**# 将目录设定成你的数据集所在的目录**

**setwd("C:/STONY/Practice/R (No.6)")**

**install.packages("quantmod")**

**library(quantmod)**

**library(ggplot2)**

**library(dplyr)**

**## 这里是利用你们老师的方法提取数据，你不用管它，我已经提出来了**

**####### 1 #######**

**symbols=c("MSFT", "IBM") #Assume we retrieve data of two stocks**

**getSymbols(symbols,from="2016-06-24",to="2018-06-24",periodicity="weekly",**

**return.class="ts")**

**mydat<-cbind(MSFT[,6], IBM[,6])**

**colnames(mydat)<-c("MSFT","IBM")**

**#save the data back to your computer**

**write.csv(mydat, file="stock\_data.csv")**

**## 重新读取已经提取出来的数据**

**### read the data ###**

**data <- read.csv("stock\_data.csv",header=T)**

**str(data)**

**## 求出每个公司每周的return**

**### calculate the MSFT return**

**return1 <- c()**

**for (i in 2:105){**

**if(i==1){**

**return1[i] <- data[,2][1] - data[,2][1]**

**}**

**else{**

**return1[i] <- (data[,2][i] - data[,2][i-1])/data[,2][i]**

**}**

**}**

**### calculate the IBM return**

**return2 <- c()**

**for (i in 2:105){**

**if(i==1){**

**return2[1] <- data[,3][1] - data[,3][1]**

**}**

**else{**

**return2[i] <- (data[,3][i] - data[,3][i-1])/data[,3][i]**

**}**

**}**

**## 因为数据里没有日期，所以我添加了一列日期，并且输入了正确的年月日**

**# generate one column "date" which represents the date**

**install.packages("lubridate")**

**library(lubridate)**

**Date <- list()**

**for (i in 1:105){**

**Date[[i]] <- ymd(20160617) + days(7 %\*% i)**

**}**

**data$date <- unlist(Date) %>% as.Date()**

**## 第一问**

**## 读取影响股票收益的数据，并且将股票return加入该表中**

**## read the factor data and merge factor and stock returns together**

**factor <- read.csv("Factors.csv",header=T)**

**factor$Date <- factor$Date %>% as.numeric()**

**factor$Date <- as.Date(as.character(factor$Date), "%Y%m%d")**

**factor <- filter(factor,Date >= "2016-06-24"& Date <= "2018-06-24")**

**new\_dat <-cbind(return1,return2,factor)**

**colnames(new\_dat) <- c("M\_rtn","I\_rtn","date","RMRF","SMB","HML",'RF')**

**new\_dat[,c(1,2)][1,] <- c(0,0)**

**## 第二问**

**## 画出每个公司股票return的柱状图，横坐标代表return，纵坐标代表出现的频数，可以看出每个公司return的一个大体分布**

**### histogram of stock's returns ###**

**ggplot(new\_dat, aes(x = I\_rtn)) +**

**geom\_histogram(fill = "lightblue", colour = "black",bins = 20) +**

**labs(title = "Histogram of stock returns Of IBM")**

**ggplot(new\_dat, aes(x = M\_rtn)) +**

**geom\_histogram(fill = "lightblue", colour = "black",bins = 20) +**

**labs(title = "Histogram of stock returns Of MSFT")**

**## 画出每个公司股票return的时间序列图，横坐标为时间，纵坐标为return，可以看出其公司return的未来趋势和走向规律**

**### time series of stock's returns ###**

**library(xts)**

**hh <- xts(new\_dat$I\_rtn, as.Date(new\_dat$date, format='%Y/%m/%d'))**

**win.graph(width = 9.5,height = 4.5,pointsize = 8)**

**plot(hh,type = 'l',main=' ')**

**title("Timeseries of stock returns of IBM")**

**kk <- xts(new\_dat$M\_rtn, as.Date(new\_dat$date, format='%Y/%m/%d'))**

**win.graph(width = 9.5,height = 4.5,pointsize = 8)**

**plot(kk,type = 'l',main=' ')**

**title("Timeseries of stock returns of MSFT")**

**## 第三问**

**## 画出这些影响因子间的pairwise图**

**####### 3 #######**

**## 这里画的是return 和这些影响因素间的关系图，可以看出一个大致规律，有些可能呈线性，有些可能相关性不大**

**### pairwise scatter plot ###**

**plot(I\_rtn~RMRF,data=new\_dat)**

**plot(I\_rtn~SMB,data=new\_dat)**

**plot(I\_rtn~HML,data=new\_dat)**

**plot(M\_rtn~RMRF,data=new\_dat)**

**plot(M\_rtn~SMB,data=new\_dat)**

**plot(M\_rtn~HML,data=new\_dat)**

**## 画出所有变量的pairwise，每行图代表一个变量与其他变量之间的关系，血有字母的是自己与自己的关系，所以用名字替代，例如SMB**

**# pairwise**

**# among the 3 factors**

**pairs(new\_dat[,c(5,6,4)])**

**# among MSFT return and 3 factor**

**pairs(new\_dat[,c(1,5,6,4)])**

**# among IBM return and 3 factor**

**pairs(new\_dat[,c(2,5,6,4)])**

**## 算出每个变量之间的相关性系数**

**### correlation function**

**cor <- cor(new\_dat[,c(5,6,4)])**

**install.packages("corrplot")**

**library(corrplot)**

**corrplot(cor, method="circle") # the darker color is, the greater the correlation will be**

**corrplot(cor, method="number")**

**## 第四问**

**## 计算出每个公司的excess return。