ORIE 4741 Final Project
Validating Value Investing Theorem

Anqi Wang (aw735),Trevor McDonald (tdm67). Yi He (yh354)

Instructor: Madeleine Udell

November 2016

1 Introduction

Value Investing is a century long investment theory proposed by Ben Graham and David Dodd in 1928. It basically says that “markets systematically undervalue companies with high cash flow but large book values and stable businesses”. If what this theory claims is true, a big challenge on current efficient market theory is posed and a profitable investment opportunity is implied. For the past decades, this theory has been advocated by many famous investors including Warrent Buffet, Laurence Tisch and Michael Larson, but whether their success is due to this theory or luck is up to further research and analysis.

In this project, our first objective is to test whether Value Investing Theorem is a decent investment strategy which will bring investors a long-term profit. A Value Investing Strategy demonstrates that value investors should choose undervalued stocks. As a result, we wish to compare if the undervalued stocks will bring investors a relatively larger return. Furthermore, another objective is to build a predictive model which could forcast stock returns based on some meaningful financial fundamentals.

2 Data Exploration

An undervalued stock is believed to be priced too low based on current indicators, such as those used in a valuation model. Should a particular company’s stock be valued well below the industry average in terms of fundamental index, it may be considered undervalued. In these instances, value investor may focus on acquiring these investments as a method of pulling in reasonable returns for a lower initial cost. Typically, value investors select stocks with lower price-to-book, price-to-earnings ratios (P/E ratios), higher dividend yields and larger market capitalization etc.. Investors invest if the comparative value is high enough. Also, value investors desire the stocks with a low earning multiple, which is a high earnings yield.

Keeping the objective in mind, our dataset will contain all public companies with their current price-to-sales, price-to-book ratios, price-to-earning, Price/Earnings to growth (PEG ratios), dividend yields, market capitalization, industry classification and stock exchange place as features. All these financial fundamentals are effective and obtained by October 28th 2016.

For a primary approach to the first objective, we used bloomberg terminal to collect all the stocks traded in Nasdaq (National Association of Securities Dealers Automated Quotations) and AMEX (American Stock Exchange), which sum into around 3700 stocks and we extracted their current corresponding fundamental indices (mentioned above). To meansure the stocks performance, we calculated their yearly price change of date October 28th 2016 and one year from that date (October 28th 2015).

The first step for handling our dataset is how to deal with missing data (marked as “N/A”). There are some fundamental financial indices missing from the datasets. First, for some of the missing data, we converted them to zero (only fpr PEG and P/E ratio categories). Take the 51Job Inc. (JOBS) as an example, it is missing the data of Dividend Yield. First, we manually go to check if we can abtain the data from other financial resources. But we found that the data is not here since the 51Job Inc has never distributed dividends to its shareholders. So it makes sense to adjust the “N/A” entry to “0” for the Dividend Yield. Second, we used the strategy to exclude the whole row which includes missing data. It’s reasonable and suitable for the situation in our project. Keeping those special circumstances in mind, some of the strategies of coping with these missing data (such as using mean or median of other available data from the same column) are not suitable for our dataset. If we filling in the missing entries with column mean, the data will become so skewed and will become meaningless. As a result, we decided to drop the rows of missing data. After dropping all these stocks, the dataset are left with 1421 stocks. It’s still enough for our analysis.

Secondly, we reformated some of the data entries to ensure that they are measured in the same scale. For example, Market Capitalization entry expressed as “34.5M” were changed to “34,500,000”. We constructed number expression rather than the way with string representation.

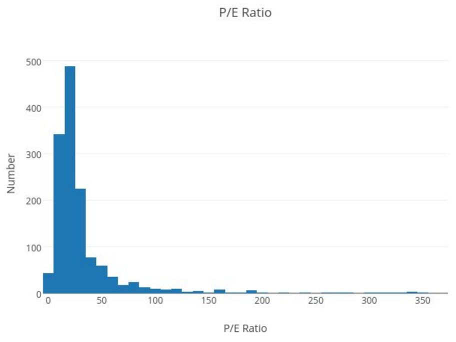
For the more detailed datasets we used for the advanced predictive model, we generally employed the same strategy of data handling and cleaning as before. To fully utilize maket volatility, we decided that daily stock data may be more appropriate for our predictive model. After we began our data extraction process, we sadly found out that there is a limit that one bloomberg terminal can only make 10k data quests everyday: in order to reduce the size, we randomly chose 200 stocks which are suitable for our model in each industry sectors. After eliminating some data points, the datasets for predictive model became $227\ \textit{stocks} \times 548\ \textit{days} \times 6\ \textit{features}$ (date ranging from 10/1/2014-10/1/2016).

