Stock Selection Strategy

Xiaoyu Liu

1 Introductioin

Factor is short for influencing factor, or simply understood as index. We all know that stock returns are affected by multiple factors, such as macro, industry, liquidity, company fundamentals, trading sentiment, and so on. The so-called "multi-factor model" is simply to find those factors that are most relevant to stock returns, and use these factors (factors or indicators) to describe stock returns and select stocks.

Multi-factor model is one of the most widely used and most mature quantitative stock selection models in the field of quantitative investment. It is based on modern financial investment theories such as portfolio, capital asset pricing (CAPM) and arbitrage pricing theory (APT). The multi-factor model assumes that the market is inefficient or weakly efficient and obtains excess returns through active portfolio management. The core idea of multi-factor stock selection is that market forces are multiple and dynamic, but there are always factors that are stable over time. In the practice of quantification, different multi-factor models are constructed because different market participants or analysts have different understandings of market dynamics and factors.

2 Data Preprocessing

2.1 Data resources

Our datasets are downloaded from [CSMAR][http://www.gtarsc.com/#/index]. The first dataset, named Fivefac, comes from stocks with different market types. Fivefac consists of 20961 samples and 12 variables, including trading date, portfolios and 10 factors. The second dataset, named SSE50, covers different stocks and their closing price. Fivefac consists of 52722 rows and 3 columns, including trading date and closing price.

2.2 Data processing

The following packages will be used in this project:

```
library(quantmod)
library(ggplot2)
library(reshape2)
library(farver)
library(dplyr)
```

Data 1: Stocks and factors

```
setwd('/Users/liuxiaoyu/Desktop/RUC/Advanced_applied_statistics/1st term/project/Fac')
Fac <- read.table('Fivefac.csv',sep="\t",header=T,fileEncoding="UCS-2LE",stringsAsFactors = F)
head(Fac)
```

```
##   MarkettypeID TradingDate Portfolios RiskPremium1 RiskPremium2      SMB1
## 1        P9709  2016-06-30          1     0.000969    -0.000031  0.001268
```

```
## 2        P9709  2016-06-30          2     0.000969    -0.000031  0.001136
## 3        P9709  2016-06-30          3     0.000969    -0.000031  0.001717
## 4        P9709  2016-06-29          1     0.004969     0.004969 -0.003154
## 5        P9709  2016-06-29          2     0.004969     0.004969 -0.003180
## 6        P9709  2016-06-29          3     0.004969     0.004969 -0.003198
##        SMB2      HML1      HML2      RMW1      RMW2      CMA1      CMA2
## 1  0.000969 -0.001541 -0.000689  0.000548  0.001227  0.002268  0.002815
## 2  0.000857 -0.000003  0.000085 -0.000101  0.000107  0.001627  0.002121
## 3  0.001698 -0.000865 -0.000152 -0.000263  0.000057  0.001752  0.002164
## 4 -0.003181  0.003817  0.004004 -0.001822 -0.002097 -0.000526  0.000133
## 5 -0.003019  0.002722  0.002881 -0.000911 -0.000700 -0.001249 -0.001022
## 6 -0.003055  0.002378  0.002492 -0.001388 -0.001357 -0.002020 -0.001691
```

Let us have a quick look at our data, especially the variables:

```r
str(Fac)
```

```
## 'data.frame':    20961 obs. of  13 variables:
##  $ MarkettypeID: chr  "P9709" "P9709" "P9709" "P9709" ...
##  $ TradingDate : chr  "2016-06-30" "2016-06-30" "2016-06-30" "2016-06-29" ...
##  $ Portfolios  : int  1 2 3 1 2 3 2 1 3 1 ...
##  $ RiskPremium1: num  0.000969 0.000969 0.000969 0.004969 0.004969 ...
##  $ RiskPremium2: num  -0.000031 -0.000031 -0.000031 0.004969 0.004969 ...
##  $ SMB1        : num  0.00127 0.00114 0.00172 -0.00315 -0.00318 ...
##  $ SMB2        : num  0.000969 0.000857 0.001698 -0.003181 -0.003019 ...
##  $ HML1        : num  -0.001541 -0.000003 -0.000865 0.003817 0.002722 ...
##  $ HML2        : num  -0.000689 0.000085 -0.000152 0.004004 0.002881 ...
##  $ RMW1        : num  0.000548 -0.000101 -0.000263 -0.001822 -0.000911 ...
##  $ RMW2        : num  0.001227 0.000107 0.000057 -0.002097 -0.0007 ...
##  $ CMA1        : num  0.002268 0.001627 0.001752 -0.000526 -0.001249 ...
##  $ CMA2        : num  0.002815 0.002121 0.002164 0.000133 -0.001022 ...
```

