ISyE 6416 Homework 1: Regression in R

Linear regression analysis and model selection

Shixiang Zhu
Georgia Institute of Technology
Atlanta, Georgia
shixiang.zhu@gatech.edu

ABSTRACT

For this analysis, the aim of linear regression is to model the relationship between three auto manufacturers: Toyata Motor Corp., Ford Motor Corp., and GM. Treat the log returns of GM as response and log returns of Toyota and Ford as predictors. We fit a linear regression model in R and perform necessary model diagnostics. By analyzing the given dataset, we draw the conclusion on whether linear regression is a useful tool for this dataset and find out which predicotrs are playing important roles in this relationship.

KEYWORDS

linear regression, model selection, log returns

1 INTRODUCTION

In this report, we build a linear regression model for the given dataset to help a banker use the log returns of Toyota and Ford to interpret the log returns of GM. First we evaluate the statistical significance for Linear regression is used to predict the value of an outcome variable (response) based on one or more input variables (predictor). The goal of linear regression model is to establish a linear relationship between predictors and responses.

2 PROBLEM FORMULATION

The aim of linear regression is to model the log returns of GM. Y as a function of the log returns of Toyota X_1 and the log returns of Ford X_2. The function of their relationship can be defined as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where β_0 is the intercept and β_1, β_2 are the regression coefficients of X_1, X_2 respectively. ϵ is the error term.

By definition of linear regression, the objective function for this problem is:

$$\min_{\beta} \|Y - X\beta\|_2^2 \iff \min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

3 MODEL ESTIMATION

A large number of procedures have been developed for parameter estimation inference in linear regression. Those methods are different in some aspects like computational simplicity, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, etc. In this report, we apply the simplest and thus most common estimator to our linear regression model, ordinary least squares (OLS), which is the exact way how we estimate parameters in R.

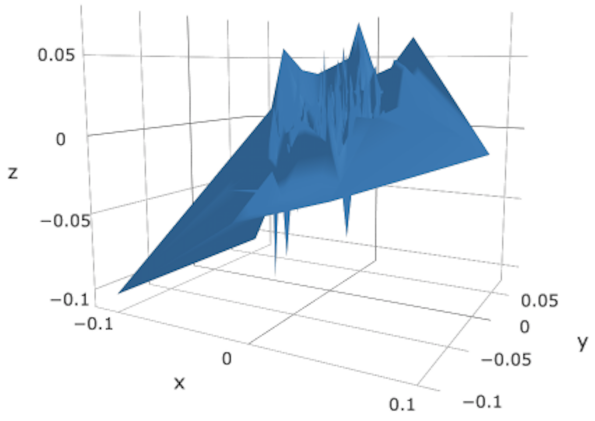

Figure 1: Surface of GM. vs. Toyota Motor Corp. and Ford Motor Corp.

Table 1: Summary of raw data

	Toyota Motor Corp.	Ford Motor Corp.	GM.
Min.	-0.1132272	-9.481e-02	*-0.1061545*
1st Qu.	-0.0084958	-9.368e-03	*-0.0119514*
Median	0.0005180	-2.082e-04	*0.0004089*
Mean	0.0008693	7.535e-05	*0.0001701*
3rd Qu.	0.0105164	1.122e-02	0.0116914
Max.	0.1049635	7.429e-02	0.0744415

The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown parameter β:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = (\sum \mathbf{x}_i \mathbf{x}_i^\top)^{-1} (\sum \mathbf{x}_i y_i)$$

In addition, the OLS estimaor is also identical to the maximum likelihood estimator (MLE) under the normality assumption for the error terms [?] (i.e. if the errors are i.i.d normal distribution $\mathcal{N}(0, \sigma^2)$.)

4 EXPERIMENTS

In experiments, we use the dataset provided on class that comes with R. Before building the linear regression model, we start to take a look at the raw dataset first to analyze and understand the variables intuitively. First of all, Surface plot can help visualize the linear relationships between the response variables and predictor

Table 2: Summary of linear regression

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
'Ford Motor Corp.'	6.145e-01	3.132e-02	19.619	<2e-16
'Toyota Motor Corp.'	6.132e-02	3.784e-02	1.621	0.106
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	

variables. Ideally, in linear regression, their relationship can be model as a plane in this space. As shown in Fig. 1, x-axis and y-axis mean the log returns of Toyota Motor Corp. and Ford Motor Corp. respectively, z-axis is the the log returns of GM. Besides graphical analysis, we can also study their statistical property. As reported in Table 1, the log returns of GM. seems to be highly correlated with Toyota.

4.1 Linear Regression Diagnostics

Before actually using linear regression model, we need to ensure it is statistically significant. Since a larger t-value indicates it is less likely that the coefficient is not equal to zero purely by chance and the corresponding p-value would be relatively low, which means the coefficients are significant. As reported in Table 2, the p-value of Ford Motor Corp. is significantly lower than 0.05, we can safely reject the null hypothesis that coefficient β_2 of the predictor is zero. However, the p-value of Toyota Motor Corp. is slightly larger than 0.05, which means the log returns of Toyota Motor Corp. is not statistically significant to the log returns of GM. As to R-Squared and Adj R-Squared value, they are kind of lower than empirical value 0.70 (actually R-Squared is not a good criterion, since it always increase with the model size).