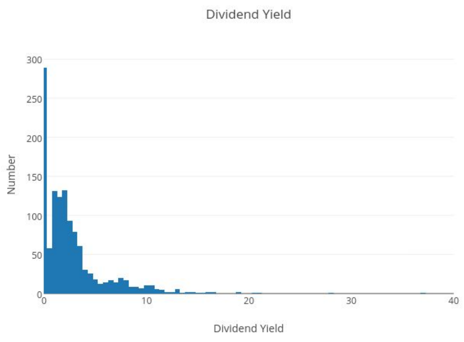
The daily data sets contains more NA value than yearly data. Since this is daily data, our intuition is to fill the NA value with their nearest neighbor. Another approach is simply to delete one data entry if any of its features shows a NA. As we will see in the predictive model section, these two methods produced 2 cleaned datasets that leading to very different results.

Moreover, PEG value is removed in daily dataset for predictive model because bloomberg does not provide this value in daily frequency, which makes sense because earning growth itself is a yearly index.

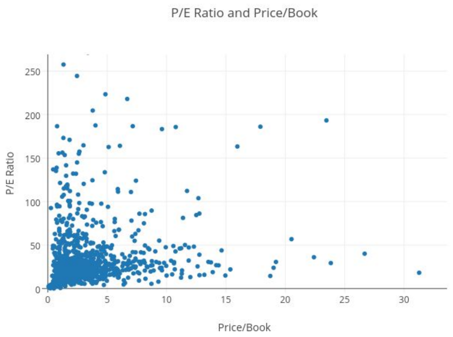
2.3 Descriptive Statistics

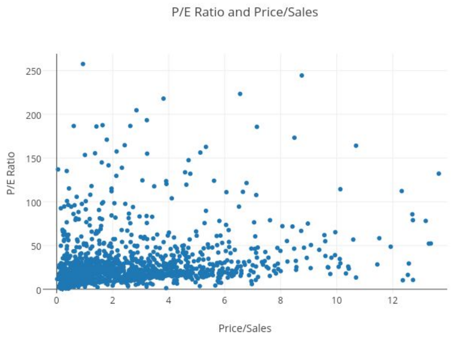
In this section, we generally creates data visualization to have a better understanding of the first dataset.



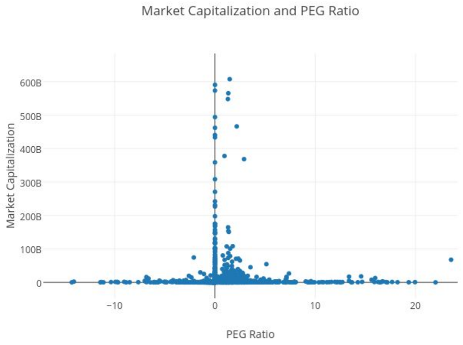


(i) (ii)





(iii) (iv)



(v)

We first decided to have a look at our data. In chart (i), we saw the range of the price per earnings ratio of the companies we werelooking at, one of the key measures in value-based investing. The histogram of p/e ratio is generally left skewed, with mean around 24. We spotified an outlier with p/e ratio equals to 340. In chart (ii), it shows the distribution of the dividend yields for the companies we are looking at, expressed as a percentage of the share price. It's also left skewed. The mode is at "0", indicates that most of the companies never distribute dividend. The scatter plot in (iii) shows the P/E ratio vs the P/B ratio. A lower P/B ratio could mean that a stock is undervalued, an indicator to a value-investor that it could be time to buy. On visual inspection, it looks like these two could be slightly positively correlated. In plot (iv), Price/Sales reflects how "valuable" each dollar of a company's revenue may be. This shows a possible slight positive correlation as well. In plot (v), market capitalization is the total value of shares for a company, and PEG ratio combines price per earnings with the growth rate over our time period. Interestingly, These two seem to show some sort of guassian relationship, something that may require further investigation.

