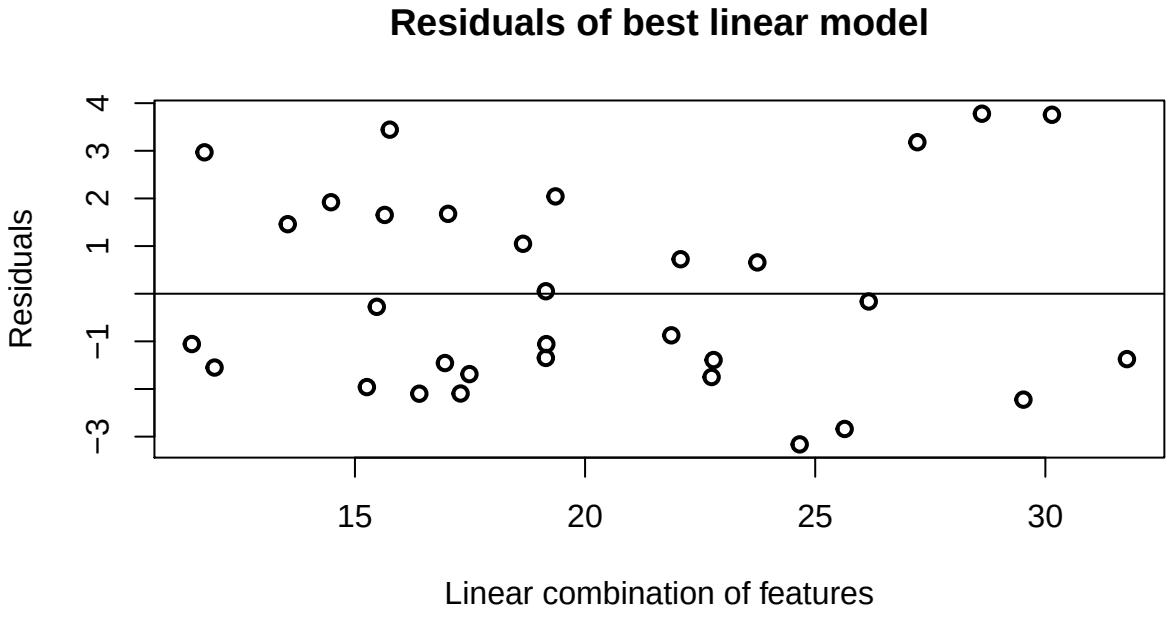
* Figure1 Relation between every two features

From the pairs of figures, we can see the # from the picture we can see that "cyl", "disp", "hp", and "wt" are negatively correlated with "mpg". "drat" is positively correlated with "mpg". We

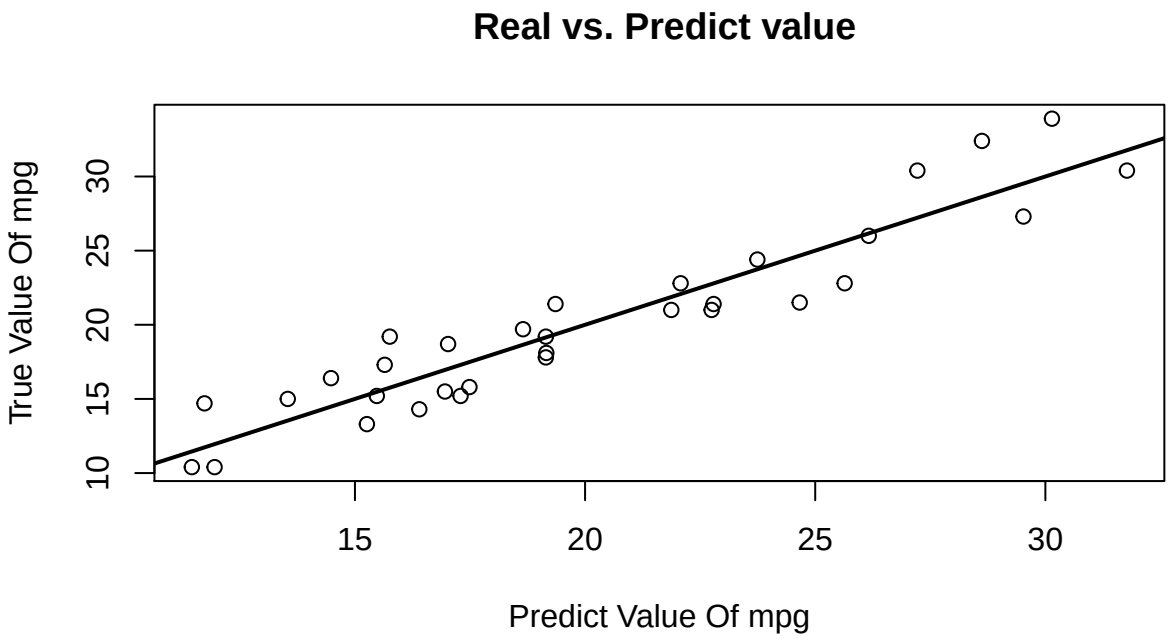
can also notice that blue points which represent manual cars have generally higher mpg than green points which represent automatic cars and we can also see that the “drat” variance distribution is like Gaussion distribution which means that it might have little relationship with others “disp” and “wt” might have linear relation and “hp” and “wt” might have linear relation.