In additin, we also check the diagnostic plots for further analysis of this regression model, in particular, they show residuals in four different ways: `1. Normal Q-Q plot`: Fig. 2 shows residuals follow a straight line well when residuals lie between -2 to 2, within this range, the residuals perfectly matched normal distribution. `2. Residuals vs Fitted plot`: Fig. 3 shows there is no obvious non-linear pattern in residuals, which is a good indication to our linear regression model. `3. Residuals vs Leverage plot`: Fig. 4 is the typical look when there is no influential case, or cases. we can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines. `4. Scale vs Location plot`: as shown in Fig. 5, we can tell there is a parabola (the red line) clearly. Residuals are spread inequally along the ranges of predictors, it begins to spread wider along the x-axis as it passes to two end.

4.2 Model Selection

Generally speaking, the linear regression model we built is fairly good in terms of the results of diagnostics. But we can still try to improve it by model selection. As we mentioned above, we can remove the predictor of Toyota Motor Corp. from the model according to their p-values. The rebuilt model shows a minor improvement on BIC and AIC criterion and other statistics, however, on the whole, their performance on prediction or interpolation are essentially the same. We can draw the conclusion that our first linear regression model can be utilized to model the log returns of these three motor companies, and the model can gain a small benefit by removing the predictor of Toyota Motor Corps, but it is not required.

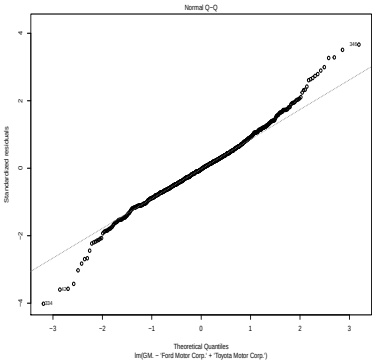

Figure 2: Normal Q-Q Plot

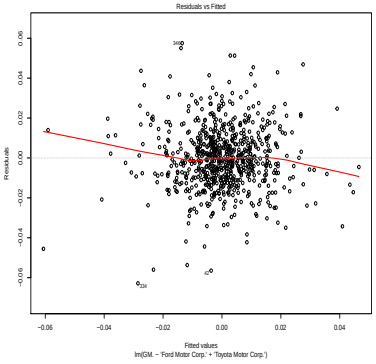

Figure 3: Residuals-Fitted Plot

5 CONCLUSIONS

In this report, we build a linear regression model for three motor companies and briefly analyze the diagnostics of the model. In the basis of the results of the experiments, we reach the conclusion that it is a good model for intercepting the log returns of GM. by using the log returns of Toyota and Ford. We can also remove the predictor of Toyota to improve our model a little bit.

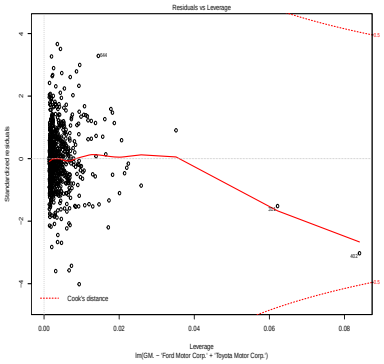

Figure 4: Residuals-Leverage Plot

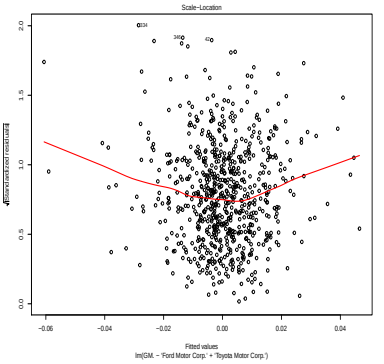

Figure 5: Scale-Location Plot

A CODES FOR THE EXPERIMENTS

```
# Load data
filePath <- "/Users/woodie/Documents/Courses/ISyE␣6416␣Computational␣Statistics␣(Spring␣2018)/HW/ISYE-6416/hw1/w_logr
rawdata  <- read.csv(file=filePath, header=FALSE, sep=",")
colnames(rawdata) <- c("Toyota␣Motor␣Corp.", "Ford␣Motor␣Corp.", "GM.")
summary(rawdata)

# Plot the 3d surface for response (GM.)
plot_ly(x=rawdata$`Toyota Motor Corp.`,
        y=rawdata$`Ford Motor Corp.`,
        z=rawdata$`GM.`, type="mesh3d")

# Linear regression
lmResult1 <- lm(`GM.`~`Ford Motor Corp.`+`Toyota Motor Corp.`,
                data=rawdata)
summary(lmResult1)

# Model Selection
lmResult2 <- lm(`GM.`~`Ford Motor Corp.`,
                data=rawdata)
summary(lmResult2)

lmResult3 <- lm(`GM.`~`Toyota Motor Corp.`,
                data=rawdata)
summary(lmResult3)

# Other diagnostics analysis
AIC(lmResult1)
BIC(lmResult1)
```