Construction of a Sample Portfolio to Track an Index in Passive Investment

Yunqing Hu, Cheng Chen

Executive Summary

- There are many passive funds whose strategy is to replicate an index or pool of indexes. This project provides a solution to track an index with a small number of stocks, instead of holding all constitutes in the index. In this way, the portfolio is easier to construct and will not be influenced by the stock adding and removing of the index.
- It is widely believed that market is efficient and the market portfolio has the highest Sharpe ratio, and actively selecting stocks will not perform better than just following the market, As we know, index is a proxy for market, so it is important to find a way to mimic the index effectively. In this project, we use regression analysis to solve this problem. The data we use is S&P 500 index and stock prices. We regress the index against all stocks prices and use stepwise method to select a subset. Based on the stepwise result, we choose to include 10 stocks in the tracking portfolio. We then do out-of-sample study to evaluate this performance of this portfolio, and find that the portfolio needs to be rebalanced periodically. We come up with three rebalancing strategies and compare their advantages and disadvantages. If our company is an ETF that tracks S&P 500, this project will have some practical use.

Introduction

As we know, if market is perfectly efficient, the market portfolio is the optimal portfolio, and no other portfolios will beat the market. So in passive management, the investor does not attempt to outperform the market through changes in security holdings or market timing. Research supporting passive management is grounded in the logic of economic theory and the efficiency of the capital markets. “The Arithmetic of Active Management”, William F.Sharpe (1991), for example, explains why the average investor cannot hope to beat a comprehensive equity index.

The dominant passive approach is indexing, which tries to buy and hold every security from the market. Pioneered in the 1970s, indexing has quickly grown and today indexed portfolios often function as the core holding in an investor’s overall equity allocation. There are funds whose explicit strategy is to replicate an index or pool of indexes, as for example the Vanguard 500 index fund which, on Nov 2012, showed a total asset value of 117 billions of dollars.

However, there are some drawbacks of replication.

- Too many stocks to hold. Some stocks may be in very small quantities, hence not efficient
- Stocks are added and removed from the index from time to time, and the portfolio needs to adjust to these changes

Due to the drawbacks of replicating an index, there is a simple approach called index tracking. The idea is to use a small number of stocks, instead of holding all the stocks, to track the index.

In this project, we are doing a study of constructing a sample portfolio that tracks S&P 500 index. We use regression model to construct a portfolio containing a small number of stocks. Followed by back-testing and some further discussions, such as rebalancing and hedging specific risk factor. At last, we evaluate our model and point out the strengths and weaknesses. The rest of this report is organized as following. In Section I, we build a regression model, using the index as response variable, and stock prices as predictor variables. Three selection methods are introduced in model selection. In Section II, we use the constructed model to form a tracking portfolio. In Section III, we do out-of-sample study analysis to evaluate the prediction ability of the tracking portfolio. In Section IV, we discuss some rebalance strategies. In Section V, we consider a factor portfolio to hedge a specific risk.

Data

We use S&P 500 data in this project. The data includes S&P's weekly index levels and weekly prices of the 500 constitutes. Data from 2000-2006 is used to build the model, and back-testing study uses 2007-2011's data.

First, we downloaded the daily data of S&P index and 500 stocks from 2000/01/01-2011/12/31 using Wharton Research Data Services, and transformed them to weekly by extracting the data on Fridays. Because of the effects of stock dividend and split, the actual price is not consistent. To solve this problem, we use "Cumulative Factor to Adjust Price" to make it consistent through time. Then we get rid of stocks that are added or removed from the S&P 500 list during 2000-2006, and about four hundreds stocks are left. We use Java to process the data and transform it to the eligible format.

Section I Model Selection

Denote Y the S&P 500 index level, and $X_1, X_2, ..., X_k$ the price (stock dividend and split adjusted) of K stocks we use to track the index. The following regression model is used to find the relationship between Y and $X_1, X_2, ..., X_k$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon .$$

We can also use return instead of price, the result will be slightly different. It is shown in the appendix, with the comparison of the two methods.

Since there are hundreds of potential predictor variables, a tradeoff has to be made between K, the number of predictor variables, and ε, the error term. We use forward search, backward search and stepwise search methods to select a good subset of variables. Figure 1 is the plot of adjusted R^2 against the size of subsets in three approaches.

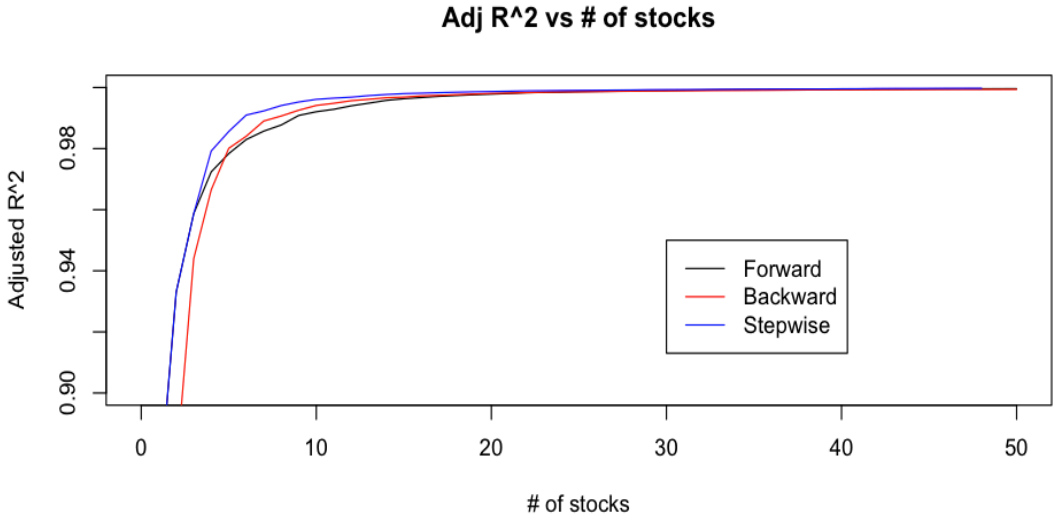

Figure 1

From Figure 1, we can see that stepwise search provides higher adjusted R^2 than the other two given a certain level of subset size, and after k=10, the adjusted R^2 improves slowly. So we select the subset of size 10 given by stepwise approach. The TICKER of the 10 stocks selected are listed below.

