A study of the auto manufacturers' returns with regression analysis

Junfeng Li[1,2]
(1.School of Mathematics and Statistics； 2.Student ID:1120132819)

15th.MAY.2016

Abstract

The purpose of this paper is to study the quantity relationship of weekly log returns among Toyota corp.,Ford corp.,and GM.By analyzing the data, we construct a multiple linear regression model whose dependent variable is the log returns of GM.Then,we do model diagnostic and test in R and get the final model using stepwise regression after removing the outliers.The results show that the weekly log returns of Ford is significant in the regression model, but its goodness of fit is only 0.359.

Key words:multiple linear regression model;R

1 Introduction

In statistics,linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X.For more than one explanatory variable, the process is called multiple linear regression.

In linear regression,the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

2 Model Building

Multiple linear model:

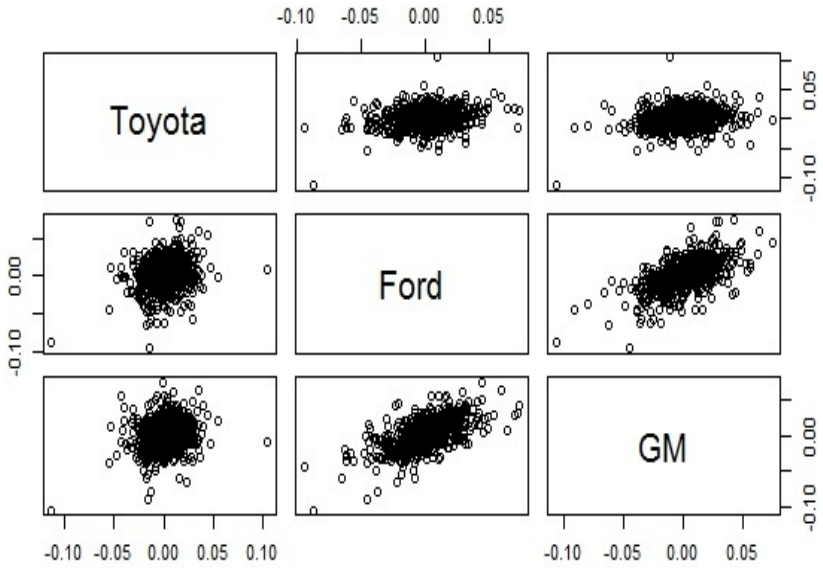
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

y - dependent variable; x_i - explanatory variables; $\beta_0, \ldots, \beta_p$ - parameters in the linear regression model; ε - random error.

In this study,we choose the weekly log returns of GM to be the dependent variable,and treat the weekly log returns of Toyota and Ford as predictors.Sample capacity is 709.

3 Regression analysis and model diagnostics

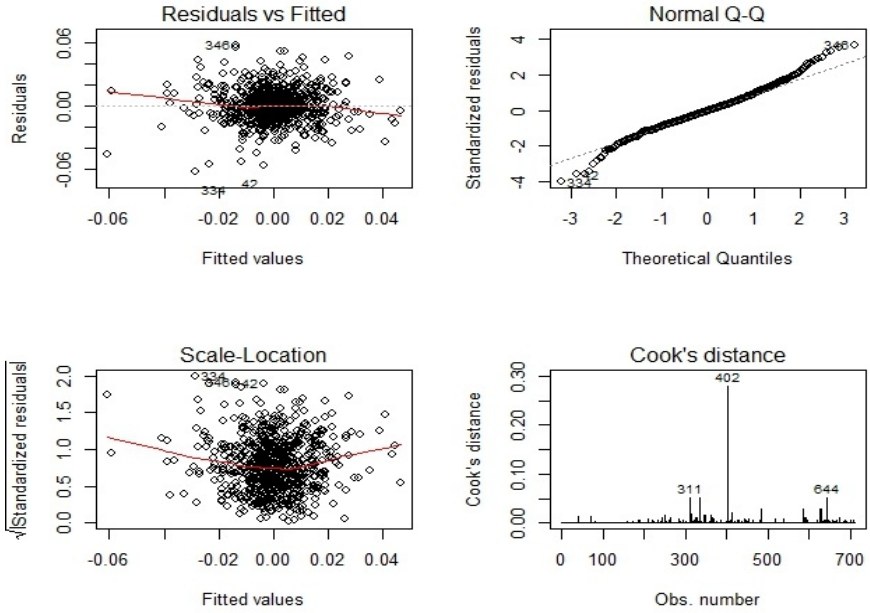
First of all,we use R to get the scatterplot matrix among the weekly log returns of Toyota corp.,Ford corp.,and GM:



From the scatterplot above,we can find that there is an obvious linear relationship between the weekly log returns of Ford corp.and GM.However,it is not conspicuous between Toyota corp.and GM.

