Developing Data Products Demo

freestander

Saturday, October 24, 2015

Executive Summary

We want to build a model to predict the miles per gallon (MPG) by using a set of car features.

There is an "mtcars" dataset in R package. 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).

Exploratory Data Analysis

First, we review each field in the dataset using the summary function and also draw a pariwise scattor plot between the variables (shown in appendix).

```
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec       vs             am     gear   carb  
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7  
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10  
##  Median :3.325   Median :17.71                         5: 5   3: 3  
##  Mean   :3.217   Mean   :17.85                                4:10  
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1  
##  Max.   :5.424   Max.   :22.90                                8: 1  
```

Data Preprocessing

Next, we transform some categorical variables into factor types to prepare for the regression analysis in the later steps.

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels=c('Automatic','Manual'))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Regression Model

We will use regression model to identify the variables that account for MPG differences.

First, we try the multivariate linear regression with all variables included. The p-value is 0.000124, and the adjusted R-squared is 0.779, which means the model accounts for 77.9% variance. From the coefficients below, we can see some of the variables have insignificant p-value thus may bring noise to the model if included in the regression.

```
m1 <- lm(mpg ~ ., data = mtcars)
summary(m1)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Next, we perform a stepwise model selection using backward elimination.

```
m2 <- step(m1, direction="backward")
```

The remaining variables ("cyl", "hp", "wt", "am") are significant and fit the model best. The p-value is 1.506e-10, and the adjusted R-squared is 0.8401, which means the model accounts for 84.01% variance. Compared to automatic transmission, MPG increases by 1.8 if having a manual transmission. Moreover, the regression result shows MPG decreases -3.03 for "cyl6", -2.16 for "cyl8", -0.03 for "hp", -2.5 for "wt", respectively.

```r
summary(m2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

We also do the residual plots (shown in appendix) to check how well the regression model fits. The “Residuals vs Fitted” plot shows no pattern, supporting the independence assumption. The “Normal Q-Q” plot shows that residuals can be approximated by normal distribution. The “Scale-Location” plot shows that the points are randomly distributed, supporting constant variance assumption. The “Residuals vs Leverage” plot shows that no particular outlier is observed.

Conclusion

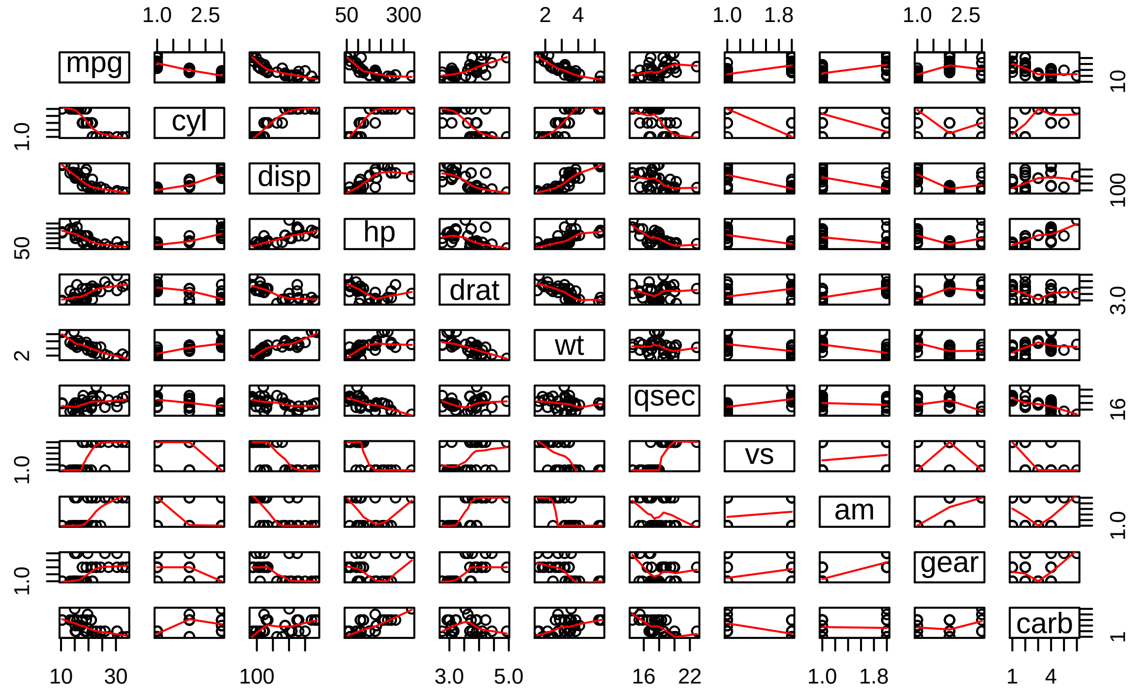
We can see that variables (“cyl”, “hp”, “wt”, “am”) fit the regression model best and and count for most MPG differences.