"MSFT" "TROW" "PFE" "CMS" "HAL" "BDX" "HD" "C" "CSCO" "EMN"

From the R results (in Appendix), all the coefficients are significant, and adjusted R-squared is 0.9961, suggesting this model is very good.

Model Diagnostics

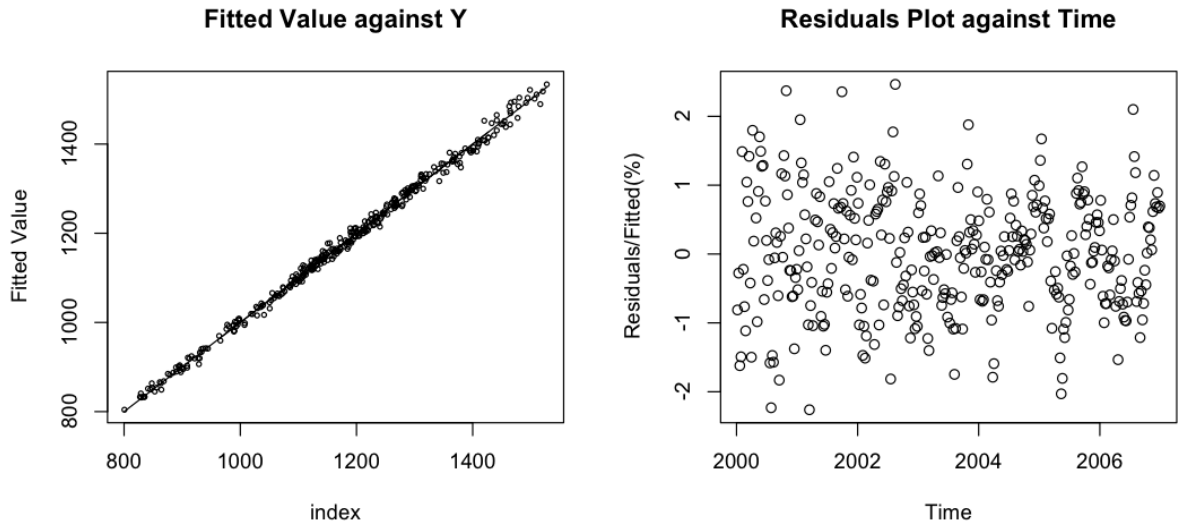

Figure 2

Figure 2 shows a plot of fitted value against the actual observations and a plot of residuals against time. Residuals are transformed to percentage by dividing the index. The residuals are almost within the $\pm 2\%$ interval. No outliers present in this plot.

Section II Tracking Portfolio

Now we have the regression model $\hat{Y} = b_0 + b_1 X_{\text{MSFT}} + b_2 X_{\text{TROW}} + ... + b_{10} X_{\text{EMN}}$, we can use it to construct a tracking portfolio.

Let's invest in b_1 shares of MSFT stock, b_2 shares of TROW stock, and so on, and hold b_0 amount of cash. Define this investment combination a share of the tracking portfolio. The value of one share is $b_0 + b_1 X_{\text{MSFT}} + b_2 X_{\text{TROW}} + ... + b_{10} X_{\text{EMN}}$, also the fitted value of the index.

Cash	MSFT	TROW	PFE	CMS	HAL	BDX	HD	C	CSCO	EMN
251.407	2.684	3.468	1.968	4.797	3.344	1.964	2.302	0.350	4.975	4.362

Table 1

Plot the trends of the index and the value of one share of tracking portfolio from 2000-2006 on the same diagram.