3 Primary Models

In this section, we decided to use linear regression model to determine the relationship between stocks returns (as Y) with it's corresponding financial fundamentals (as X). Before starting our regression, we checked and reformated our data: all stocks with entries missing are excluded from the incoming regression. All stocks are grouped by their sectors: Basic Industries, Capital Goods, Consumer Durables, Consumer Non-Durables, Consumer Services, Energy, Finance, Health Care, Miscellaneous, Public Utilities, Technology, or Transportation. Then we calculated the sector average for all those fundamental index listed above. If any stock miss one or more of those index, we will not include that stock in the corresponding index mean calculation, but will include it in other index mean calculation.

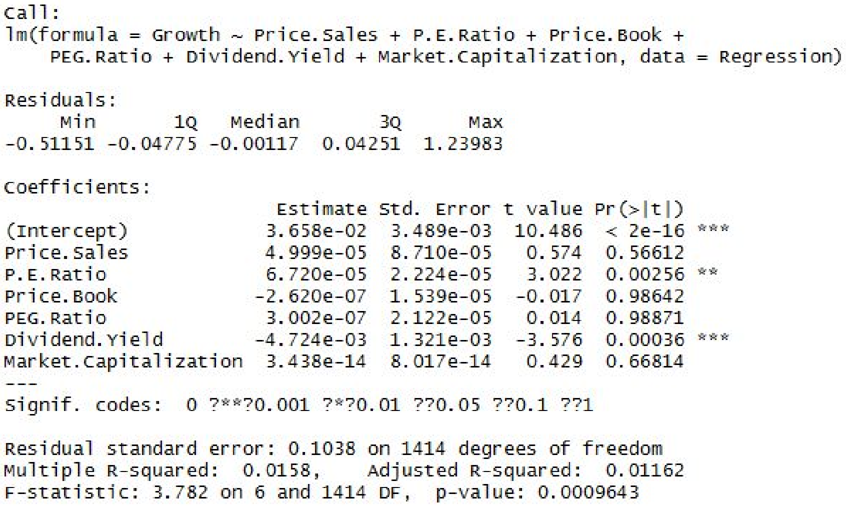
We did two linear regressions with respect to our collected dataset.

3.1 The First Linear Regression

The first regression is regressing the return rate for the past one year on its Price-Sales (P/S), Price-Earning (P/E), Price-Book (P/B), Price-Earing-Per-Growth ratio (PEG), Dividend Yield (DI) and Market Capitalization (CAP). Model written as:

$$Y_{return} = \beta_0 + \beta_1 X_{P/S} + \beta_2 X_{P/E} + \beta_3 X_{P/B} + \beta_4 X_{PEG} + \beta_5 X_{DI} + \beta_6 X_{CAP}$$

By running this regression, we got:



```
Call:
lm(formula = Growth ~ Price.Sales + P.E.Ratio + Price.Book +
    PEG.Ratio + Dividend.Yield + Market.Capitalization, data = Regression)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51151 -0.04775 -0.00117  0.04251  1.23983

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.658e-02  3.489e-03  10.486  < 2e-16 ***
Price.Sales            4.999e-05  8.710e-05   0.574  0.56612
P.E.Ratio              6.720e-05  2.224e-05   3.022  0.00256 **
Price.Book            -2.620e-07  1.539e-05  -0.017  0.98642
PEG.Ratio              3.002e-07  2.122e-05   0.014  0.98871
Dividend.Yield        -4.724e-03  1.321e-03  -3.576  0.00036 ***
Market.Capitalization  3.438e-14  8.017e-14   0.429  0.66814
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.1038 on 1414 degrees of freedom
Multiple R-squared:  0.0158,    Adjusted R-squared:  0.01162
F-statistic: 3.782 on 6 and 1414 DF,  p-value: 0.0009643
```