Then we extract certain part of the original dataset, that is a specific type of stock P9709 and five factors renamed as 'Trddt','MAR','SMB','HML', 'RMW' and 'CMA'.

```r
Fac <- Fac[Fac$MarkettypeID == 'P9709'& Fac$Portfolios == 1,c(2,4,6,8,10,12)]
Fac$TradingDate <- as.Date(Fac$TradingDate)
colnames(Fac) <- c('Trddt','MAR','SMB','HML','RMW','CMA')
head(Fac)
```

```
##         Trddt       MAR       SMB       HML       RMW       CMA
## 1  2016-06-30  0.000969  0.001268 -0.001541  0.000548  0.002268
## 4  2016-06-29  0.004969 -0.003154  0.003817 -0.001822 -0.000526
## 8  2016-06-28  0.006969  0.005902 -0.005268 -0.001505  0.000915
## 10 2016-06-27  0.016969  0.006622 -0.008590 -0.007090 -0.001344
## 14 2016-06-24 -0.010031  0.003053  0.001309  0.003050 -0.000368
## 18 2016-06-23 -0.003031  0.001458 -0.001011 -0.001303 -0.001958
```

Data 2: Stocks and Returns

```r
setwd('/Users/liuxiaoyu/Desktop/RUC/Advanced_applied_statistics/1st term/project/Fac')
stk <- read.table('SSE50.csv',sep="\t",header=T,fileEncoding="UCS-2LE",stringsAsFactors = F)
stk$Stkcd <- as.character(stk$Stkcd)
stk$Trddt <- as.Date(stk$Trddt)
head(stk)
```

```
##    Stkcd      Trddt Clsprc
```

```
## 1 600000 2014-08-01   9.76
## 2 600000 2014-08-04   9.92
## 3 600000 2014-08-05   9.87
## 4 600000 2014-08-06   9.74
## 5 600000 2014-08-07   9.56
## 6 600000 2014-08-08   9.54
```

