Final Exam

Galen Byrd

5/6/2018

1

A

```r
library(readxl)
firms <- read_excel("firms.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/New_York'
```

```r
bankrupt <- subset(firms, Status=="B")[,-5]
sound <- subset(firms, Status=="S")[,-5]
#Calculating the basic statistics
p <- ncol(bankrupt)
n1 <- nrow(bankrupt)
n2 <- nrow(sound)
#Calculating the mean vectors and covariance matrices
mean.b <- colMeans(bankrupt)
mean.s <- colMeans(sound)
S.b <- var(bankrupt)
S.s <- var(sound)
S.pl <- ((n1-1)*S.b+(n2-1)*S.s)/(n1+n2-2)
#Calculating Hotelling's T2
T2 <- n1*n2/(n1+n2)*t(mean.b-mean.s)%*%solve(S.pl)%*%(mean.b-mean.s)
#Calculating the critical value
a <- p*(n1+n2-2)/(n1+n2-p-1)
crit.val <- a*qf(.95,p,n1+n2-p-1)
(p.val <- 1-pf(1/a*T2,p,n1+n2-p-1))
```

```
##              [,1]
## [1,] 1.359661e-05
```

```r
# Reject null. There is evidence (p-val=1.36e-05) to suggest a difference among bankrupt
# and financially sound banks in at least one of the variables x1,x2,x3,x4.
```

B

```r
# We perform a hypothesis test to make sure the groups are different enough
# in at least one of the given variables.
```

C

```r
#Finding the E and H matrix using MANOVA
m1 <- manova(cbind(x1,x2,x3,x4)~as.factor(Status),data=firms)
```

```r
H <- summary(m1)$SS[[1]]
E <- summary(m1)$SS[[2]]
#Calculating the eigenvalues and vectors for the discriminant analysis
e.vals <- Re(round(eigen(solve(E)%*%H)$values,digits=4))
e.vecs <- Re(round(eigen(solve(E)%*%H)$vectors,digits=4))
(a1 <- e.vecs[,1])
```

```
## [1] -0.1323 -0.9412 -0.1867  0.2484
```

```r
# So the variables that contribute to the separation from most to least are: x2,x4,x3,x1
```

D

```r
t(a1)%*%mean.b
```

```
##               [,1]
## [1,] -0.06032907
```

```r
t(a1)%*%mean.s
```

```
##              [,1]
## [1,] -0.4613045
```

```r
(zc <- .5*t(a1)%*%(mean.b+mean.s))
```

```
##              [,1]
## [1,] -0.2608168
```

```r
# for new data point, if its mean is >-.26 it's Bankrupt, <-.26 it's Financially stable
```

E

```r
library(MASS)
k <- 2
LDA <- lda(Status~x1+x2+x3+x4 , data=firms, prior=rep(1,k)/k)
(error <- mean(firms$Status != predict(LDA)$class) )
```

```
## [1] 0.08695652
```

```r
#Apparent error rate = .087
LDA.CV <- lda(Status~x1+x2+x3+x4 , data=firms, prior=rep(1,k)/k, CV=T)
(error <- mean(firms$Status != LDA.CV$class) )
```