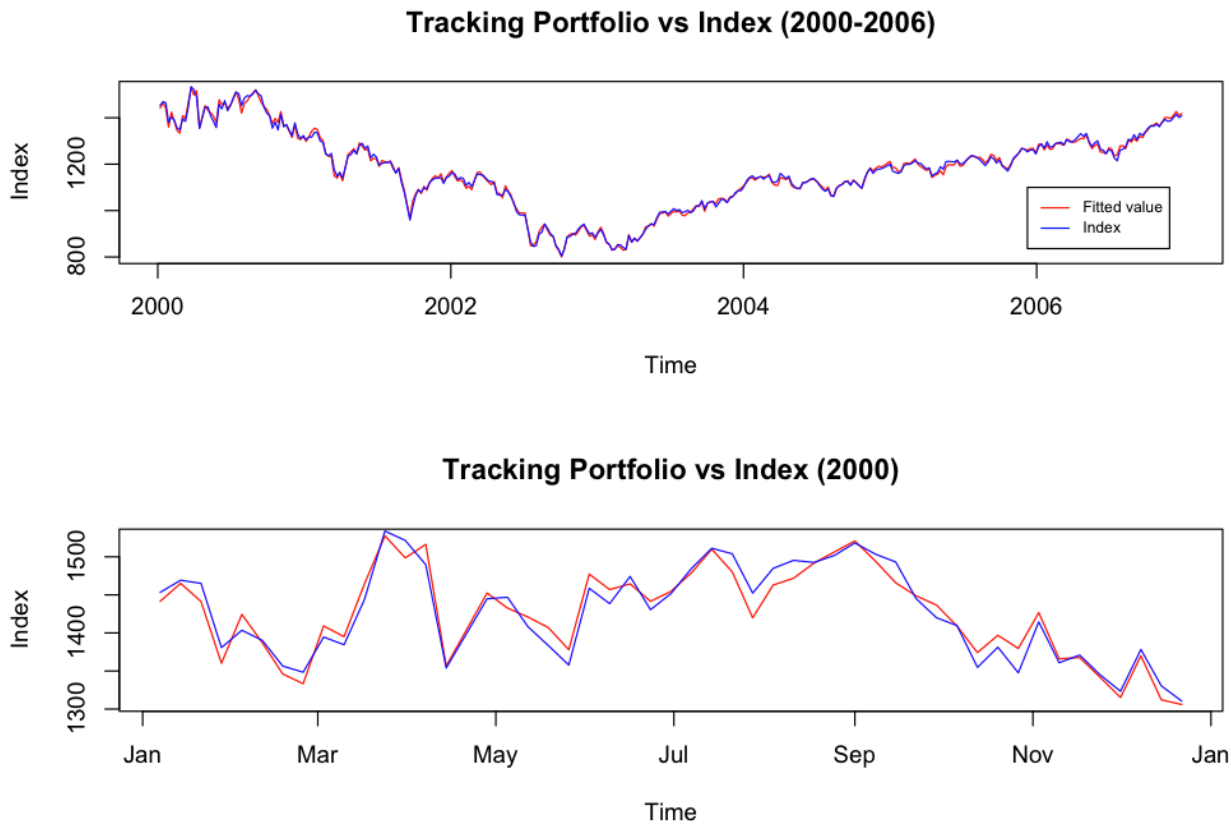

Figure 3

We see from Figure 3 that the portfolio tracks the index very closely and almost catches every small movements of the index.

Section III Out-of-Sample Study

Although the portfolio constructed above looks really nice, it's impossible to build a portfolio at the beginning of 2000 based on 2000-2006 data. So we have to do out-of-sample study to investigate the predict ability of the tracking portfolio beyond 2006.

At the beginning of 2007, we set up the portfolio and hold it without any adjustments. Figure 4 presents the performance of the tracking portfolio against the index and the deviation between them through time.

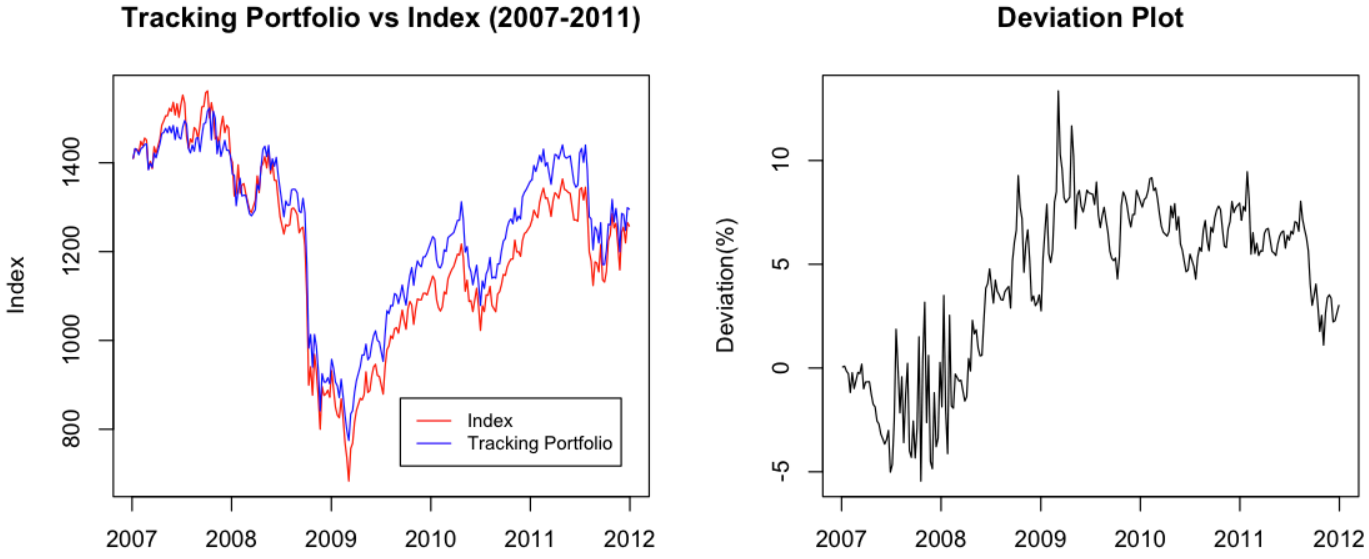

Figure 4

The deviation is small at the beginning, but increases to about 10% after two years. Although the plot suggests the tracking portfolio outperforms the market, it is not a good sign if our goal is to track the index, because it is also possible that the portfolio underperforms by that amount.

Section IV Rebalance Strategy

IV.1 Whole Rebalance

As we discussed above, the tracking portfolio performs well when it is initially established, but performs badly as time goes by. This is mainly because the companies' situations have changed during these years and thus they are no longer efficient in tracking the index. So it is necessary to rebalance the tracking portfolio periodically. In whole rebalance, whenever the original tracking portfolio is not efficient, we take recent price data into account to form a new one.

The newly built portfolio is constructed based on data from past N weeks. The reason why we don't use the whole sample starting from 2000 is that when the period is too long, the influence of recent data, which is essential for rebalancing, will be relatively small.

In order to determine the time when a tracking portfolio needs to be updated, we define the Absolute Tracking Deviation (ATD) as $\left|\frac{Portfolio - Index}{Index}\right| * 100\%$. When the moving average of ATD in the last 3 months (13 weeks) $\sum_{t-12}^{t} ATD_t / 13$ is greater than a certain level, the rebalance procedure is called.

Figure 5 shows the rebalancing schedule for 5% and 3% level of moving average ATD when N=100.