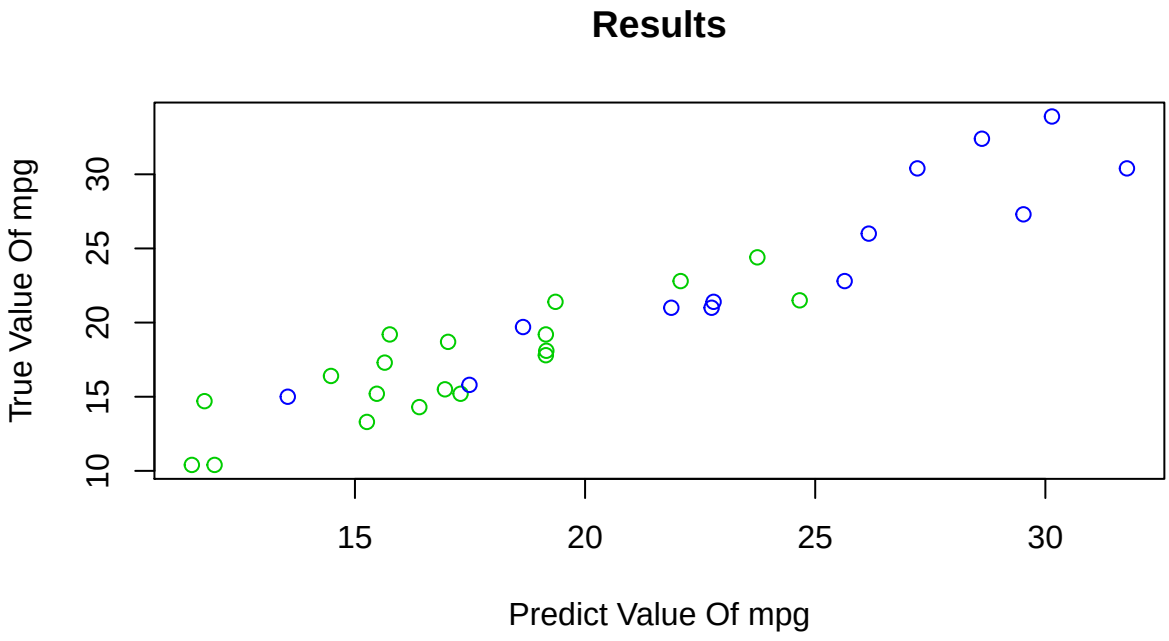
* Figure2 The Residuals of linear model between "mpg" and "wt"



* Figure3 The Residuals of the linear model mpg ~ wt + cyl + hp + disp + hp:disp



* Figure4 The Real and Predict value of mpg



* Figure5 Results

Blue points represent the manual cars and the green points represents the automatic cars