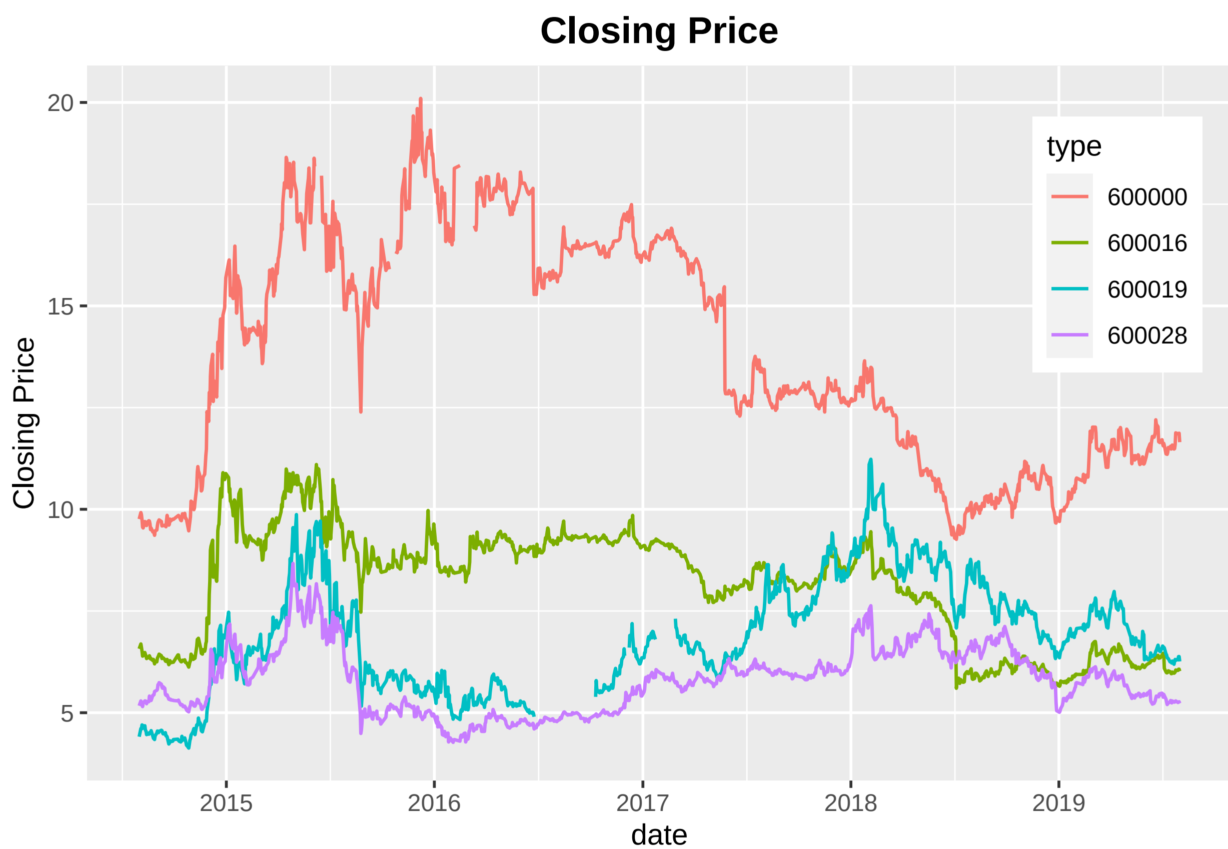
For the convenience of analysis, we transpose the data into the following format, with all stocks as columns.

```
stkCls <- dcast(stk,Trddt~Stkcd,value.var = 'Clsprc')
stkCls[1:5,1:5]
```

```
##        Trddt 600000 600016 600019 600028
## 1 2014-08-01   9.76   6.57   4.41   5.17
## 2 2014-08-04   9.92   6.69   4.61   5.29
## 3 2014-08-05   9.87   6.61   4.60   5.24
## 4 2014-08-06   9.74   6.51   4.70   5.22
## 5 2014-08-07   9.56   6.39   4.69   5.15
```

The following picture demonstrates four stocks, including 600000, 600016, 600019 and 600028, and their closing prices.

```
da <- stkCls[,1]
title <- 'Closing Price'
p1<-data.frame(date=da,ma=stkCls[,2],type=rep('600000',length(da)))
p2<-data.frame(date=da,ma=stkCls[,3],type=rep('600016',length(da)))
p3<-data.frame(date=da,ma=stkCls[,4],type=rep('600019',length(da)))
p4<-data.frame(date=da,ma=stkCls[,5],type=rep('600028',length(da)))
pdata<-rbind(p1,p2,p3,p4)
ggplot(pdata,aes(x=date, y=ma,color=type))+
    theme(legend.position=c(0.9,0.75))+geom_line(size=0.6)+ylab('Closing Price')+
    ggtitle(title)+
    theme(plot.title=element_text(size=14,hjust=0.5,colour='black',face='bold'))
```



The picture tells us that among the four stocks, it is 600000 whose price changes most dramatically. Next we turn to compute and analyze the returns, which is of our interest, rather than the closing price itself.

