STAT 631 Homework 8

Jack Cunningham (jgavc@tamu.edu)

11/1/24

```
source("FM_Functions.R")
source("Factor_Tests.R")
load("HW08.RData")
attach(FF5)
```

1)

The Fama-French 3 factor model is the below:

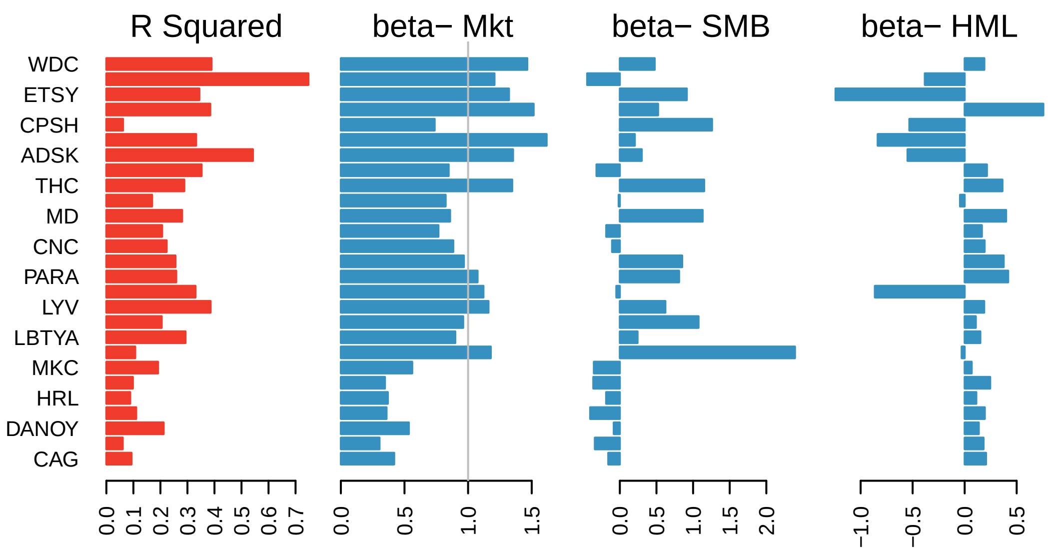
$$Y_t = \alpha + B^T F_t + \epsilon_t, \quad E[\epsilon_t|F_t] = 0, \quad E[\epsilon_t \epsilon_t^T|F_t] = \Sigma_\epsilon$$

Where $F = \begin{bmatrix} \text{Excess Return Market Portfolio} & \text{Small Minus Big} & \text{High Minus Low} \end{bmatrix}$, these are each economic vectors with length n.

```
Yt = apply(Rt,2, function(x) x-RF); dimnames(Yt)[[2]] = syb;
n = dim(Yt)[1]; N = dim(Yt)[2]; p = 3
fit = lm(Yt ~ Mkt.RF + SMB + HML); sfit = summary(fit)
```

a)

```
betas = coef(fit)[-1,]
R.Squared = c(); for(i in 1:N) R.Squared[i] = sfit[[i]]$r.squared
names(R.Squared) <- syb
coef.plot(R.Squared, coef(fit)[-1,])
```



From the R squared plot we see that the three factor Fama French model performance varies greatly. Let's take a look at the breakdown by industry:

```
table(Hi_R.Sq = R.Squared > 0.5, by_industry)
```

```
       by_industry
Hi_R.Sq Ent Food HCare Tech
  FALSE   7    7     6    5
  TRUE    0    0     0    2
```

Generally R-Squared isn't very high for these assets. There are only two that exceed 0.5, both are in the technology industry Microsoft and Autodesk.

```
table(Hi_R.Sq = R.Squared < 0.2, by_industry)
```

```
       by_industry
Hi_R.Sq Ent Food HCare Tech
  FALSE   6    1     5    6
  TRUE    1    6     1    1
```

R-Squared is particularly low for the Food industry, six of the seven stocks have an R-Squared beneath 0.2. This indicates that the three factor Fama French model does not perform well for this industry.

```r
table(Aggressive = coef(fit)[2,] > 1, by_industry)
```

```
          by_industry
Aggressive Ent Food HCare Tech
     FALSE   3    7     5    1
     TRUE    4    0     1    6
```

On an industry level we see that that Food and Heath Care are not aggressive compared to market returns while Technology generally is. Entertainment is more of a mixed bag.