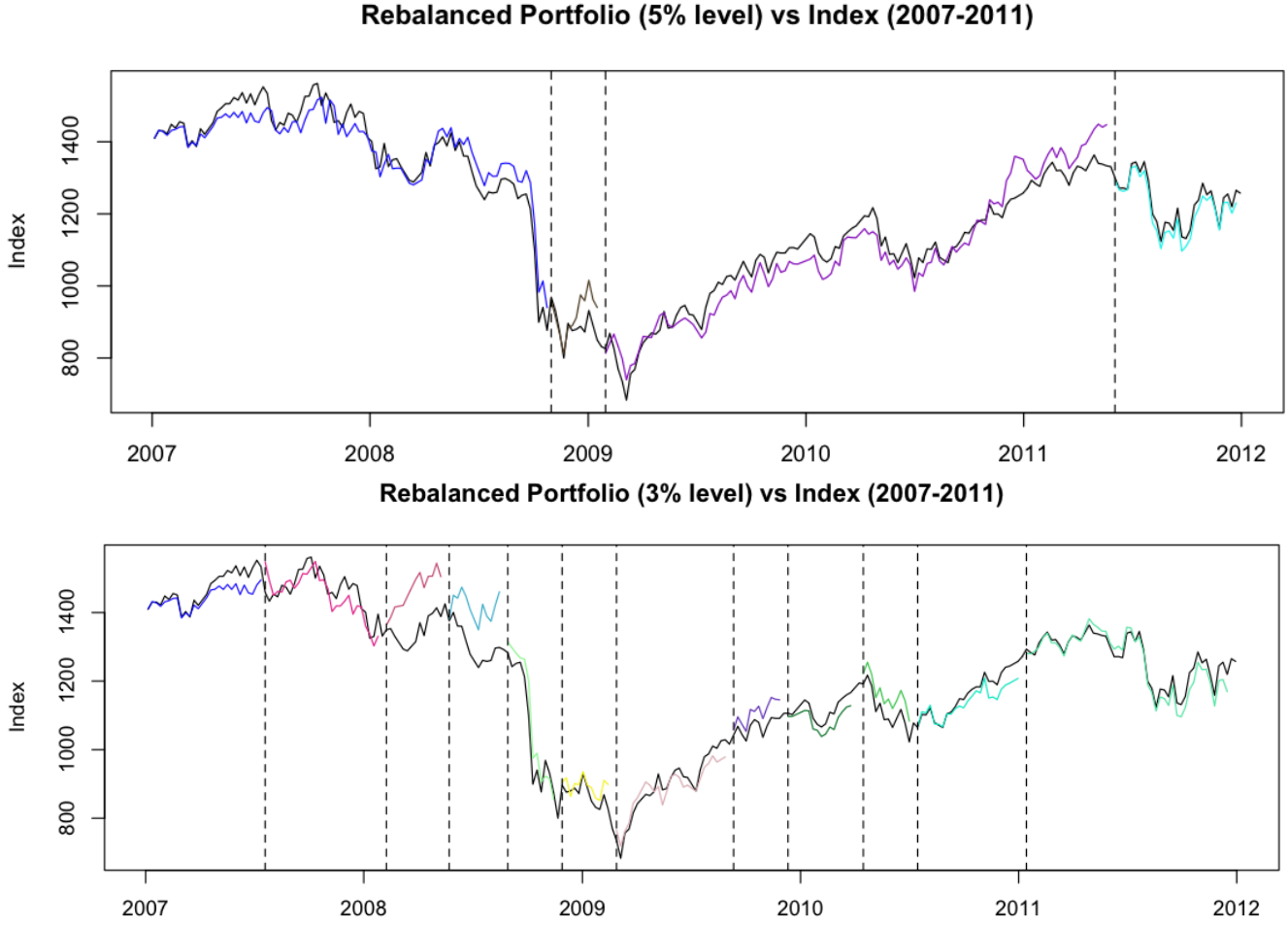

Figure 5

When the level is 5%, we only rebalance 3 times during 2007-2011, but at the level of 3%, we need to rebalance 11 times. It means when the level is lower, we need to rebalance more frequently. Meanwhile, the deviation seems smaller under 3% level. We compare different rebalancing strategies based on two criterions.

- The lower average ATD the better.
- The cost of rebalancing is high, so the less number of rebalancing the better.

We do scenarios test for control factors level and N, and summarize the results below.

Level \ N	100	150	200	250	300	Mean
3	3.18	3.34	2.76	1.85	2.37	2.70
4	3.04	1.97	2.36	2.12	2.21	2.34
5	3.04	2.86	2.94	2.02	2.50	2.67
6	3.19	2.75	2.27	3.29	2.63	2.83
Mean	3.11	2.73	2.58	2.32	2.43	

Table 2 Average ATD

Level \ N	100	150	200	250	300	Mean
3	11	10	8	6	8	8.6
4	8	1	3	3	4	3.8
5	3	3	4	1	3	2.8
6	3	2	1	3	2	2.2
Mean	6.25	4	4	3.25	4.25	

Table 3 # of Rebalancing

According to Table 2 and 3, N=250 has the minimum ATD, and level=5 or 6 has small # of rebalancing, so we prefer the pair (5.5,250).

Figure 6 plots the rebalancing time frame for Level=5.5% and N=250. The portfolio is rebalanced twice from 2007-2012, and it tracks the index well.