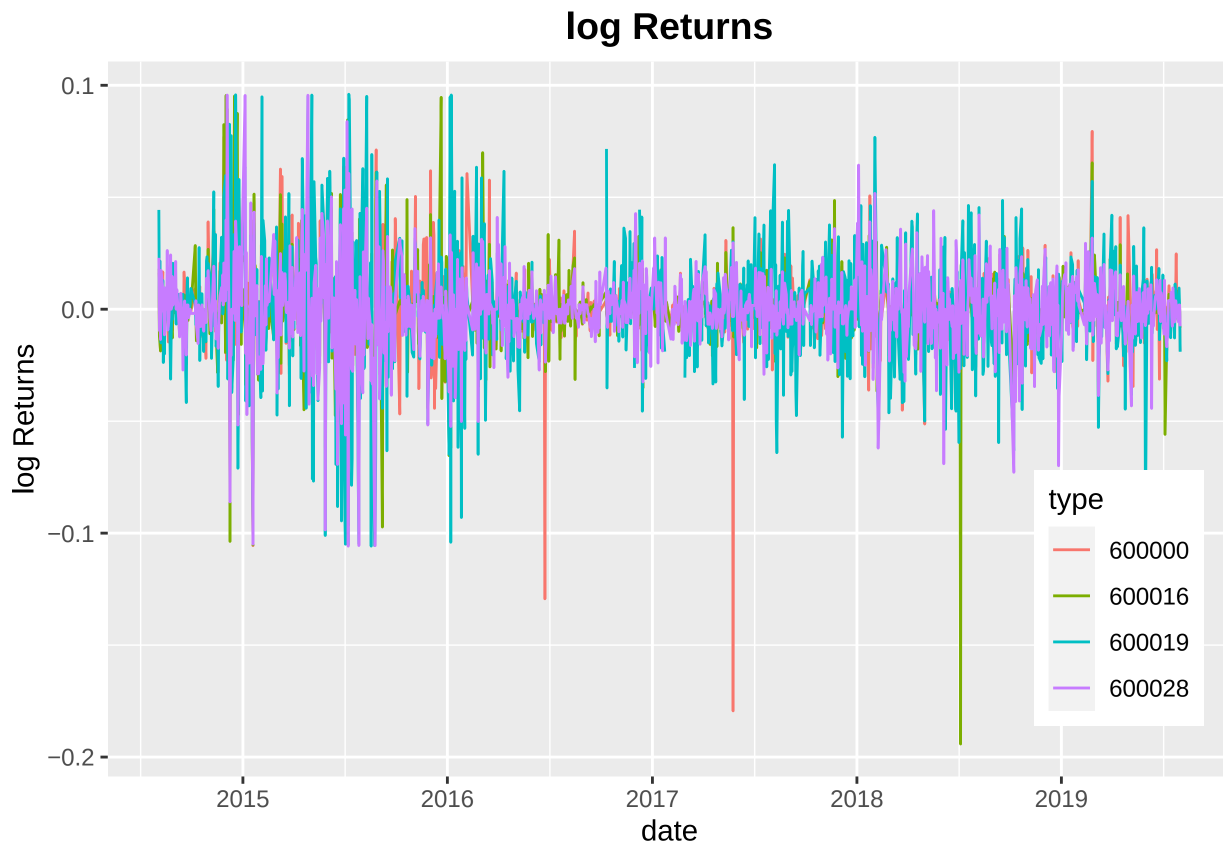
```
names <- colnames(stkCls)
myfun <- function(x){x <- c(NA,diff(log(x)))}
stkRet <- as.data.frame(apply(stkCls[2:ncol(stkCls)],2,myfun))
stkRet <- cbind(stkCls[,1],stkRet)
colnames(stkRet) <- names
stkRet$Trddt <- as.Date(stkRet$Trddt)
stkRet[1:5,1:5]
```

```
##        Trddt       600000      600016       600019       600028
## 1 2014-08-01           NA          NA           NA           NA
## 2 2014-08-04  0.016260521  0.01810004  0.044353168  0.022945557
## 3 2014-08-05 -0.005053068 -0.01203022 -0.002171554 -0.009496748
## 4 2014-08-06 -0.013258736 -0.01524420  0.021506205 -0.003824096
## 5 2014-08-07 -0.018653391 -0.01860519 -0.002129926 -0.013500687
```

Also, we present the log returns for the same four stocks.

```
da <- stkRet[-1,1]
title <- 'log Returns'
r1<-data.frame(date=da,ma=stkRet[-1,2],type=rep('600000',length(da)))
r2<-data.frame(date=da,ma=stkRet[-1,3],type=rep('600016',length(da)))
r3<-data.frame(date=da,ma=stkRet[-1,4],type=rep('600019',length(da)))
r4<-data.frame(date=da,ma=stkRet[-1,5],type=rep('600028',length(da)))
rdata<-rbind(r1,r2,r3,r4)
ggplot(rdata,aes(x=date, y=ma,color=type))+
    theme(legend.position=c(0.9,0.25))+geom_line(size=0.5)+ylab('log Returns')+
```

```r
    ggtitle(title)+
    theme(plot.title=element_text(size=14,hjust=0.5,colour='black',face='bold'))
```