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
compare <- data.frame(
  Stock = syb,
  Beta = coef(fit)[2,],
  Industry = by_industry
)
compare |>
  group_by(Industry) |>
  summarise(Average_Beta = mean(Beta))
```

```
# A tibble: 4 x 2
  Industry Average_Beta
  <chr>           <dbl>
```

```
1 Ent              1.05
2 Food             0.413
3 HCare            0.921
4 Tech             1.32
```

By taking a look at the average Beta we can see that the Food Industry has a Beta of 0.4132 on average. Healthcare, despite being not aggressive compared to the market, is far closer to 1 in comparison.

b)

To identify the individual assets that don't follow the FF-3-factor model we use the t-test for $H_0 : \alpha_i = 0$ that is automatically computed from the lm function.

```
Alpha = c()
for(i in 1:N){
  Alpha = rbind(Alpha, sfit[[i]]$coef[1, ])
}
dimnames(Alpha)[[1]] = syb
Alpha_df <- data.frame(Alpha, Industry = by_industry)
Alpha_df |>
  filter(Pr...t.. < .05)
```

```
         Estimate Std..Error   t.value   Pr...t.. Industry
LBTYA -0.08814986 0.03719910 -2.369677 0.01789040      Ent
PARA  -0.11743008 0.05692884 -2.062752 0.03925461      Ent
WBD   -0.10721466 0.05359548 -2.000442 0.04557659      Ent
MD    -0.12999721 0.05171332 -2.513805 0.01201531    HCare
```

There are four individual assets that do not follow the FF-3 factor model, Live Nation Entertainment, Paramount, Warner Brothers Discovery and Pediatric Medical Group. The first three are in the entertainment industry and the last is in healthcare.

c)

We are testing the hypothesis that $H_0 : \alpha = 0$. If we reject this hypothesis this indicates that the FF-3 factor does not hold for all 27 assets. We perform the Wald and Likelihood Ratio Tests.

```
alpha <- coef(fit)[1, ]
res = resid(fit); Sig.e = 1/n*t(res)%*%res
m11 = sfit[[1]]$cov.unscaled[1,1]
var.alpha = m11*Sig.e
```

```r
p = 3

wald.fun(est = alpha, est.var = var.alpha, n =  n, p = p)
```

```
      Wald      p.value          df1          df2
 1.1490913    0.2718611   27.0000000 2150.0000000
```

```r
res.0 = resid(lm(Yt~Mkt.RF + SMB + HML - 1))
Sig.e0 = 1/n*t(res.0)%*%res.0
lrt.fun(sig = Sig.e, sig0 = Sig.e0,n = n)
```

```
       LRT    p.value         df
31.0114868  0.2706672 27.0000000
```

Both the Wald and Likelihood test ratios have a similar result with p value $\approx .271$. We cannot reject the null hypothesis that the FF-3 factor holds for all 27 assets.

d)

```r
wald = c(); lrt = c()
for(i in industry){
  ind = which(by_industry == i)
  wald = rbind(wald, wald.fun(alpha[ind], m11*Sig.e[ind,ind],n = n, p = p))
  lrt = rbind(lrt, lrt.fun(Sig.e[ind,ind], Sig.e0[ind,ind], n = n))
}

rownames(wald) = rownames(lrt) = industry
cat("Wald test by industry:"); wald
```

```
Wald test by industry:

           Wald    p.value df1  df2
Food  0.2752167 0.96372392   7 2170
Ent   1.7837085 0.08628134   7 2170
HCare 1.8053977 0.09425210   6 2171
Tech  1.3190840 0.23693005   7 2170
```