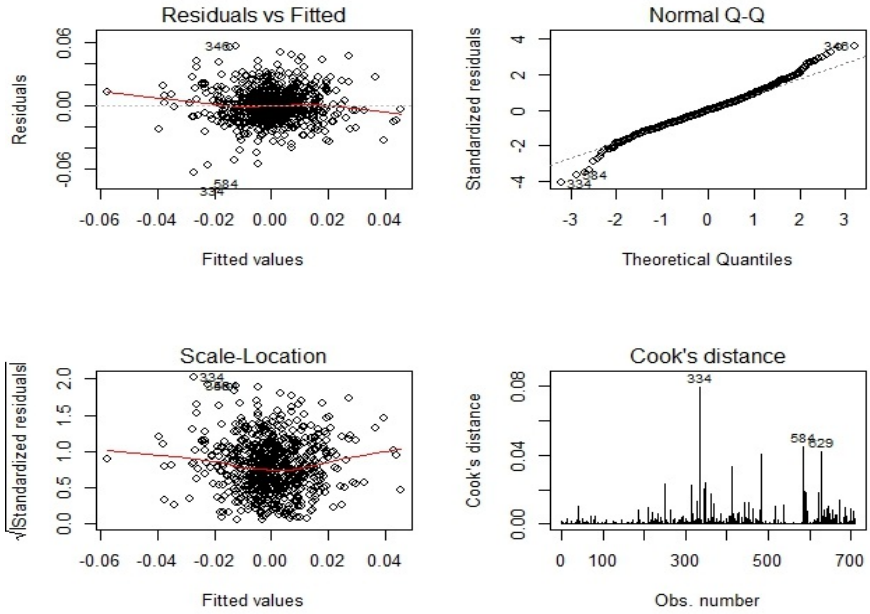
In model diagnostic,it shows that the multiple R-squared of this model is 0.3775 and the adjusted R-squared is 0.3757 which means the explanatory variables can just response 37.5% of dependent variable.And,the F-statistic shows that the regression equation is significant.The t-value of β_2 is 19.619,but the t-value of β_1 is 1.621 which can not pass t-test.Then,we got dw-statistic is 2.0158.It's near 2 with a very close distance,so we consider that there is no autocorrelation in the model.Meanwhile,the multicollinearity test by condition number passed.(More information in Appendix.)

On the other hand,we need to check whether there is a heteroscedasticity or outliers,and the random error also need to be normal.



From the Q-Q plot,we could consider that the random error is normally distributed.But,there is a outlier seeing in the cook's distance figure.

Finally,we use the method of stepwise regression after removing the outliers and got the following result.It removes one explanatory variable and remain weekly log returns of Ford.In model diagnostic,it shows that the multiple R-squared of this model is 0.359 which means the weekly log returns of Ford can response 35.9% of he weekly log returns of GM.And,both the F-test and t-test passed.Residual plot and Q-Q plot are given below.



4 conclusion

Through the above analysis,we can know that multiple linear regression model is not a perfect model to study the quantity relationship of weekly log returns among Toyota corp.,Ford corp.,and GM.Although there seems to be a linear relationship between Ford corp.,and GM,the multiple R-squared of this model can just add up to 0.359.So,it may be a bad idea to utilize linear regression in this study.

Appendix

```
> D<-read.csv("E:/课件/统计计算/w_logret.csv",header=T)#录入数据
> pairs(D)                                                    #scatterplot matrix
> lm1=lm(GM~Toyota+Ford,data=D)
> summary(lm1)                                               #multiple linear regression analysis

Call:
lm(formula = GM ~ Toyota + Ford, data = D)

Residuals:
      Min        1Q    Median        3Q       Max
-0.062848 -0.009649 -0.000405  0.008977  0.057515

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.049e-05  5.914e-04   0.119    0.905
Toyota      6.132e-02  3.784e-02   1.621    0.106
Ford        6.145e-01  3.132e-02  19.619   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01572 on 706 degrees of freedom
Multiple R-squared:  0.3775,    Adjusted R-squared:  0.3757
F-statistic: 214.1 on 2 and 706 DF,  p-value: < 2.2e-16

> library(lmtest)
> dwtest(lm1)                                                #DW-test

        Durbin-Watson test

data:  lm1
DW = 2.0158, p-value = 0.586
alternative hypothesis: true autocorrelation is greater than 0

> kappa(D)                                                   #multicollinearity-test
[1] 2.490097
> par(mfrow=c(2,2))
> plot(lm1,which=c(1:4))        #check whether there is a heteroscedasticity or outliers,and the
                                                             random error also need to be normal.
> D=D[-402,]
> lm2=lm(GM~Toyota+Ford,data=D)
> summary(lm2)
```

```
Call:
lm(formula = GM ~ Toyota + Ford, data = D)

Residuals:
      Min       1Q   Median       3Q      Max
-0.063880 -0.009832 -0.000253  0.008723  0.056845

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0001656  0.0005888   0.281    0.779
Toyota      0.0335251  0.0387131   0.866    0.387
Ford        0.6038200  0.0313369  19.269   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01563 on 705 degrees of freedom
Multiple R-squared:  0.3597,    Adjusted R-squared:  0.3579
F-statistic:   198 on 2 and 705 DF,  p-value: < 2.2e-16

> plot(lm2,which=c(1:4))                    #remove the outliers and redo regression analysis
> lm.aic=step(lm2,trace=F)
> summary(lm.aic)

Call:
lm(formula = GM ~ Ford, data = D)

Residuals:
      Min       1Q   Median       3Q      Max
-0.064212 -0.009971 -0.000120  0.008845  0.056505

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0001990  0.0005874   0.339    0.735
Ford        0.6094575  0.0306479  19.886   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01563 on 706 degrees of freedom
Multiple R-squared:  0.359,  Adjusted R-squared:  0.3581
F-statistic: 395.4 on 1 and 706 DF,  p-value: < 2.2e-16

> plot(lm.aic,which=c(1:4))                 #Screen explanatory variables using stepwise regression
```