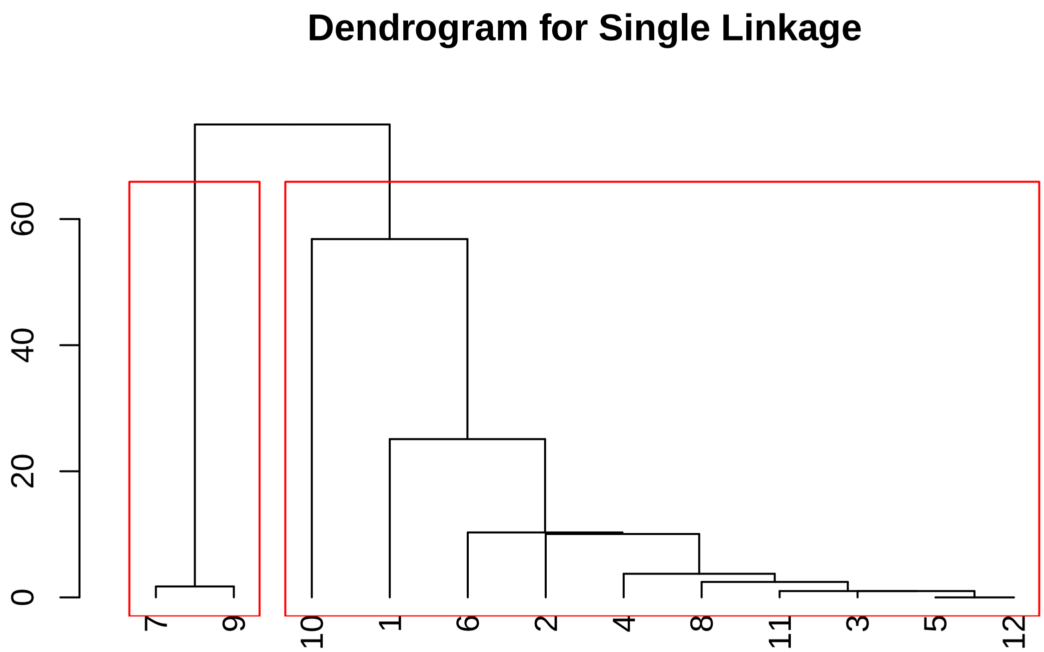
```
## [1] 0.1086957
```

```r
#Error rate using cross validation = .109

# Apparent error rate underestimates actual error rate. Using cross validation we remove
# each observation individially and recalculate the classification rules, which should
# pull the apparent error rate towards the actual error rate, which it does.
```

