46-926 Homework #3, Part I

Jingyi Guo, Pittsburgh Campus

2/3/2017

I choose to analyze the excess return of PNC.

Get Data

```
library(quantmod)
source("http://www.stat.cmu.edu/~cschafer/MSCF/getFamaFrench.txt")
ffhold=getFamaFrench(from="2012-1-1",to="2012-6-30")
PNC = getSymbols("PNC", from="2012-1-1", to="2012-6-30", auto.assign=F)
ffhold$PNCexret = 100*dailyReturn(PNC) - ffhold$RF
yielddata = read.table(
  "http://www.stat.cmu.edu/~cschafer/MSCF/YieldCurves2012.txt", header=T)
yielddata$Date = as.Date(as.character(yielddata$Date), format="%m/%d/%y")
keep = yielddata$Date <= "2012-6-30" & yielddata$Date >= "2012-1-1" &
  yielddata$Date != "2012-4-6"
yielddatasub = yielddata[keep,]
yieldcurves=yielddatasub[,2:12]
```

Models

Model 1

```
fitmodel1=lm(PNCexret ~ Mkt.RF, data=ffhold)
```

Model 2

```
fitmodel2=lm(PNCexret ~ Mkt.RF + SMB + HML, data=ffhold)
```

Model 3

```
pcaout=princomp(yieldcurves)
PCA1=pcaout$scores[,1]
PCA2=pcaout$scores[,2]
PCA3=pcaout$scores[,3]
fitmodel3=lm(PNCexret ~ Mkt.RF + SMB + HML + PCA1 + PCA2 + PCA3, data=ffhold)
```

Model 4

```
fitmodel4=lm(PNCexret ~ Mkt.RF + SMB + HML +pcaout$scores[,1:6], data=ffhold)
```

1. Summary

```
summary(fitmodel1)
```

```
##
## Call:
## lm(formula = PNCexret ~ Mkt.RF, data = ffhold)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0082 -0.5185  0.0816  0.5577  3.1729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04943    0.08748  -0.565    0.573
## Mkt.RF       1.18436    0.09848  12.026   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9745 on 123 degrees of freedom
## Multiple R-squared:  0.5404, Adjusted R-squared:  0.5367
## F-statistic: 144.6 on 1 and 123 DF,  p-value: < 2.2e-16
```

```
summary(fitmodel2)
```

```
##
## Call:
## lm(formula = PNCexret ~ Mkt.RF + SMB + HML, data = ffhold)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1844 -0.3588  0.0651  0.4858  2.6850
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04113    0.07903  -0.520    0.604
## Mkt.RF       1.14211    0.10224  11.170  < 2e-16 ***
## SMB          0.07927    0.19963   0.397    0.692
## HML          1.09418    0.19976   5.478 2.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8795 on 121 degrees of freedom
## Multiple R-squared:  0.6317, Adjusted R-squared:  0.6226
## F-statistic: 69.19 on 3 and 121 DF,  p-value: < 2.2e-16
```

```
summary(fitmodel3)
```

```
##
```

```
## Call:
## lm(formula = PNCexret ~ Mkt.RF + SMB + HML + PCA1 + PCA2 + PCA3,
##     data = ffhold)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8634 -0.3875  0.0656  0.4612  2.4782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04078    0.07759  -0.526   0.6002
## Mkt.RF       1.13544    0.10144  11.193  < 2e-16 ***
## SMB          0.12914    0.19764   0.653   0.5148
## HML          1.09054    0.19652   5.549 1.79e-07 ***
## PCA1        -0.37420    0.18139  -2.063   0.0413 *
## PCA2        -0.42942    0.94297  -0.455   0.6497
## PCA3        -2.47048    1.41120  -1.751   0.0826 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8634 on 118 degrees of freedom
## Multiple R-squared:  0.6539, Adjusted R-squared:  0.6363
## F-statistic: 37.16 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
summary(fitmodel4)
```

```
##
## Call:
## lm(formula = PNCexret ~ Mkt.RF + SMB + HML + pcaout$scores[,
##     1:6], data = ffhold)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8620 -0.3742  0.0717  0.4182  2.5633
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.0419     0.0778  -0.539   0.5912
## Mkt.RF                         1.1496     0.1022  11.242  < 2e-16 ***
## SMB                            0.1087     0.1989   0.546   0.5859
## HML                            1.0662     0.1991   5.356 4.42e-07 ***
## pcaout$scores[, 1:6]Comp.1    -0.3714     0.1819  -2.042   0.0434 *
## pcaout$scores[, 1:6]Comp.2    -0.4313     0.9455  -0.456   0.6491
## pcaout$scores[, 1:6]Comp.3    -2.4828     1.4150  -1.755   0.0820 .
## pcaout$scores[, 1:6]Comp.4    -7.1798     5.3174  -1.350   0.1796
## pcaout$scores[, 1:6]Comp.5     1.7682     6.0821   0.291   0.7718
## pcaout$scores[, 1:6]Comp.6     4.9511     7.3670   0.672   0.5029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8657 on 115 degrees of freedom
## Multiple R-squared:  0.6609, Adjusted R-squared:  0.6344
## F-statistic: 24.91 on 9 and 115 DF,  p-value: < 2.2e-16
```