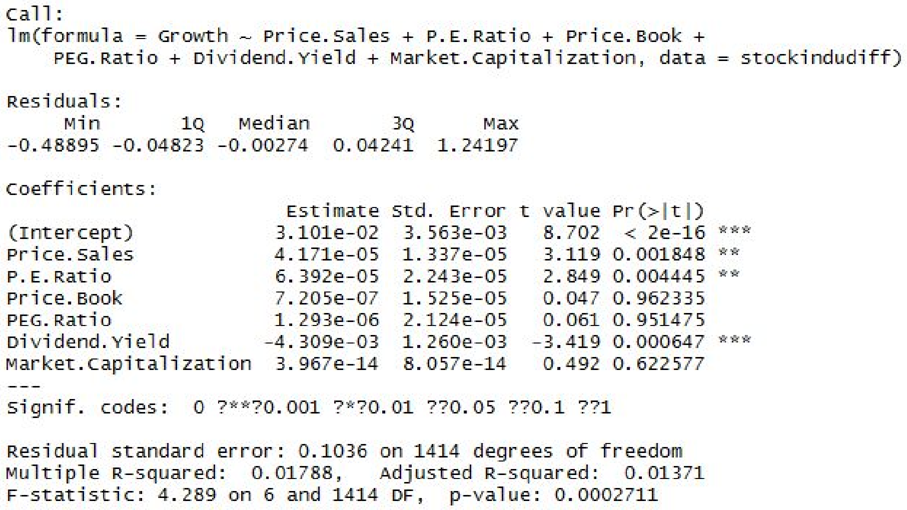
As we can see, Price-Sales, PE, PEG ratio and Market Capitalization do contribute to a higher return. To our surprise, Price-Book value has a negative coefficient but the t-value is very small, which means the overall negative coefficient for Price-Book ratio is not an counterexample of value investing theorm. Besides that, this regression is more of a reference than any direct evidence for or against value investing theorm as the fundamental ratios here is their absolute value, not comparative value, as needed in the value investing theorem.

3.1 The Second Linear Regression

The second regression is regressing the return of each stock based on the difference of their fundamental index and the sector average level. Model written as:

$$Y_{return} = \beta_0 + \Delta\beta_1 X_{P/S} + \beta_2 \Delta X_{P/E} + \beta_3 \Delta X_{P/B} + \beta_4 \Delta X_{PEG} + \beta_5 \Delta X_{DI} + \beta_6 \Delta X_{CAP}$$

The result we got is:



```
Call:
lm(formula = Growth ~ Price.Sales + P.E.Ratio + Price.Book + 
    PEG.Ratio + Dividend.Yield + Market.Capitalization, data = stockindudiff)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.48895 -0.04823 -0.00274  0.04241  1.24197 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.101e-02  3.563e-03   8.702  < 2e-16 ***
Price.Sales           4.171e-05  1.337e-05   3.119 0.001848 ** 
P.E.Ratio             6.392e-05  2.243e-05   2.849 0.004445 ** 
Price.Book            7.205e-07  1.525e-05   0.047 0.962335    
PEG.Ratio             1.293e-06  2.124e-05   0.061 0.951475    
Dividend.Yield       -4.309e-03  1.260e-03  -3.419 0.000647 ***
Market.Capitalization 3.967e-14  8.057e-14   0.492 0.622577    
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.1036 on 1414 degrees of freedom
Multiple R-squared:  0.01788,   Adjusted R-squared:  0.01371 
F-statistic: 4.289 on 6 and 1414 DF,  p-value: 0.0002711
```

As we can see, the coefficient gets positive in this regression, though it’s t-value is still small. All of the Price-fundamental coefficient are positive, which is not quite what value investing theorem predicted while the dividend yield having a negative effect on return rate. The market capitalization does have a prositive effect on stock return (same as the value investing theory predicted). However, overall, the result of this regression cannot serve as a primary evidence of the value investing theorem, at least not in the time period of the recent one year.

4. Predictive Model

After we got an overview of the primary yearly data and made a rough understanding of the relationship between each of the predictive factors and the yearly price change, we decided to move on to daily stock data and build a predictive model that may be useful in day trading or other estimation of price fluctuation.

With what the professor taught us in mind, we decided to try five methods on the data. We tried simple linear regression, linear regression with ridge regularizer, lasso regularizer and huber regression. Then we decided to jump out of linear regression field and using SVM, an improved perceptron method to predict whether the price will increase or decrease tomorrow based on fundamental data today. i.e. a categorization problem with positive price change and negative price change as two categories.

All data is randomly splitted into test and training with 20% and 80% in proportion. For the purpose of comparison, only test result (t-value, r-squared value etc.) will be shown and discussed.

First, we used the cleaned data with NA filled with nearest neighbor and regressed a simple learn model. The coefficients for the remaining five variables: PE, PS, CAP, PB and Dividend is

Coefficients: [6.64606137e-08 1.08176398e-04 3.38084653e-10 3.66845035e-06 -7.73313163e-05]

We can see that these results are all much smaller to two or three power compared with the results we got from the yearly data’s linear model. So we used it as a reference point and move on.