```r
cat("LRT by industry:"); lrt
```

```
LRT by industry:
```

```
            LRT    p.value df
Food   1.929655 0.96363090  7
Ent   12.475994 0.08595263  7
HCare 10.825360 0.09392620  6
Tech   9.233106 0.23635053  7
```

All industries cannot reject the null hypothesis that the FF-3 factor model holds for their respective stocks at a significance level of 0.05. However there is still a significant difference between the industries. At a significance level of 0.1 both entertainment and healthcare would reject the null hypothesis. The Food industry however has a p.value ≈ 0.96, the evidence strongly suggests the FF-3 factor model holds well for this industry.

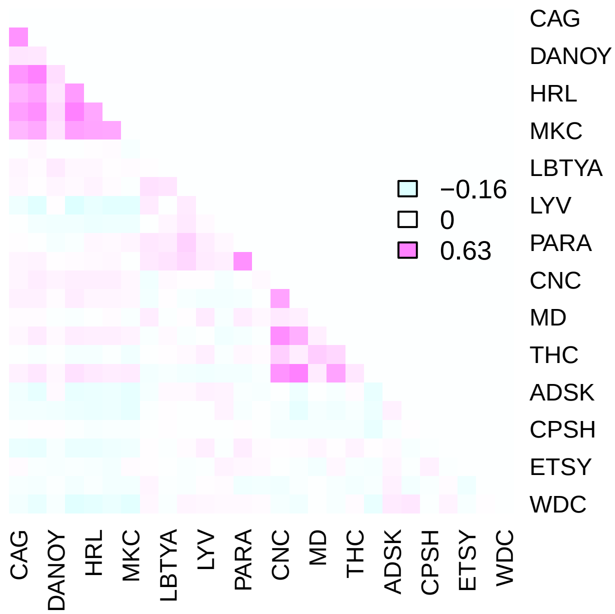
e)

The sample covariance approach has $N(N+1)/2$ estimates. With $N = 27$ this is 378 estimates.

The model based approach has $(p+1)(N+p/2)$ estimates. With $N = 27, p = 3$ this is 114 estimates.

```r
resid.summary(res)
```

```
Significant pairs at 1% level:  133 of  351 pairs
Significant pairs at 5% level:  180 of  351 pairs
```



We can see high correlation between stocks in the same industry but small correlation between stocks in different industries. This indicates that our assumption of a diagonal covariance matrix could be unreasonable. This would call into question any inference we obtain from our model. We should formally check for this and consider an industry factor model as a possible option.

f)

The test for block-diagonal matrices tests $H_0 : \Sigma = \text{diag}\{\Sigma_{11}, \dots, \Sigma_{kk}\}$ has test statistic:

$$\text{LRT} = -\log \frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_{11}) \dots \det(\hat{\Sigma}_{kk})}$$

This statistic is approximately χ^2_v, with degrees of freedom $v = \frac{1}{2}(d^2 - \sum_{i=1}^k d_i^2)$.

```
cov.diag.test(Sig.e, Ns = Ns, n = n, p = p)
```

```
*** Testing if the matrix is block diagonal ***
LRT -statistic: 846.2345     p-value: 0     DF: 273
```

This test rejects the null hypothesis of block-diagonal matrices.

We also test whether the full matrix is diagonal. This is an adaption of the previous test, we have $d_i = 1, i = 1, ..., d$. Then the statistic is $-\log\det(\widehat{\text{Corr}}(\text{y}))$, with degrees of freedom $v = \frac{1}{2}d(d-1)$.

