ST599 Group Project #3

Spring 2015

Jeremiah Cloud

Vanessa Lepe

Faraz Niyaghi

Kate Williams

Data

The data for this project came from “QuantQuote Historical Data”. This contains data on all National Association of Securities Dealers Automated Quotations (NASDAQ), New York Stock Exchange (NYSE), and AMEX securities from 1998 to the present. The data consists of six variables: date, open, high, low, close and volume. Date was provided in an integer format where 20100527 represents May 27th, 2010. Open is the opening price for that company for that day. High is the high price for that company for that day. Low is the low price for that day and close is the closing price for that day. There were originally 500 excel files for 500 separate companies.

To ease the analysis of the data, we combined all 500 excel files for the separate companies into one data file and added company name as a variable. We used the “lubridate” package in R to change the date variable from a string of integers to the form 2010-05-27.

Questions

Can we get reasonable predictions for average weekly closing stock price based off the closing stock prices of the previous 10 weeks of a given company?

Which company predictions had the lowest MSPE using ridge regression on the previous 10 weeks of daily closing stock price?

Does differencing using first order random walk help reduce season trends and give better predictions?

Developing Predictors

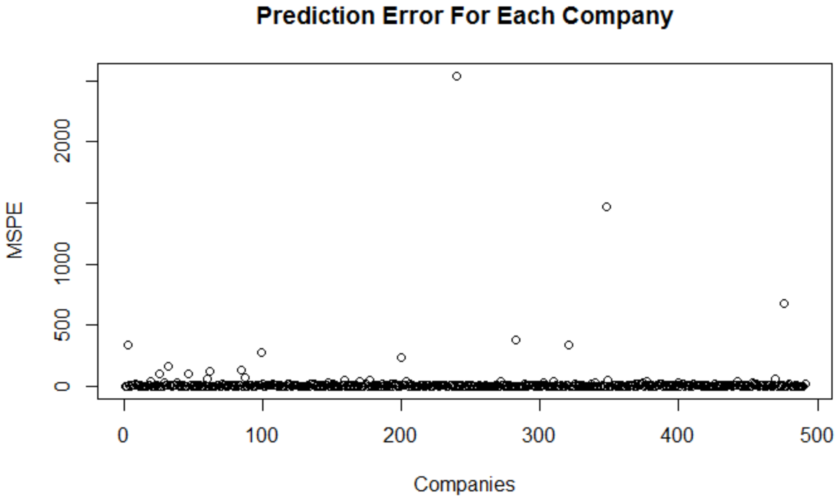
Initially we needed to come up with time series predictors of interest and decide which variable we wanted to predict out of those available from daily, high, low, opening, and closing stock price for 500 companies that had data from 1999 to 2013. We decided weekly averages of the closing stock price at the company level would be reasonable due to the very large number of observations. This allowed us to have a bit over 300,000 observations instead of 1.8 million in order to reduce computation time. We also transformed the year/month/day numeric date provided into a formatted YMD() using the R package "lubridate" and then generated the unique date of every Monday corresponding to the weekly average closing price we calculated.

The next step was to make sure we kept the companies separated from each other by creating a column that had the company abbreviations repeated for each corresponding row of averages and date.

Ridge Regression without Differencing

Upon formatting the data in a way we could use it, we separated the data into a test and training set where the test set was the eight most recent weekly average closing prices and their corresponding previous 10 week weekly average closing prices. For the training, it was the rest of the previous 10 weeks and the corresponding 11th week for a response. R does not have a predict function for ridge regression so we used lm.ridge() within a loop that would subset the observations from a unique company and store the model produced from the ridge function. This allowed us to have a regression model fitted to each unique company. We had to remove companies without enough weeks of data for prediction so we only kept 491 of the 500 companies.

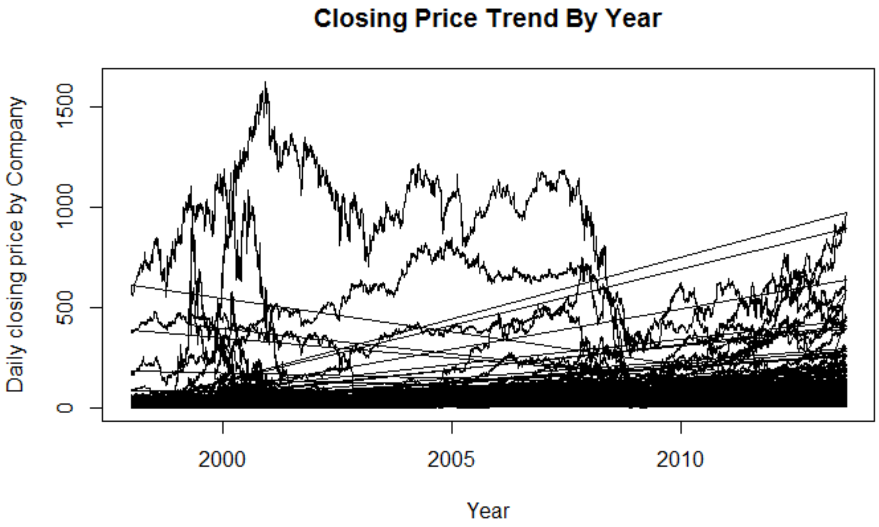
We then calculated and stored all of the predicted weekly average closing prices corresponding to the eight weekly averages in the test set followed by another variable storing the respective mean square prediction errors.



1st Order Random Walk

The range of MSPE values was from .02 to around 3000 indicating decent predictions for some companies and very poor predictions for others. In order to achieve better predictions, we noticed that the time period for the stocks had very different trends in prices during the periods before and after the recession and during the stock market issues during 2009-2010. To correct for these seasonal trends, we decided to try taking the iterative difference from the first two weekly average closing prices up to the last two weekly average closing prices. We are currently still working on this due to some issues with our code giving errors we haven't figured out yet.

Theoretically, this differencing should make the data appear more stationary. It is possible that there will not be a massive difference since there were only a small amount of companies with a large fluctuation in price out of the 500 possible companies. If we can get this part to run correctly, then it is a simple matter of just rerunning our previous ridge/prediction loops with the stationary data. The following is a graph of closing price over time for the 491 companies.



Complications and Future Work

Due to only having access to the free data, we didn't have the entire data set, nor all of the variables. Having access to more information could have led to better predictions. Also, we considered using other companies to predict the future of a different company but we only had access to 500 out of the 4000~ companies listed in the company names file. We do not know if the 500 were randomly selected from the 4000 possible.

We are not up to date with time series, so we had many issues figuring out how to difference the data to make it stationary. Given more time, we would learn better methods for making the data stationary.