Merge Data 1 and 2

```r
Fac <- Fac[Fac$Trddt %in% stkRet$Trddt,]
dt <- merge(Fac,stkRet)
dt[1:5,1:12]
```

```
##        Trddt       MAR       SMB       HML       RMW       CMA       600000
## 1 2014-08-01 -0.008072 -0.001871  0.000157  0.004936 -0.001748           NA
## 2 2014-08-04  0.016928  0.000557  0.001639 -0.000003 -0.000895  0.016260521
## 3 2014-08-05  0.000928  0.006372 -0.001631 -0.004278  0.000748 -0.005053068
## 4 2014-08-06 -0.000072  0.002378 -0.003937 -0.008152  0.003019 -0.013258736
## 5 2014-08-07 -0.013072  0.004299 -0.002968 -0.002810  0.000028 -0.018653391
##        600016       600019       600028       600029      600030
## 1          NA           NA           NA           NA          NA
## 2  0.01810004  0.044353168  0.022945557  0.016129382  0.05901864
## 3 -0.01203022 -0.002171554 -0.009496748  0.003992021 -0.01331381
## 4 -0.01524420  0.021506205 -0.003824096  0.015810606  0.00297398
## 5 -0.01860519 -0.002129926 -0.013500687 -0.007874056 -0.02938064
```

Take the stock 600000 as an example, we now present the intuitive correlation of returns and five factors.

```r
qq <- quantile(dt$`600000`, seq(0, 1, 0.2), na.rm = TRUE)
qq
```

```
##           0%          20%          40%          60%          80%         100%
```

```
## -0.179352472 -0.008887726 -0.001979587  0.002346780  0.009540151  0.081640964
```

```
mutate(dt, return.quint = cut(`600000`, qq)) %>%
  group_by(return.quint) %>%
  summarize(f1 = mean(MAR, na.rm = TRUE),
            f2 = mean(SMB, na.rm = TRUE),
            f3 = mean(HML, na.rm = TRUE),
            f4 = mean(RMW, na.rm = TRUE),
            f5 = mean(CMA, na.rm = TRUE))
```

```
## Warning: Factor `return.quint` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 6 x 6
##   return.quint                f1        f2        f3          f4         f5
##   <fct>                    <dbl>     <dbl>     <dbl>       <dbl>      <dbl>
## 1 (-0.179,-0.00889]   -0.0132     0.00288  -0.00182  -0.000626   0.000404
## 2 (-0.00889,-0.00198] -0.00126    0.00189  -0.00152  -0.000754   0.0000122
## 3 (-0.00198,0.00235]   0.000673   0.00104  -0.000204 -0.000265   0.000182
## 4 (0.00235,0.00954]    0.00362   -0.000199  0.000537  0.0000638  0.000162
## 5 (0.00954,0.0816]     0.0123    -0.00436   0.00328   0.00190   -0.000914
## 6 <NA>                 0.00218    0.000470  0.000975 -0.000277   0.000517
```

3 Fama-French Five Factor model

Step1: The Fama-Frentch five-factor model was used to regression the return series of 50 stocks in Shanghai in the first 100 trading days to obtain the corresponding α of each stock. More precisely,

$$R_t = \alpha + m \cdot MAR_t + s \cdot SMB_t + h \cdot HML_t + r \cdot RMW_t + c \cdot CMA_t + e_t$$

Step 2: Rank the alpha of each stock, taking the top five stocks with alpha largest.

Step 3: Equal-weighted allocation of 5 stocks acquired and held for 30 trading days.

Step 4: Portfolio reallocation every 30 days.

Step 5: Return to Step 1.

The algorithm can be presented as in the following figure.

```
stk_Sel <- function(dtSmp){
  sym <- rep(NA,100)
  alpha <- rep(NA,100)
  tgstk <- data.frame(sym,alpha)
  for(i in 7:(ncol(dtSmp)-1)){
    tra <- dtSmp[,c(1:6,i)]
    colnames(tra) <- c(names(Fac),'logRet')
    glm <- glm(formula = 'logRet~MAR+SMB+HML+RMW+CMA',data = tra)
    tgstk$sym[i] <- colnames(dtSmp)[i]
    tgstk$alpha[i] <- glm$coefficients[1]
  }
  tgstk<- na.omit(tgstk[order(-tgstk$alpha),])[1:5,]
  return(tgstk$sym)
}

dt1 <- NA
for(i in seq(101,nrow(dt),30)){
```