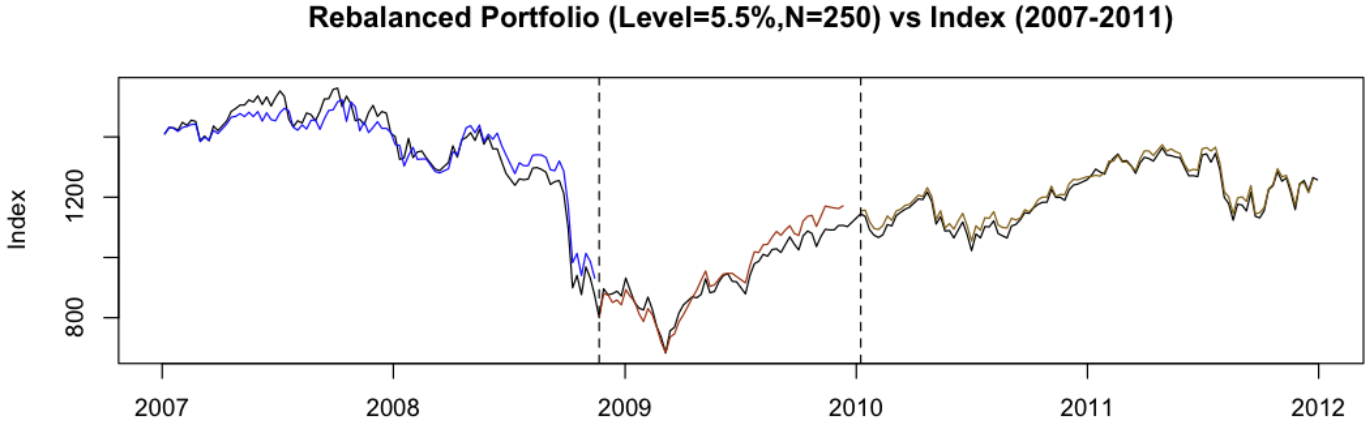

Figure 6

Figure 7 is a plot comparing the absolute deviation of non-rebalanced and rebalanced (5% level) portfolios. Originally, the deviation is quite high after 2009, and after periodic rebalancing, the deviation maintains at low level at all time.

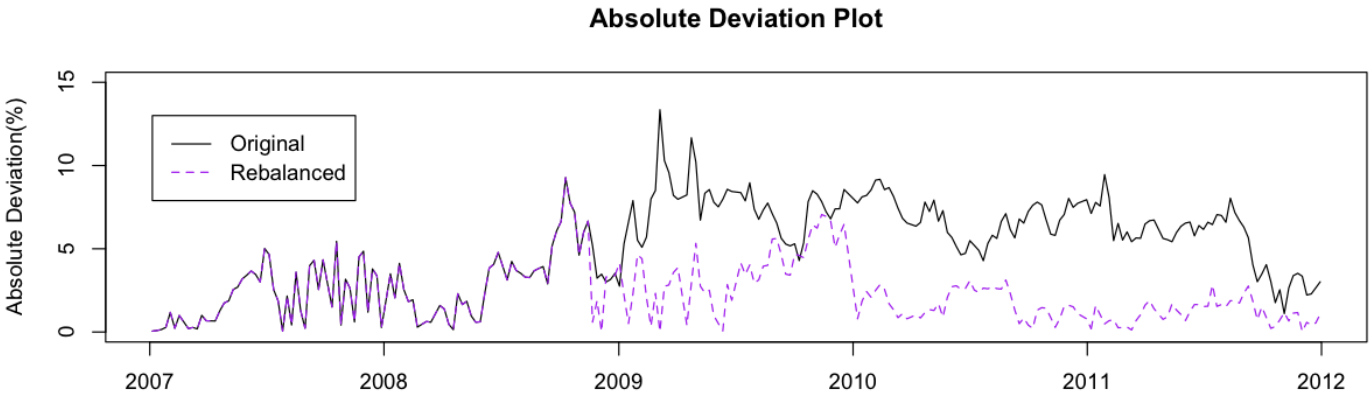

Figure 7

The following table is the detail of the rebalance schedule.

Date	$	Shares									
05/01/07	Cash	MSFT	TROW	PFE	CMS	HAL	BDX	HD	C	CSCO	EMN
	251.407	2.684	3.468	1.968	4.797	3.344	1.964	2.302	0.350	4.975	4.362
21/11/08	Cash	MSFT	LH	IR	APD	PKI	IGT	AXP	NWL	TIF	ALL
	210.732	2.676	2.459	1.808	1.736	4.480	0.792	4.113	3.200	1.930	3.197
08/01/10	Cash	KO	UTX	NBR	IFF	GPC	LNC	CI	AIG	BBT	CSCO
	159.698	2.690	2.776	3.908	2.628	4.303	2.960	1.403	0.067	1.204	4.206

Table 4

IV.2 Weight Rebalance

From Table 4, it is surprising to find that every time the portfolio is rebalanced, almost all 10 stocks are replaced. As we know, the cost of rebalancing is high since we have to sell the stocks hold and buy new ones. This can involve huge transaction cost when the capital of the portfolio is large. So another idea is using the initial stocks and rebalance their weights.

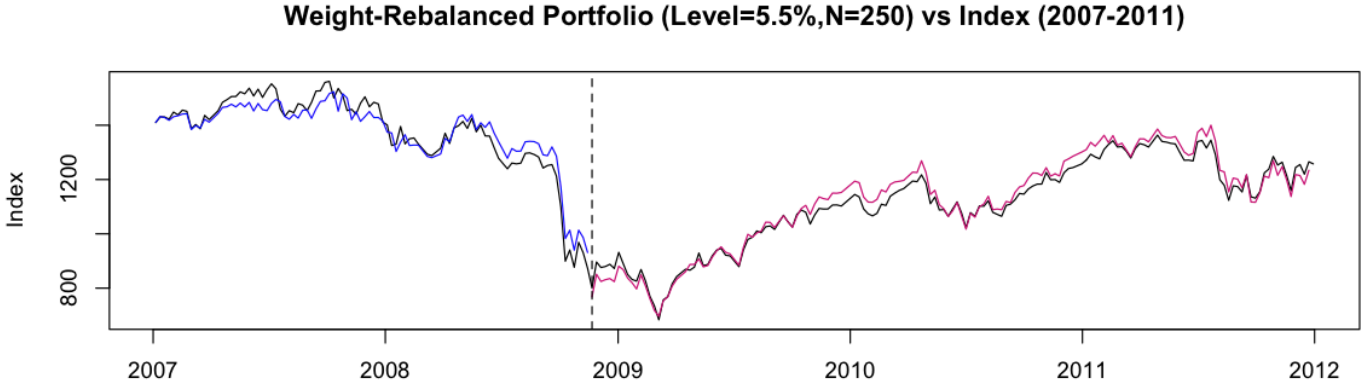

Figure 8

Date	$	Shares									
05/01/07	Cash	MSFT	TROW	PFE	CMS	HAL	BDX	HD	C	CSCO	EMN
	251.407	2.684	3.468	1.968	4.797	3.344	1.964	2.302	0.350	4.975	4.362
21/11/08	Cash	MSFT	TROW	PFE	CMS	HAL	BDX	HD	C	CSCO	EMN
	95.573	6.461	1.640	1.410	7.157	3.687	2.403	1.351	0.487	6.000	5.146