However, there are some limitations that we need to address to further improve the analysis result. For example, the residual plots show some transformation of the variables are needed to achieve linearity, and the sample size is too small (only 32 records) to arrive at a reliable conclusion.

Appendix

The following is a pairwise scattor plot between different variables.

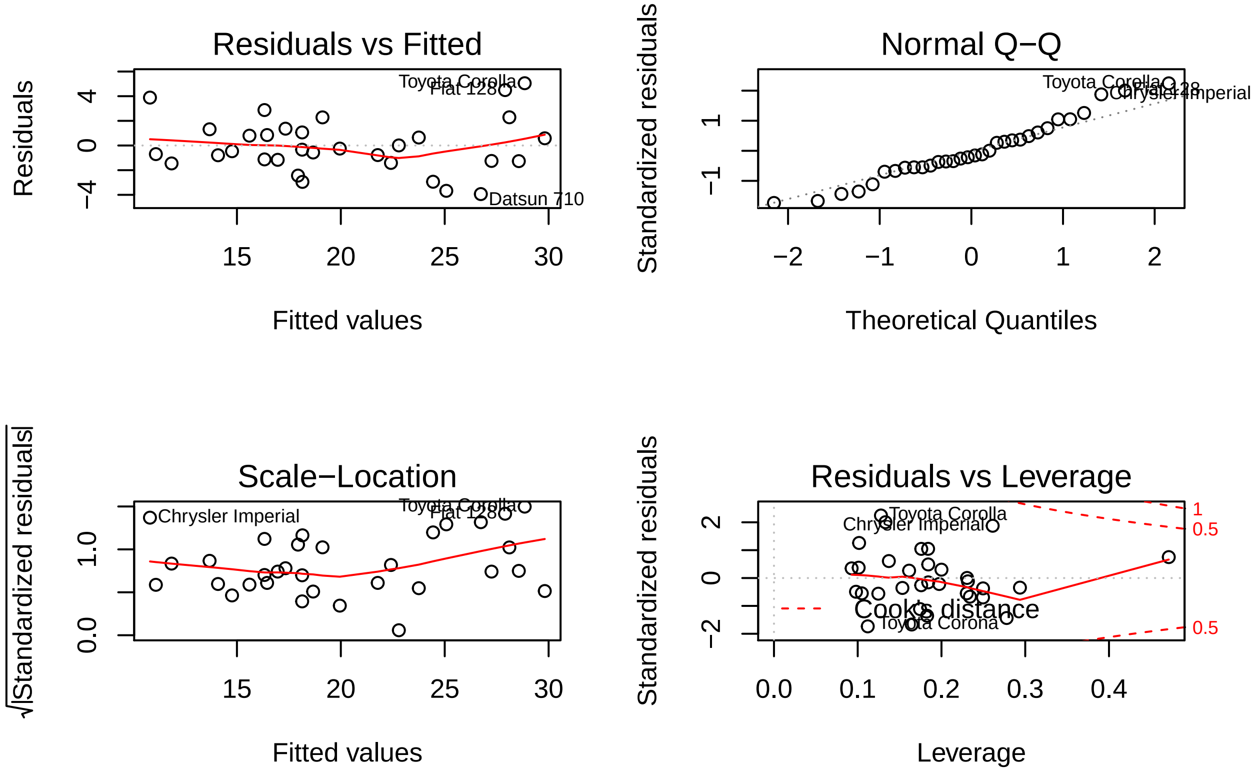
```r
pairs(mtcars, panel=panel.smooth, main="Pairwise Scattor Plot of Motor Trend Car Road Tests")
```



Pairwise Scatter Plot of Motor Trend Car Road Tests

The following are the residual plots to check how well the regression model fits.

```r
par(mfrow = c(2, 2))
plot(m2)
```



Source

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