```
cov.diag.test(Sig.e, Ns = rep(1,N), n = n, p = p)
```

```
*** Testing if the matrix is diagonal ***
LRT -statistic: 10095.56     p-value: 0     DF: 351
```

This test rejects the null hypothesis of a diagonal covariance matrix.

2)

```
fa.none = factanal(Yt,3,rotation = "none")
print(fa.none)
```

```

Call:
factanal(x = Yt, factors = 3, rotation = "none")

Uniquenesses:
  CAG   CPB DANOY   GIS   HRL     K   MKC   AMC LBTYA LGF-A   LYV  NFLX  PARA
0.532 0.409 0.759 0.285 0.572 0.409 0.544 0.887 0.683 0.743 0.567 0.755 0.678
  WBD   CNC   HUM    MD   MOH   THC   UNH  ADSK   AMD  CPSH   DXC  ETSY  MSFT
0.693 0.403 0.389 0.775 0.516 0.705 0.238 0.499 0.728 0.950 0.665 0.758 0.506
  WDC
0.595

Loadings:
      Factor1 Factor2 Factor3
CAG    0.485  -0.458   0.153
CPB    0.453  -0.618
DANOY  0.472           0.134
GIS    0.531  -0.655
HRL    0.472  -0.443
K      0.485  -0.589
MKC    0.544  -0.382   0.117
AMC    0.201   0.142   0.229
LBTYA  0.483   0.178   0.228
LGF-A  0.363   0.195   0.295
LYV    0.465   0.381   0.268
```

```
NFLX   0.369   0.237   0.229
PARA   0.414   0.215   0.322
WBD    0.423   0.199   0.298
CNC    0.657   0.180  -0.364
HUM    0.629   0.148  -0.440
MD     0.400   0.230   0.109
MOH    0.586   0.154  -0.342
THC    0.464   0.277
UNH    0.753   0.132  -0.422
ADSK   0.545   0.346   0.289
AMD    0.384   0.265   0.233
CPSH   0.119   0.104   0.159
DXC    0.471   0.291   0.170
ETSY   0.377   0.235   0.211
MSFT   0.631   0.239   0.196
WDC    0.470   0.333   0.270

               Factor1 Factor2 Factor3
SS loadings      6.362   2.822   1.575
Proportion Var   0.236   0.105   0.058
Cumulative Var   0.236   0.340   0.398

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 2924.13 on 273 degrees of freedom.
The p-value is 0
```

The first factor has all positive coefficients and are relatively similar, it seems to be a shared market component.

The second factor has negative, with the exception of a near zero coefficient for Danone SA, coefficient for all stocks in the food industry. This appears to be an industry factor.

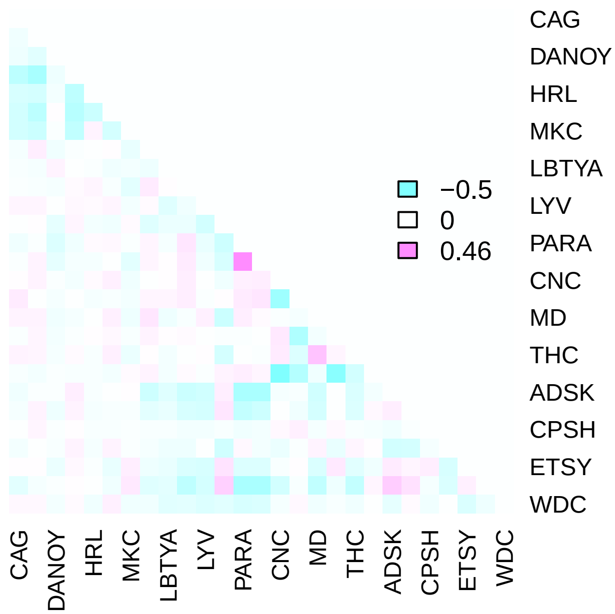
The third factor looks like an aggressiveness component. It is positive for all tech and entertainment stocks with generally negative values for health and food stocks. This comports with our previous analysis of betas for each stock.

b)

```
p = 3
Zt = apply(Yt, 2, function(u) (u-mean(u))/sd(u))
fa = factanal(Zt, p, scores = "Bartlett", rotation = "none")
B = t(fa$loading)
Ft.fa = fa$scores
```