Table 5

From Figure 8 and Table 5, we see that the weight-rebalanced-portfolio also does a good job and the stocks' weights don't change much after rebalancing, thus the transaction cost to rebalance is much lower.

IV.3 Combined Rebalance

Model	Whole Rebalance					Weight Rebalance				
		Estimate	Std. Error	t value	Pr(>|t|)		Estimate	Std. Error	t value	Pr(>|t|)
Coefficients	MSFT	2.6760	0.5093	5.254	3.28e-07 ***	MSFT	6.46126	0.74582	8.663	6.87e-16 ***
	LH	2.4585	0.1773	13.864	< 2e-16 ***	TROW	1.63988	0.48488	3.382	0.000840 ***
	IR	1.8079	0.3724	4.855	2.17e-06 ***	PFE	1.40953	0.67190	2.098	0.036967 *
	APD	1.7356	0.2104	8.250	1.05e-14 ***	CMS	7.15683	0.93114	7.686	3.87e-13 ***
	PKI	4.4799	0.6338	7.068	1.70e-11 ***	HAL	3.68719	0.31322	11.772	< 2e-16 ***
	IGT	0.7917	0.2373	3.336	0.000984 ***	BDX	2.40290	0.35117	6.843	6.43e-11 ***
	AXP	4.1125	0.4262	9.649	< 2e-16 ***	HD	1.35128	0.59986	2.253	0.025184 *
	NWL	3.2003	0.8065	3.968	9.57e-05 ***	C	0.48665	0.03489	13.947	< 2e-16 ***
	TIF	1.9301	0.2446	7.891	1.06e-13 ***	CSCO	6.00044	0.68015	8.822	2.36e-16 ***
	ALL	3.1966	0.2794	11.442	< 2e-16 ***	EMN	5.14571	0.58516	8.794	2.86e-16 ***
Mean ATD	2.309084					2.34347				
Adjusted R^2	0.9908					0.9809				

Table 6

In Table 6, we compare some characters of the two models used in whole and weight rebalances. The coefficients in weight rebalancing model is not so significant and mean ATD, adjusted R^2 are higher compared to whole rebalancing model. It's still good, while after a long period, it is likely that even changing the weights of the original stocks will not work well. So we can combine the two strategies together to utilize both their advantages.

We come up with the following two combined rebalance strategies.

- Weight-Whole-Weight
  First use weight rebalances, until the R^2 of the model is less than a certain level, say 95%. Then we use whole-rebalance once, and switch back to weight rebalance afterward until the condition is satisfied again.
- Semi-Rebalance
  Every time we need to rebalance, we keep 9 of the 10 stocks, and replace the least significant one by another stock, which will make the R^2 of the new model the highest among all.

Section V Factor Portfolio

A pure factor portfolio (or simply a factor portfolio) is a portfolio that has been constructed to have sensitivity equal to 1.0 to only one risk factor, and sensitivities of zero to the remaining factors. Factor portfolios are particularly useful for speculation or hedging purposes. For example, assume a portfolio manager believes GDP growth will be stronger than expected, but wishes to hedge against all other factor risks. The manager can take a long position in the GDP "factor portfolio." The factor portfolio is exposed to the GDP risk factor, but is hedged (zero sensitivity) to all other risk factors.

Alternatively, consider a manager who wishes to hedge his portfolio against GDP factor risk. Assume the portfolio's GDP factor sensitivity equals 0.80, and the portfolio's sensitivities to the remaining risk factors are different from zero. Suppose the portfolio manager wishes to hedge against GDP risk, but remain exposed to the remaining factors. The manager can hedge against GDP risk by taking an 80% short position in the GDP factor portfolio. The 0.80 GDP sensitivity of the managed portfolio will be offset by the –0.80 GDP sensitivity from the short position in the GDP factor portfolio.

Suppose a manager in Great Britain wants to have the return of S&P 500, he is exposed to GBP/USD exchange rate risk. Instead of going into the future currency market, he can also consider a factor portfolio.