excess return就是return减去无风险利率**

**####### 4 #######**

**### regression model**

**## 微软的excess return 并建立线性回归模型**

**excs\_rtn1 <- new\_dat[,1] - new\_dat$RF**

**model1 <- {excs\_rtn1 ~ RMRF+SMB+HML}**

**fit1 <- lm(model1,data=new\_dat)**

**## IBM的excess return并建立线性回归模型**

**excs\_rtn2 <- new\_dat[,2] - new\_dat$RF**

**model2 <- {excs\_rtn2 ~ RMRF+SMB+HML}**

**fit2 <- lm(model2,data=new\_dat)**

**## 第五问**

**## 利用ANOVA 表判断模型里重要的影响因素是谁，结果有\*就代表显著，挑选出显著的影响因子**

**####### 5 #######**

**summary(fit1)**

**anova(fit1)**

**## the ANOVA table shows that RMRF and SMB are significant in this regression on MSFT stock**

**summary(fit2)**

**anova(fit2)**

**# the ANOVA table shows that RMRF and HML are significant in this regression on IBM stock**

**## 第六问**

**## 通过step()逐步回归模型的公式，我们找到每个公司的best model，并分析model里公司的利益率受哪一个因子的影响比较大。RMRF代表市场，SMB代表大小，HML代表value。公司受什么样的因素影响，就证明其是什么类型的股票，以下有我对结果的分析。**

**####### 6 #######**

**step(fit1)**

**## find out the best model for MSFT**

**M\_excs <- lm(formula = excs\_rtn1 ~ RMRF + SMB, data = new\_dat)**

**summary(M\_excs)**

**#Adjusted R-squared: 0.4748 , RMRF is \*\*\* ,SMB \*\* significant for MSFT**

**#This means the trend of the stock follows the market trend. Here SMB is less than 0, this shows**

**#Microfost's excess earning is affected by large-cap stocks.**

**#the value of MSFT larger, the more excess earning MSFT will get**

**step(fit2)**

**## find out the best model for IBM**

**I\_excs <- lm(formula = excs\_rtn2 ~ RMRF + SMB + HML, data = new\_dat)**

**summary(I\_excs)**

**#Adjusted R-squared: 0.3653, RMRF is \*\*\* ,HML \* significant for MSFT**

**#Here SMB is not significant for Pvalue is larger than 0.1**

**#the value of MSFT larger, the more excess earning MSFT will get**

**## 第七问**

**## 将上述的结果与市场模型进行比较，传统的市场模型仅仅取决于市场这一因素，而不是三因素模型。这里我们建立市场模型，通过Adjusted R-squared对模型进行比较，三因素模型的R square在上述第六问得出，越大，证明模型越好**

**##### 7 #######**

**### market model**

**M\_mrk <- lm(formula = excs\_rtn1 ~ RMRF, data = new\_dat)**

**summary(M\_mrk)**

**# Adjusted R-squared: 0.4284**

**I\_mrk <- lm(formula = excs\_rtn2 ~ RMRF, data = new\_dat)**

**summary(I\_mrk)**

**# Adjusted R-squared: 0.329**

**# we can see 3 factor model outperforms the market model,cause its R square is larger than market model.**

**# This means 3 facotr model fits the data better.**

**## 第八问**

**## 对建立的两个公司的三因素模型进行残差分析，如果残差符合正太分布，那么说明该模型是合理的。以下结果说明模型可用**

**####### 8 ######**

**### check assumptions ###**

**### MSFT**

**par(mfrow=c(2,2))**

**plot(M\_excs)**

**### IBM**

**par(mfrow=c(2,2))**

**plot(I\_excs)**

**## plot the residual diagnosis**

**install.packages("car")**

**library(car)**

**# test of independence**

**durbinWatsonTest(M\_excs) #p-value = 0.016 independent in the 99% confidence interval**

**durbinWatsonTest(I\_excs) #p-value = 0.136 independent in the 90%,95%,99% confidence interval**

**# test of normality**

**shapiro.test(M\_excs$residuals) #p-value = 0.2232 normal in the 90%,95%,99% confidence interval**

**shapiro.test(I\_excs$residuals) #p-value = 2.739e-05 abnormal in the 90%,95%,99% confidence interval**

**# test of homoscedasticity**

**ncvTest(M\_excs) #p-value = 0.95385 homo in the 90%,95%,99% confidence interval**

**ncvTest(I\_excs) #p-value = 0.18437 homo in the 90%,95%,99% confidence interval**