```r
R.Sq.fa = diag(t(B)%*%var(Ft.fa)%*%B)
resid_mat = Zt - Ft.fa %*% B
resid.summary(resid_mat)
```

```
Significant pairs at 1% level:  138 of  351 pairs
Significant pairs at 5% level:  187 of  351 pairs
```

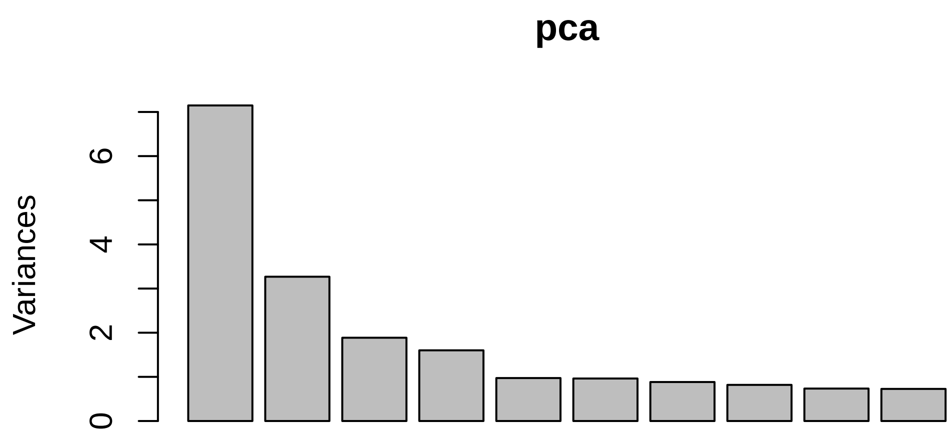


There is still correlation but it is less grouped by industry. It is less severe between particular stocks, before in the FFA-3-F model correlation was particularly strong for certain pairs. This model makes the assumption of a diagonal covariance matrix a bit more reasonable.

3)

Using the standardized excess return data means we are creating an approximate factor through PCA.

```r
pca = prcomp(Zt)
plot(pca)
```



From this plot I would choose three principal components. The difference of explained variance between three and four is rather small.

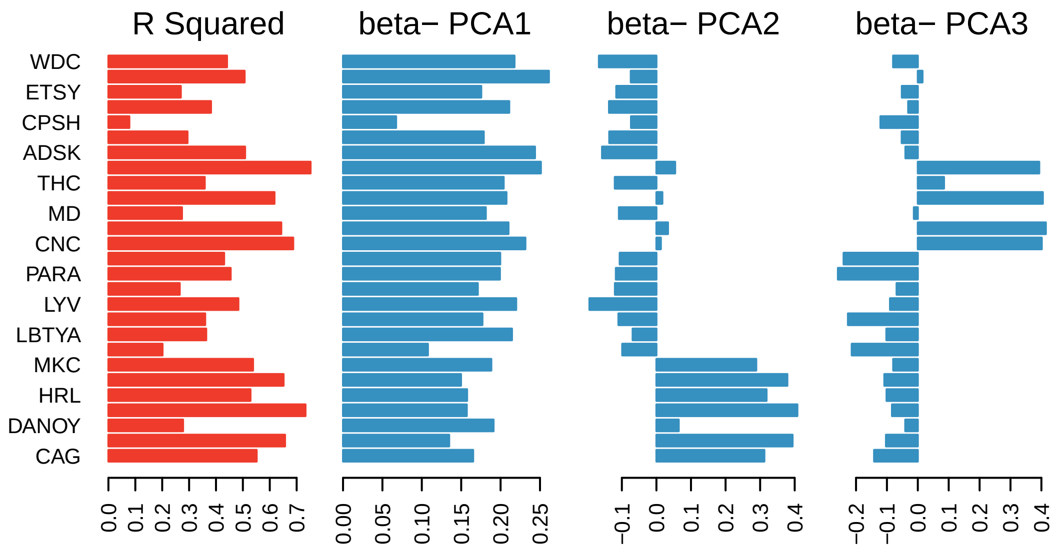
b)

We choose $p = 3$.