2. AIC

```
AIC(fitmodel1)
```

```
## [1] 352.2718
```

```
AIC(fitmodel2)
```

```
## [1] 328.5806
```

```
AIC(fitmodel3)
```

```
## [1] 326.8123
```

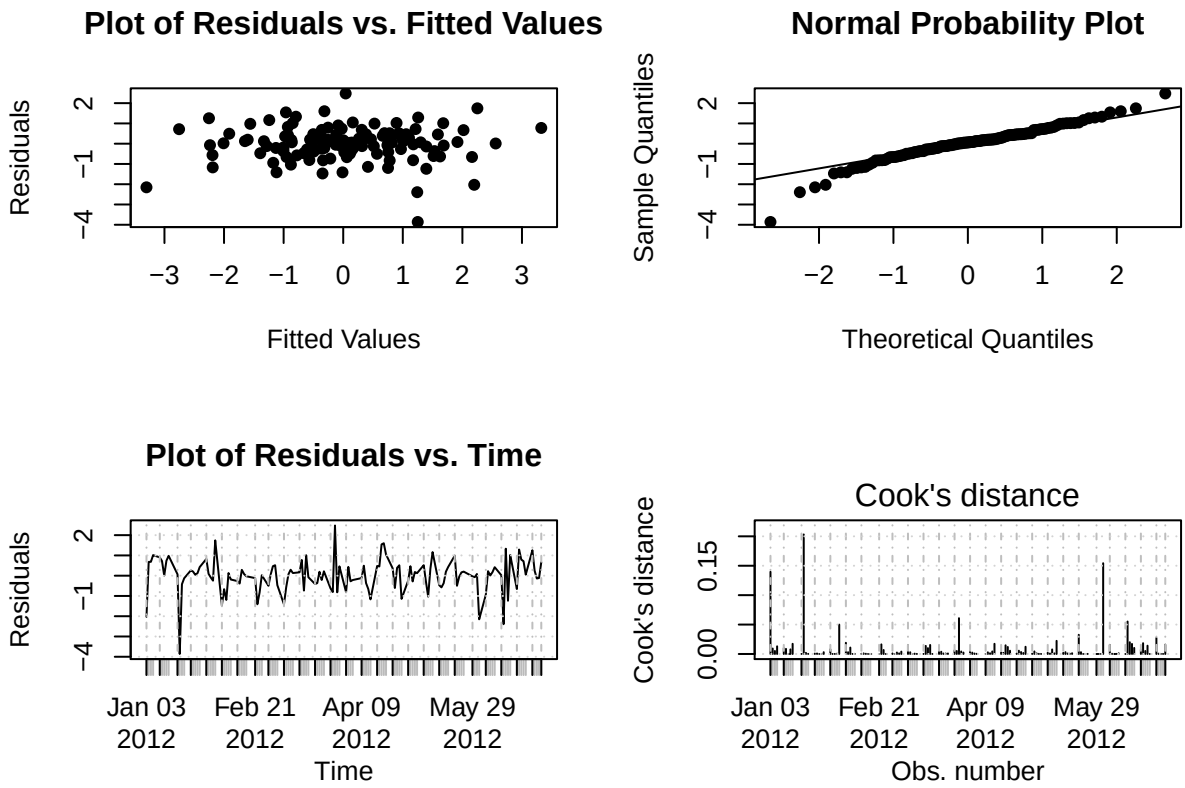
```
AIC(fitmodel4)
```

```
## [1] 330.2614
```

Since the AIC of the third model is the smallest, we conclude that Model 3 is the best choice using AIC criterion.