Then we added Lasso regularization, the result is:

```
“Lasso Regression with 3-fold Cross-validation
TEST STATISTICS
Coefficients: [  5.78095079e-08   0.00000000e+00  -6.66919151e-10   0.00000000e+00
-1.46713844e-04]
R2 score: 0.000554848134686
Mean square error for model: 0.000379071509183”
```

As we expected, Lasso performed variable selection and PS and PB value are marked with zero coefficients.

The results of Ridge regularization is:

```
“Ridge Regression with 3-fold Cross-validation
TEST STATISTICS
Coefficients: [  6.64606137e-08   1.08176398e-04   3.38084653e-10   3.66845035e-06
-7.73313163e-05]
R2 score: 1.49640024261e-06
Mean square error for model: 0.000379281385509
”
```

Which is comparable to the results in Lasso. But we have noticed that the R-Squared value for these two methods are all much smaller than our previous yearly linear model.

We are not very satisfied with these results and wondered if the tail volatility in price change and outliers can be the cause of the problem, so we tried Huber regression to decrease the effects of outliers

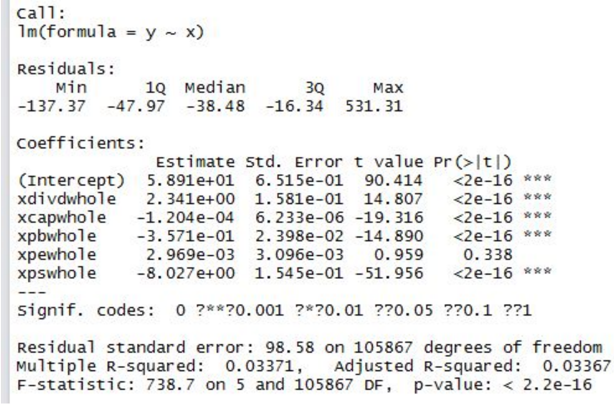
The results of Huber Regression is:

```
“Huber coefficients:  [ 7.60688999e-12   5.93235380e-13   9.93338859e-10   7.98625362e-13
  -3.18515366e-14]
Huber intercept:  1.80872771017e-13
TestingStatistics:
R2 score: 0.000398163043167
Mean square error for model: 0.000379432969123”
```

As we can see, the results is still of not major improvements from Lasso and Ridge.

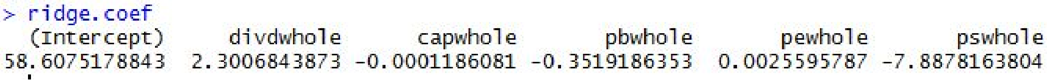
Then one member of our team proposed that it may not be a good data cleaning method simply filling NA value with nearest neighbor as many parts of the data missing entres consecutively. So we tried cleaning our data by deleting any entry with at least one features as NA. Then run all four regression above again.

The result of simple linear is:



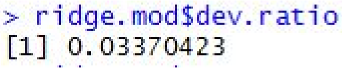
```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-137.37  -47.97  -38.48  -16.34  531.31 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.891e+01  6.515e-01  90.414   <2e-16 ***
xdivdwhole   2.341e+00  1.581e-01  14.807   <2e-16 ***
xcapwhole   -1.204e-04  6.233e-06 -19.316   <2e-16 ***
xpbwhole    -3.571e-01  2.398e-02 -14.890   <2e-16 ***
xpewhole     2.969e-03  3.096e-03   0.959    0.338    
xpswhole    -8.027e+00  1.545e-01 -51.956   <2e-16 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 98.58 on 105867 degrees of freedom
Multiple R-squared:  0.03371,	Adjusted R-squared:  0.03367 
F-statistic: 738.7 on 5 and 105867 DF,  p-value: < 2.2e-16
```

The result of ridge regression is:



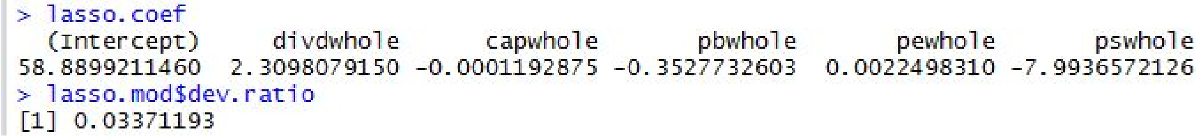
```
> ridge.coef
  (Intercept)      divdwhole       capwhole        pbwhole        pewhole        pswhole 
58.6075178843  2.3006843873 -0.0001186081 -0.3519186353  0.0025595787 -7.8878163804 
```

R-Squared value is:



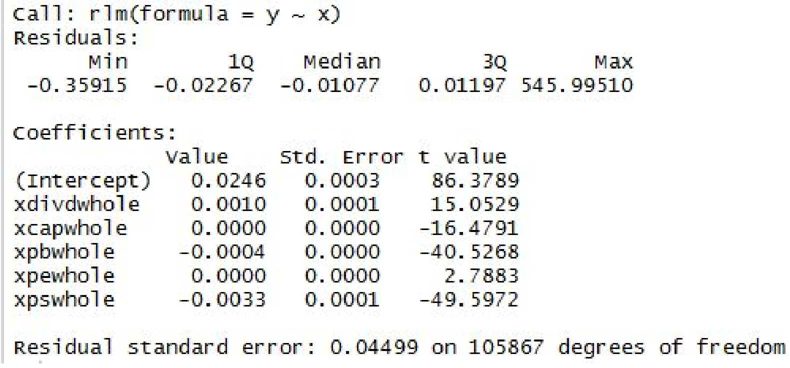
```
> ridge.mod$dev.ratio
[1] 0.03370423
```

The coefficient and R-Squared value of lasso regression is:



```
> lasso.coef
  (Intercept)      divdwhole       capwhole        pbwhole        pewhole        pswhole 
58.8899211460  2.3098079150 -0.0001192875 -0.3527732603  0.0022498310 -7.9936572126 
> lasso.mod$dev.ratio
[1] 0.03371193
```

The result of Huber robust regression is:



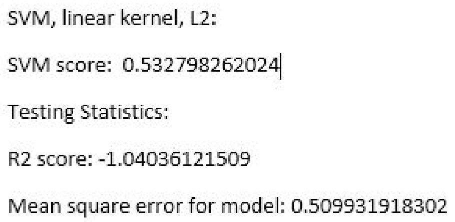
```
Call: rlm(formula = y ~ x)
Residuals:
      Min        1Q    Median        3Q       Max 
 -0.35915  -0.02267  -0.01077   0.01197 545.99510 

Coefficients:
            Value    Std. Error t value 
(Intercept)  0.0246   0.0003     86.3789
xdivdwhole   0.0010   0.0001     15.0529
xcapwhole    0.0000   0.0000    -16.4791
xpbwhole    -0.0004   0.0000    -40.5268
xpewhole     0.0000   0.0000      2.7883
xpswhole    -0.0033   0.0001    -49.5972

Residual standard error: 0.04499 on 105867 degrees of freedom
```

As we can see, the test R-Squared value became much bigger when we cleaned our data the right way. The Huber regression actually have some variable selection function in its output.

(By the time of this report, we haven't finished our Huber regression study in class, and we decided to dig deeper into the reason why it produced several zero coefficients as ourcourse study moves on).

Finally, we conducted our try to make the price change a positive-negative classification problem by performing the SVM method, an advanced perceptron non-linearly seperately datasets.

The result is:



```
SVM, linear kernel, L2:

SVM score:  0.532798262024

Testing Statistics:

R2 score: -1.04036121509

Mean square error for model: 0.509931918302
```

Since random guess produces 0.50 SVM scores, we prediction based on fundamental is just a little better than pure guess. This result is meaning in the sense that it still has around 3% chance of profitting by betting whether tomorrow's price would go up or down based on today's fundamentals. By trading consistently in the long run, we believe it will give the trader a small edge with proper leverage.

5. Conclusion

Our takeaways from this project is that:

1. fundamental indices only plays a small part in price change in yearly data and even smaller change in daily price fluctuation. Thus Value Investing Theorem may only help, but cannot promise big profit.
2. Data cleaning is very important in building predictive model and different ways of handling missing data can lead to very different results.
3. It's easier to predict the rough direction of price change than accurately measuring it quantitatively

We will keep analyzing the results of the five methods we tried in this project, combined with the material taught in class to get a deeper understanding and mastery of them.