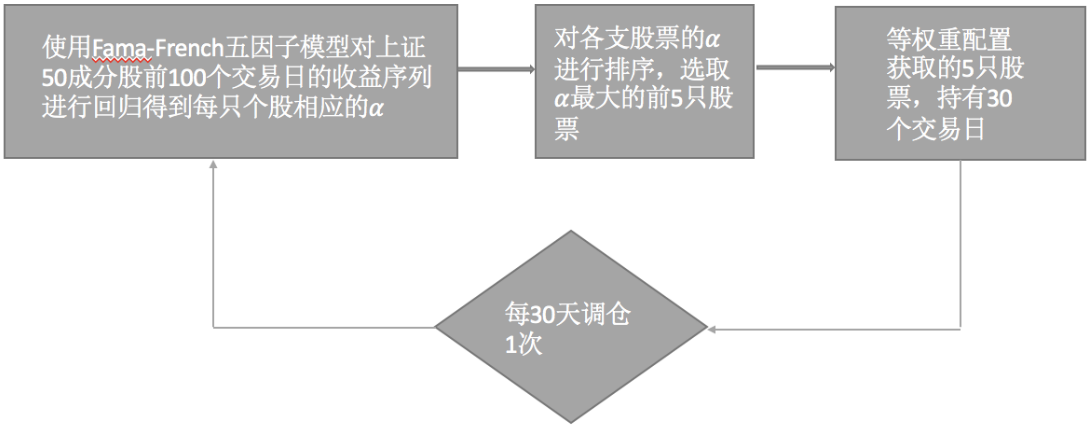


Figure 1: Fama-French stock selection strategy

```r
  tra <- dt[(i-100):(i-1),]
  stkSel <- stk_Sel(tra)
  tra2 <-dt[(i):(i+29),]
  tra2$stk1 <- stkSel[1]
  tra2$stk2 <- stkSel[2]
  tra2$stk3 <- stkSel[3]
  tra2$stk4 <- stkSel[4]
  tra2$stk5 <- stkSel[5]
  tra2$ret_daily <- rowSums(tra2[,colnames(tra2)%in%stkSel],na.rm = T)
  dt1 <- rbind(dt1,tra2)
}
dt1 <- dt1[-1,]
Trade <- dt1[,c(1,(ncol(dt1)-6):ncol(dt1))]
Trade <- na.omit(Trade)
```