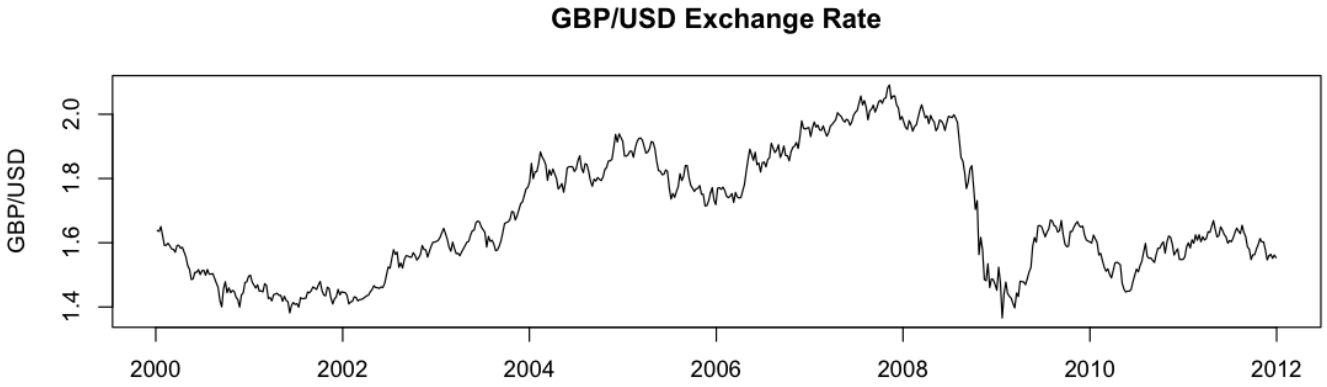

Figure 9

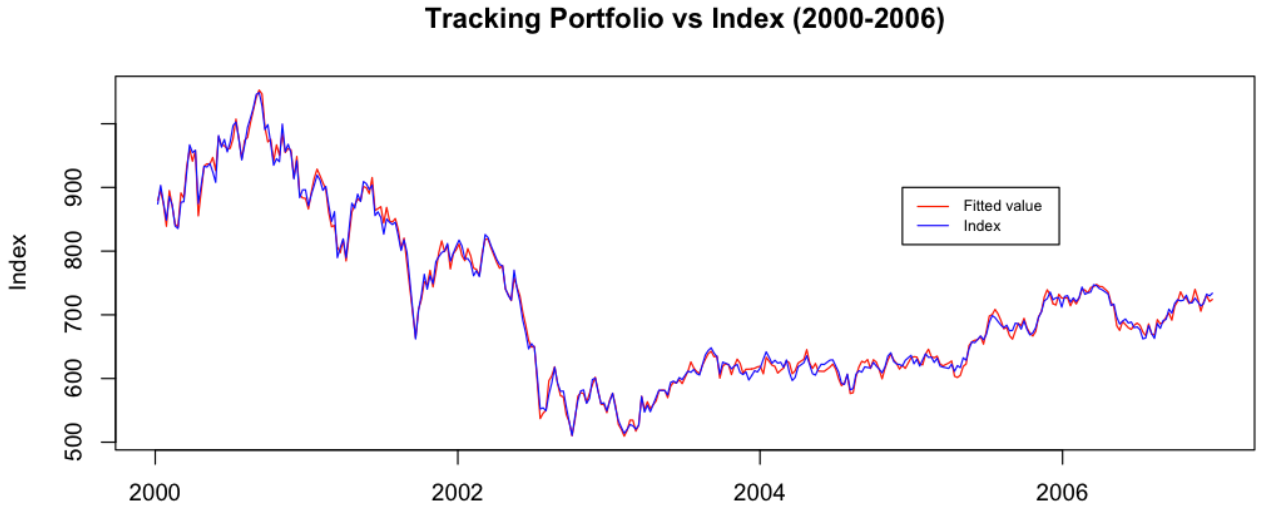

Figure 10

Figure 10 shows the in sample plot of the factor portfolio vs. the index level (in GBP). Although the in sample study has a good result, the out-of-sample result is somehow poor. This is mainly because few stocks in S&P 500 are highly correlated with GBP/USD exchange rate. So, some improvement should be made.

Conclusion

In this project, we have provided the solution to track an index with a small number of stocks using regression model. Through this project, we have learned the following things:

1. Market efficiency theory, passive investment, and portfolio management
2. Data search, and use Java, R to transform daily data to weekly and put in an organized form
3. Multiple regression model, model selection, model diagnostics and out-of-sample study

There are many ETFs are passive, investing in bonds, stocks, commodities and others. And there are kinds of bond indexes, stock indexes, etc. So if our company is an institution investor within these areas, the strategy proposed in this project will benefit our company in terms of the cost.

If we are given 6 more months, we can do the following things:

1. Use indexes in other investment areas other than stocks (bonds, commodities, alternatives) to construct portfolio
2. Evaluate the profit of the portfolio and try to outperform the index
3. Adjust the rebalance strategy to act differently when the portfolio is above of below the index
4. Get data from a longer period and do Weight-Whole-Weight Semi-Rebalance strategies to see their performances
5. Improve the factor portfolio and consider more risk factors to hedge

Appendix

Model	Price				Return			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
Coefficients	MSFT 2.68423	0.19029	14.106	<2e-16 ***	MSFT 0.0768411	0.0089003	8.634	2.29e-16 ***
	TROW 3.46756	0.28374	12.221	<2e-16 ***	XOM 0.1215246	0.0130445	9.316	< 2e-16 ***
	PFE 1.96752	0.22369	8.796	<2e-16 ***	GE 0.0940222	0.0121196	7.758	1.00e-13 ***
	CMS 4.79737	0.18230	26.315	<2e-16 ***	AMAT 0.0604379	0.0063532	9.513	< 2e-16 ***
	HAL 3.34412	0.15746	21.237	<2e-16 ***	PFE 0.0861619	0.0099716	8.641	< 2e-16 ***
	BDX 1.96393	0.18408	10.669	<2e-16 ***	CINF 0.0971844	0.0122222	7.951	2.71e-14 ***
	HD 2.30237	0.14093	16.337	<2e-16 ***	AXP 0.1082264	0.0115565	9.365	< 2e-16 ***
	C 0.34954	0.01933	18.084	<2e-16 ***	TIF 0.0609005	0.0079498	7.661	1.92e-13 ***
	CSCO 4.97454	0.07961	62.488	<2e-16 ***	VRSN 0.0278347	0.0042177	6.600	1.57e-10 ***
	EMN 4.36218	0.26057	16.741	<2e-16 ***	GS 0.0710176	0.0102350	6.939	1.99e-11 ***
Adjusted R^2	0.9962				0.9295			

Table Ap.1

We both choose 10 stocks using stepwise, the model based on price has an adjusted R^2 of 0.9962, while the model based on return has an adjusted R^2 of 0.9295. This is because the returns are more volatile than price.

