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

1 INTRODUCTION

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. In this report,

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.

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 $\mathcal{N}(0, \sigma^2)$.)

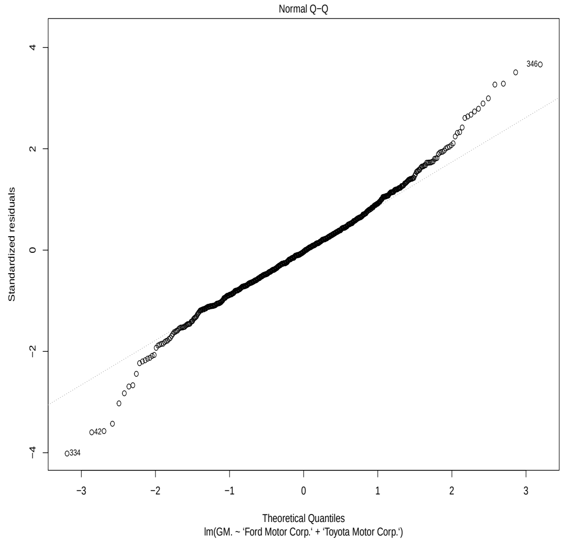

Figure 1: Normal Q-Q Plot

4 EXPERIMENTS

In experiments, we use the dataset provided on class that comes with R. Before begin building the linear regression model, we start to take a look at the raw dataset first to analyze and understand the variables intuitively.

5 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

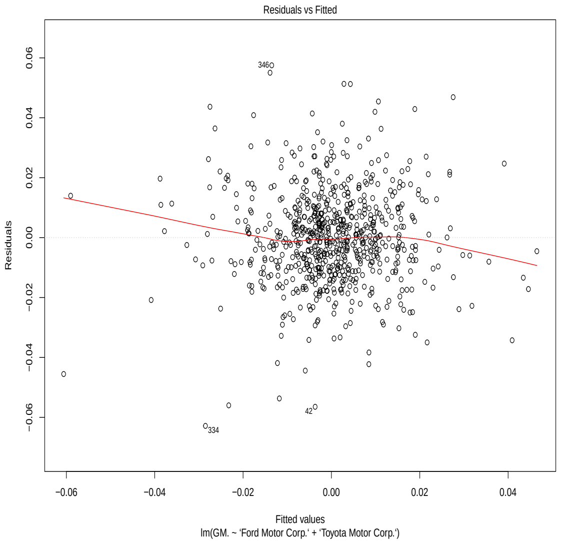

Figure 2: Residuals-Fitted Plot

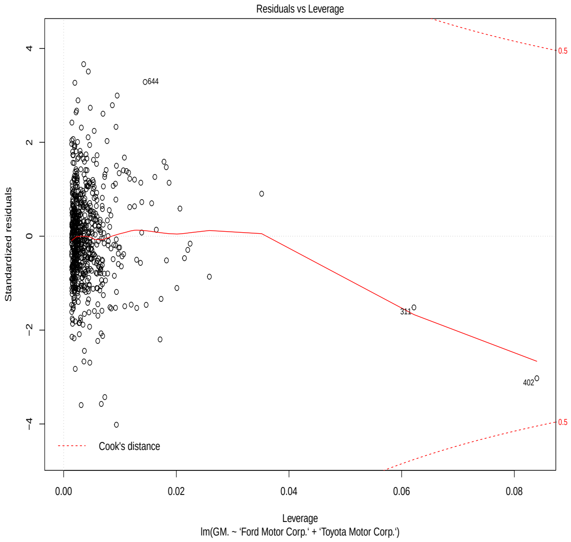

Figure 3: Residuals-Leverage Plot

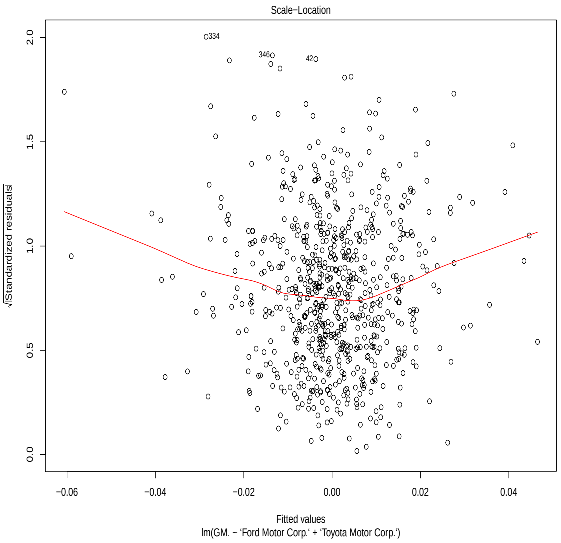

Figure 4: Scale-Location Plot