```
p = 3
B = t(pca$rotation[, 1:p])
Ft.pc = pca$x[, 1:p]
R.Sq.pc = diag(t(B)%*%diag(pca$sd[1:p]^2)%*%B)
```

c)

```
coef.plot(R.Sq.pc, B, factors = c("PCA1","PCA2","PCA3"))
```




```
compare_pca <- data.frame(
  symbol = syb,
  beta_1 = as.numeric(B[1,]),
  beta_2 = as.numeric(B[2,]),
  beta_3 = as.numeric(B[3,]),
  industry = by_industry
)
```

The weight on the first principal component is similar for most stocks across industries, it appears to reflect the general upward trend of each stock over time. The weights on the second principal component are positive for all stocks in the food industry, this approximates an industrial factor.

```
compare_pca |>
  filter(beta_2 > 0)
```

```
  symbol    beta_1     beta_2      beta_3 industry
1    CAG 0.1651490 0.31219352 -0.14148823     Food
2    CPB 0.1347282 0.39383187 -0.10311558     Food
3  DANOY 0.1910415 0.06452018 -0.04067080     Food
4    GIS 0.1570461 0.40772128 -0.08359895     Food
5    HRL 0.1574734 0.31857746 -0.10118199     Food
```

```
6       K 0.1497350 0.37874871 -0.10835499     Food
7     MKC 0.1882090 0.28877044 -0.07971293     Food
8     CNC 0.2314410 0.01280475  0.40129452    HCare
9     HUM 0.2099790 0.03344303  0.41443813    HCare
10    MOH 0.2074282 0.01713727  0.40473127    HCare
11    UNH 0.2511059 0.05413348  0.39301608    HCare
```

In fact there are also a few healthcare companies with small positive coefficients, they all are less aggressive than the market as determined in the FFA-3 factor model.

```
compare |>
  filter(Stock %in% c("CNC","HUM","MOH"))
```

```
    Stock      Beta Industry
CNC   CNC 0.8834189    HCare
HUM   HUM 0.7669530    HCare
MOH   MOH 0.8235326    HCare
```

In fact if we compute the correlation between the two coefficients we see they are strongly negatively correlated. This indicates that the B_2 estimates seem to be a combination of an industry and conservative factor.

```
cor(compare_pca$beta_2, compare$Beta)
```

```
[1] -0.8759118
```

The weights B_3 appear to firmly be an industry factor for healthcare companies, there are six stocks with a positive coefficient five of which are healthcare companies and Microsoft (with a very small positive coefficient). Perhaps Microsoft is in this group because healthcare companies are large institutions that rely on both Windows software and database solutions.

```
compare_pca |>
  filter(beta_3 > 0)
```

```
  symbol    beta_1      beta_2     beta_3 industry
1    CNC 0.2314410  0.01280475 0.40129452    HCare
2    HUM 0.2099790  0.03344303 0.41443813    HCare
3    MOH 0.2074282  0.01713727 0.40473127    HCare
4    THC 0.2039138 -0.12042617 0.08478033    HCare
5    UNH 0.2511059  0.05413348 0.39301608    HCare
6   MSFT 0.2612992 -0.07413066 0.01511342     Tech
```

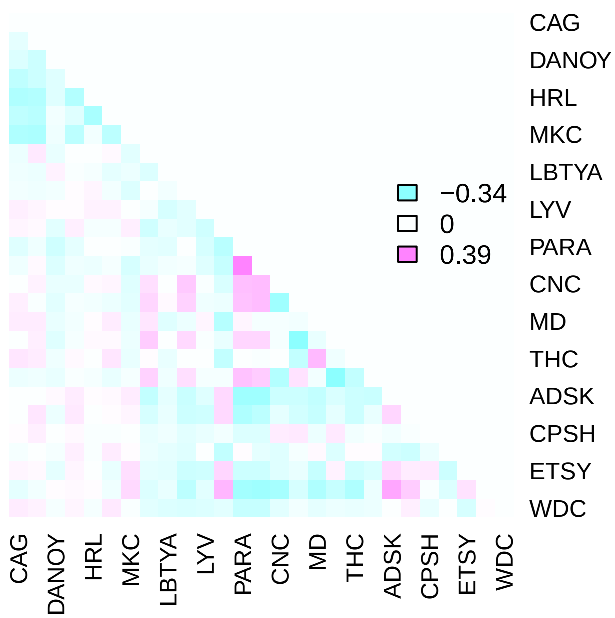
d)

With our $p = 3$ we have the following residual matrix:

$$\hat{E} = Z - \tilde{F}\hat{B}$$