Figure Ap.1 and Ap.2 are in sample and out-sample plots of tracking portfolio's return vs. index's return.

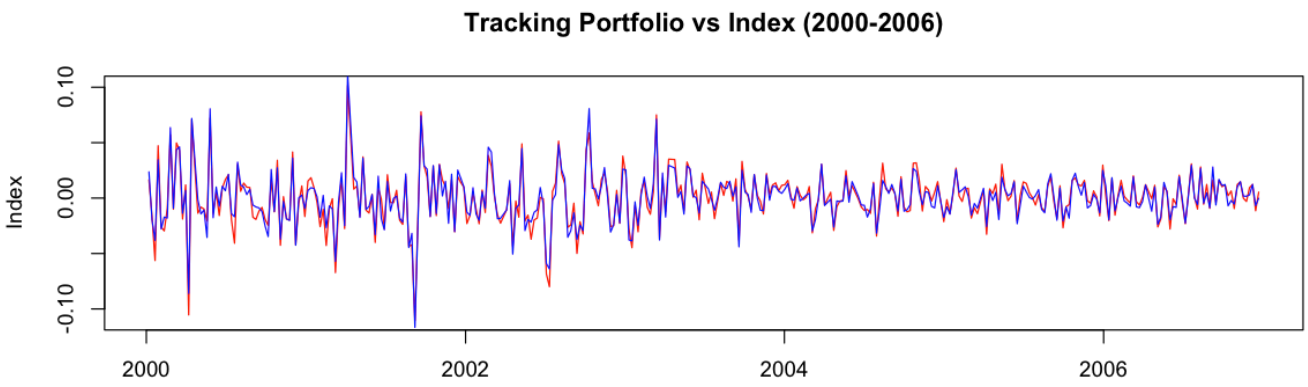

Figure. Ap.1

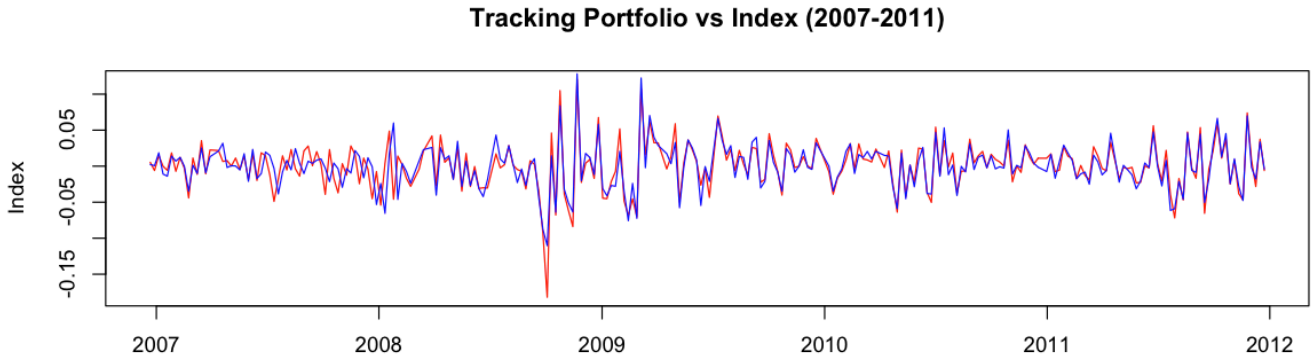

Figure. Ap.2

Figure Ap.3 shows the compound returns of tracking portfolio and index through 2007-2011, and the deviation plot. Compared to model based on price, they are quite similar.

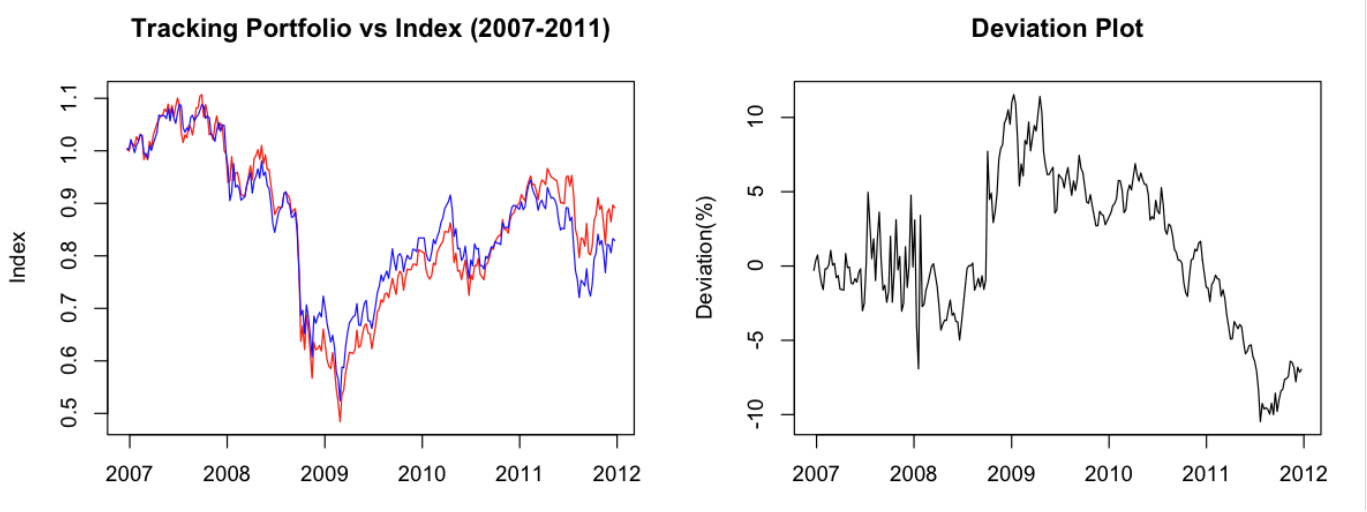

Figure Ap.3