3. Diagonistic Plot

```
par(mfrow=c(2,2))
# plot of residuals vs. fitted values
plot(as.numeric(fitmodel3$fit),as.numeric(fitmodel3$resid),pch=16,xlab="Fitted Values",
     ylab="Residuals",main="Plot of Residuals vs. Fitted Values")
# normal prob. plot
qqnorm(as.numeric(fitmodel3$resid),pch=16,main="Normal Probability Plot")
qqline(as.numeric(fitmodel3$resid))
# plot of residuals vs. time
plot(fitmodel3$resid, xlab="Time",ylab="Residuals",pch=16,main="Plot of Residuals vs. Time")
# Plot of Cook Distance
plot(fitmodel3, which=4)
```




```
cookd = as.numeric(cooks.distance(fitmodel3))
sort(pf(cookd,7,118),decreasing=TRUE)[1:5]
```

```
## [1] 0.0158068374 0.0069099438 0.0050801022 0.0003470540 0.0002494861
```

Comment:

There is no prevalent pattern in the plot of residuals vs. fitted values, so Model 3 seems to be an acceptable fit. The normal distribution plot suggests that errors have distribution with tails significantly higher than normal distribution. The plot of residuals versus time does not reveal significant concern. In the Cook's Distance plot, two observations might be considered influential. The largest Cook's Distance is at the 1.6% of the F distribution. So there is not need for concern from influential dots.

4. Prediction

```
# obtain data
ffhold1=getFamaFrench(from="2012-7-1",to="2012-7-31")
keep1 = yielddata$Date <= "2012-7-31" & yielddata$Date >= "2012-7-1"
yielddatasub1 = yielddata[keep1,]
yieldcurves1=yielddatasub1[,2:12]
newcoords = predict(pcaout, yieldcurves1)
# prediction
result=predict.lm(fitmodel2,newdata=data.frame(ffhold1,PCA1=newcoords[,1],PCA2=newcoords[,2],
                                               PCA3=newcoords[,3]),interval="prediction")
print(result)
```

```
##               fit        lwr       upr
## 22757  0.35922389 -1.4079092 2.1263570
```

```
## 22758  0.95975113 -0.8057173 2.7252195
## 22759 -1.35825238 -3.1579031 0.4413984
## 22760 -0.80185766 -2.5680551 0.9643398
## 22761 -0.60891580 -2.3625479 1.1447163
## 22762 -0.70237685 -2.4634230 1.0586693
## 22763  0.64627692 -1.1310363 2.4235901
## 22764 -1.49197733 -3.2758327 0.2918781
## 22765  2.37997683  0.5744124 4.1855413
## 22766 -0.69029216 -2.4469991 1.0664147
## 22767  0.95482313 -0.8194214 2.7290677
## 22768 -0.03881912 -1.8169316 1.7392933
## 22769 -0.81012711 -2.6081318 0.9878776
## 22770 -1.20134000 -2.9614463 0.5587663
## 22771 -1.07398262 -2.8414359 0.6934706
## 22772 -0.91063934 -2.6766604 0.8553817
## 22773 -0.06476748 -1.8165221 1.6869871
## 22774  2.02426329  0.1936611 3.8548655
## 22775  2.33092470  0.5508986 4.1109508
## 22776  0.06396434 -1.6925365 1.8204652
## 22777 -0.28249654 -2.0380295 1.4730364
```

```
PNC1 = getSymbols("PNC", from="2012-7-1", to="2012-7-31", auto.assign=F)
ffhold1$PNCexret = 100*dailyReturn(PNC1) - ffhold1$RF
# count correct prediction intervals
count=0
for (i in 1:21)
{
  if (ffhold1$PNCexret[i]>=result[i,2] && ffhold1$PNCexret[i]<=result[i,3])
    count=count+1
}
print(count/21)
```

```
## [1] 0.9047619
```

Therefore 90.4% of the intervals include the true values.