5 Results

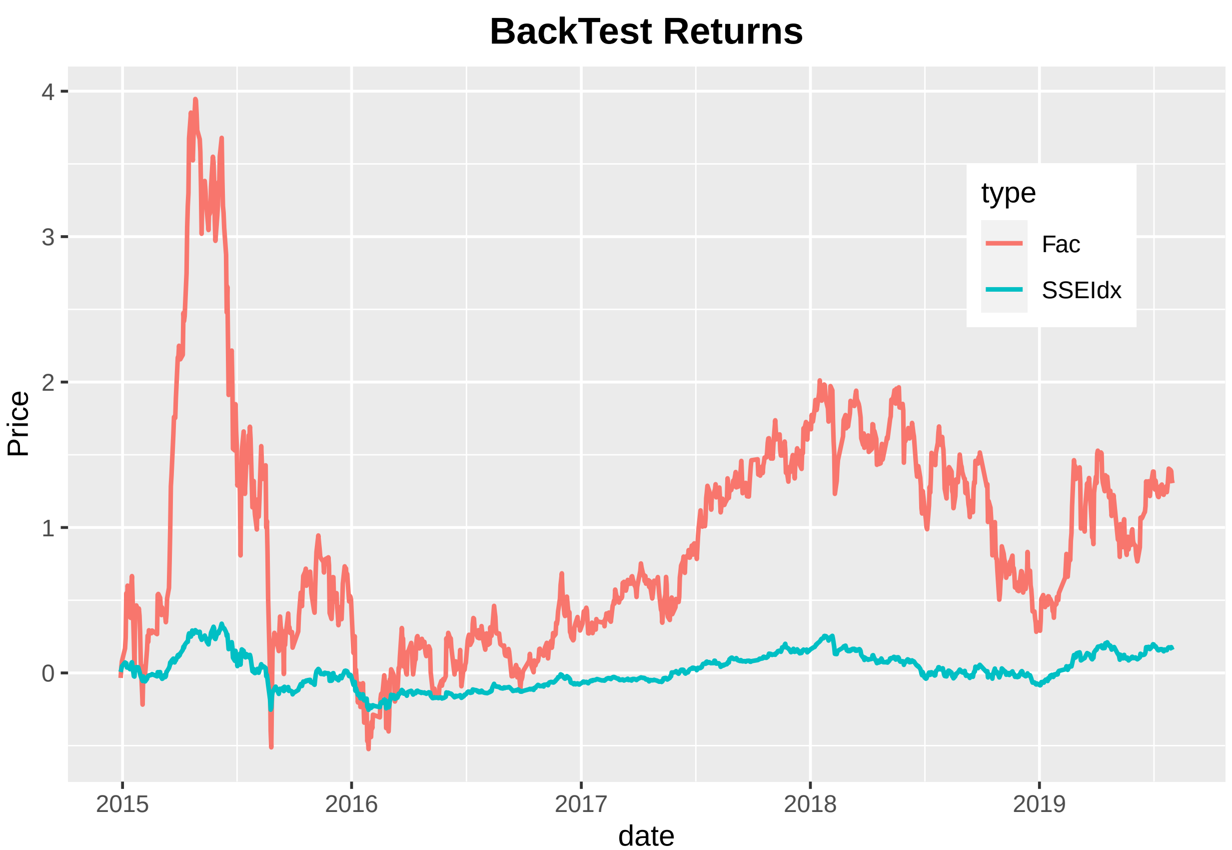
```r
Trade$ret_acml <- 0
Trade$ret_SSEIdx <- 0
for(i in 1:nrow(Trade)){
  Trade$ret_acml[i] <- (sum(Trade$ret_daily[1:i]))
  Trade$ret_SSEIdx[i] <- sum(Trade$SSEIdx[1:i])
}

head(Trade)
```

```
##          Trddt       SSEIdx   stk1   stk2   stk3   stk4   stk5   ret_daily
## 101 2014-12-29  0.006808909 601800 601688 600837 600030 601766 -0.03458773
## 102 2014-12-30  0.016770650 601800 601688 600837 600030 601766  0.07264898
## 103 2014-12-31  0.022699400 601800 601688 600837 600030 601766  0.04116241
## 104 2015-01-05  0.026026041 601800 601688 600837 600030 601766  0.08881439
## 105 2015-01-06 -0.007668098 601800 601688 600837 600030 601766  0.07434736
## 106 2015-01-07  0.002214781 601800 601688 600837 600030 601766  0.30505095
```

```
##         ret_acml  ret_SSEIdx
## 101 -0.03458773 0.006808909
## 102  0.03806125 0.023579559
## 103  0.07922366 0.046278958
## 104  0.16803805 0.072304999
## 105  0.24238541 0.064636901
## 106  0.54743636 0.066851682
```

Trade data set shows our stock selection process, each selected stock combination will be held for 30 days.

```r
Lokup <- function(Trade){
  dd <- Trade$Trddt
  title <- 'BackTest Returns'
  k1<-data.frame(date=dd,ma=Trade$ret_acml,type=rep('Fac',length(dd)))
  k2<-data.frame(date=dd,ma=Trade$ret_SSEIdx,type=rep('SSEIdx',length(dd)))
  kdata<-rbind(k1,k2)
  ggplot(kdata,aes(x=date, y=ma,color=type))+
    theme(legend.position=c(0.85,0.75))+geom_line(size=0.8)+ylab('Price')+
    ggtitle(title)+
    theme(plot.title=element_text(size=14,hjust=0.5,colour='black',face='bold'))
}
Lokup(Trade)
```



6 Discussion and Future Plan

From the backtest chart, we can also see that when using the Fama-French-based stock selection strategy, our return is greater than the average return of the Shanghai 50 constituent stocks. However, we mention that this strategy is far from satisfactory. First of all, the difference of log return between our selected stocks and Shanghai 50 constituent stocks is not numerically obvious at. Secondly, the performance of stocks selected by stock selection strategies is significantly more unstable, that is, stockholders need to bear greater risks, which is obviously unacceptable to many risk aversion. Therefore, in future work we will choose a more advanced model as our new stock selection strategy.