```
lambda_diag <- diag(pca$sd[1:p]^2)
O_matrix <- t(B)
Ft = pca$x[,1:p]
resid.pca = Zt - Ft %*% B
resid.summary(resid.pca)
```

```
Significant pairs at 1% level:  185 of  351 pairs
Significant pairs at 5% level:  220 of  351 pairs
```

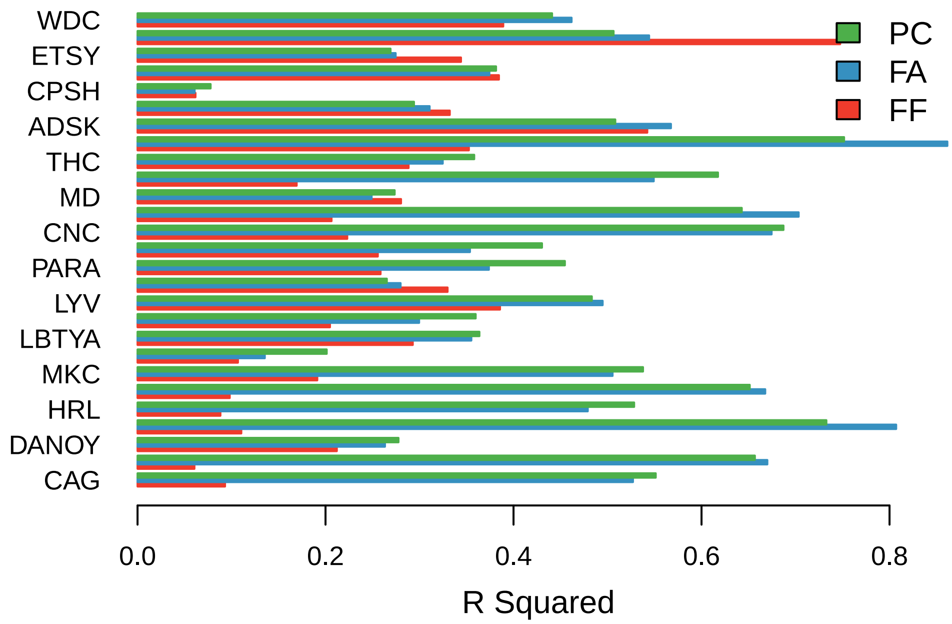


The PCA model with $p = 3$ has a covariance matrix with more significant pairs than both of the previous models, correlation is less grouped within industry than the FF3 Factor model though. This indicates the assumption of a diagonal covariance matrix may not be appropriate.

4)

a)

```r
RSq.all <- cbind(R.Squared, R.Sq.fa, R.Sq.pc)
RSq.plot(RSq.all)
```



The FF3 factor model is the worst out of the three we've tested. The biggest discrepancies can be seen in certain industries. Visually we can see how low R^2 was in the food industry in the bottom 7 stocks on the graph and how much better the two other models, which are able to factor in industry differences, perform. The PCA and FA models perform similarly for the return data.

The overall takeaway is that when we are dealing with companies that belong to multiple known industries we should extend past the FF3 factor model and opt for ones that can take into account industry factors. The main concern about the PCA and FA models is their lack of interpretability but with comparisons to a default model, like the FF3 factor model, we can get an idea of what each generated factor represents.