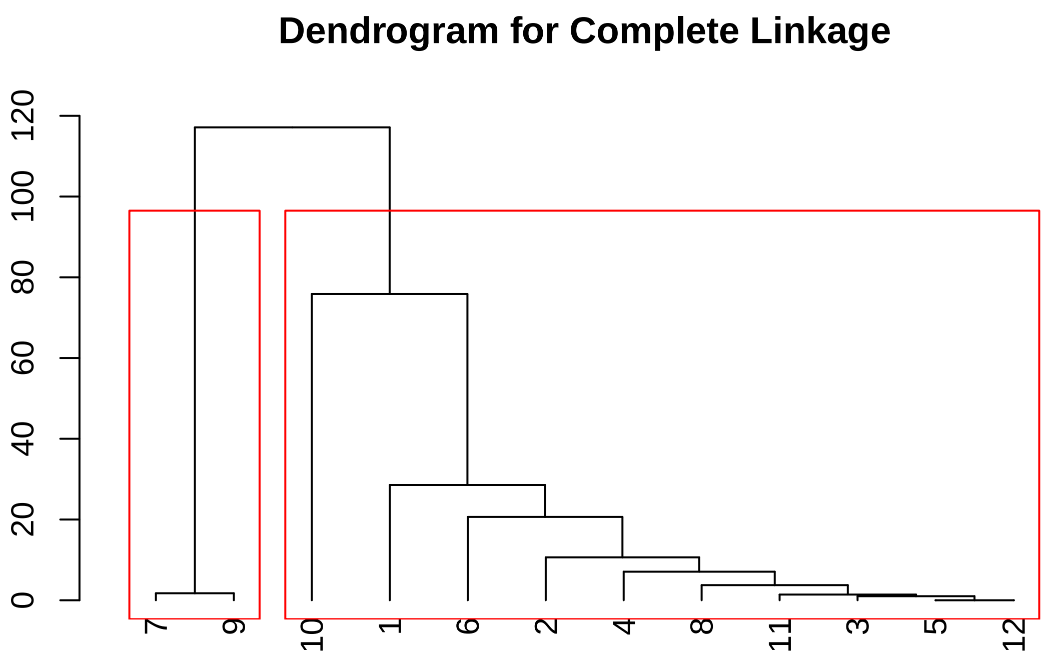
2

A

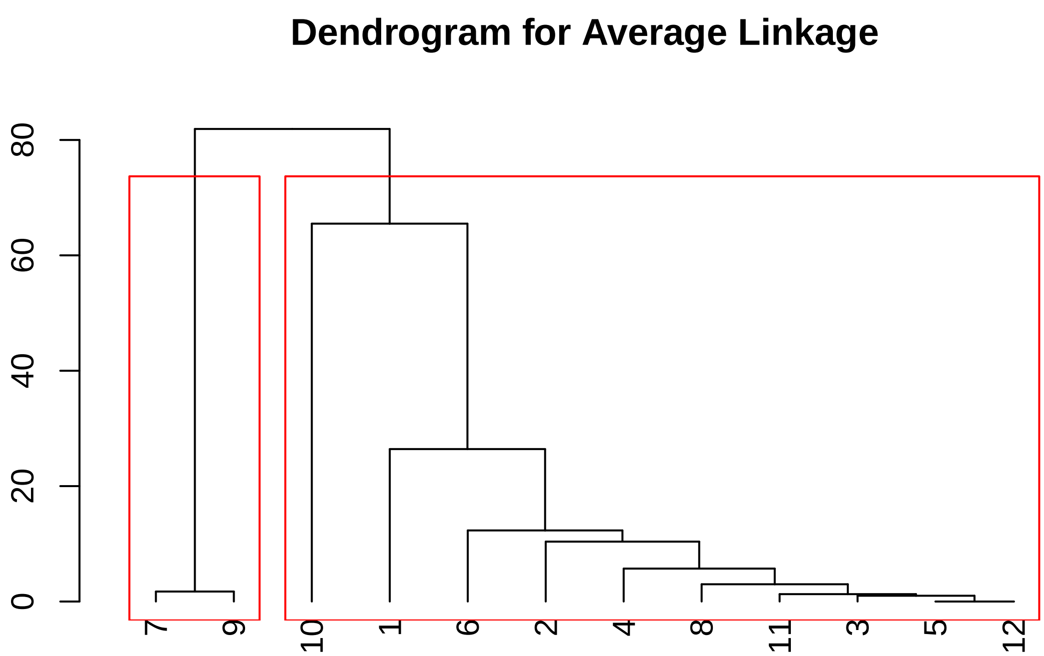
```r
cereal <- read_excel("cereal.xlsx")
D <- dist(cereal[,-1],diag=T, upper=T)

m.sl <-hclust(d = D, method="single")
plot(as.dendrogram(m.sl), main="Dendrogram for Single Linkage")
rect.hclust(m.sl,k=2,border="red")
```




```r
m.cl <-hclust(d = D, method="complete")
plot(as.dendrogram(m.cl), main="Dendrogram for Complete Linkage")
rect.hclust(m.cl,k=2,border="red")
```




```r
m.al <-hclust(d = D, method="average")
plot(as.dendrogram(m.al), main="Dendrogram for Average Linkage")
rect.hclust(m.al,k=2,border="red")
```



B

```r
# I prefer the average method because it does not stretch/shrink the data
# like single and complete linkage do
```

C

```
# I would use 2 clusters
```

D

```
# Based on that answer, Product/Total are in a group, and the rest are in one big group.
# This makes sense as Product/Total both have a lot more vitamins than the others.
```

E

```
c1<-rbind(cereal[7,-1],cereal[9,-1])
c2<-rbind(cereal[1:6,-1],cereal[8,-1],cereal[10:12,-1])
#Calculating the basic statistics
p <- ncol(c1)
n1 <- nrow(c1)
n2 <- nrow(c2)
#Calculating the mean vectors and covariance matrices
mean1 <- colMeans(c1)
mean2 <- colMeans(c2)
S.pl <- var(cereal[,-1])
S1 <- var(c1)
S2 <- var(c2)
S.pl <- ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
#Calculating Hotelling's T2
T2 <- n1*n2/(n1+n2)*t(mean1-mean2)%*%solve(S.pl)%*%(mean1-mean2)
#Calculating the critical value
a <- p*(n1+n2-2)/(n1+n2-p-1)
crit.val <- a*qf(.95,p,n1+n2-p-1)
(p.val <- 1-pf(1/a*T2,p,n1+n2-p-1))
```

```
##              [,1]
## [1,] 5.838216e-05
```

```
# Reject null. There is evidence (p-val=5.84e-05) to suggest a difference in means among
# cereal groups, meaning we should not combine the clusters.
```

3

A

```
house.med <- read.table("housdat.txt",header=T)
house <- house.med[,-14]
n <- nrow(house)
p <- ncol(house)
diag(cov(house))
```

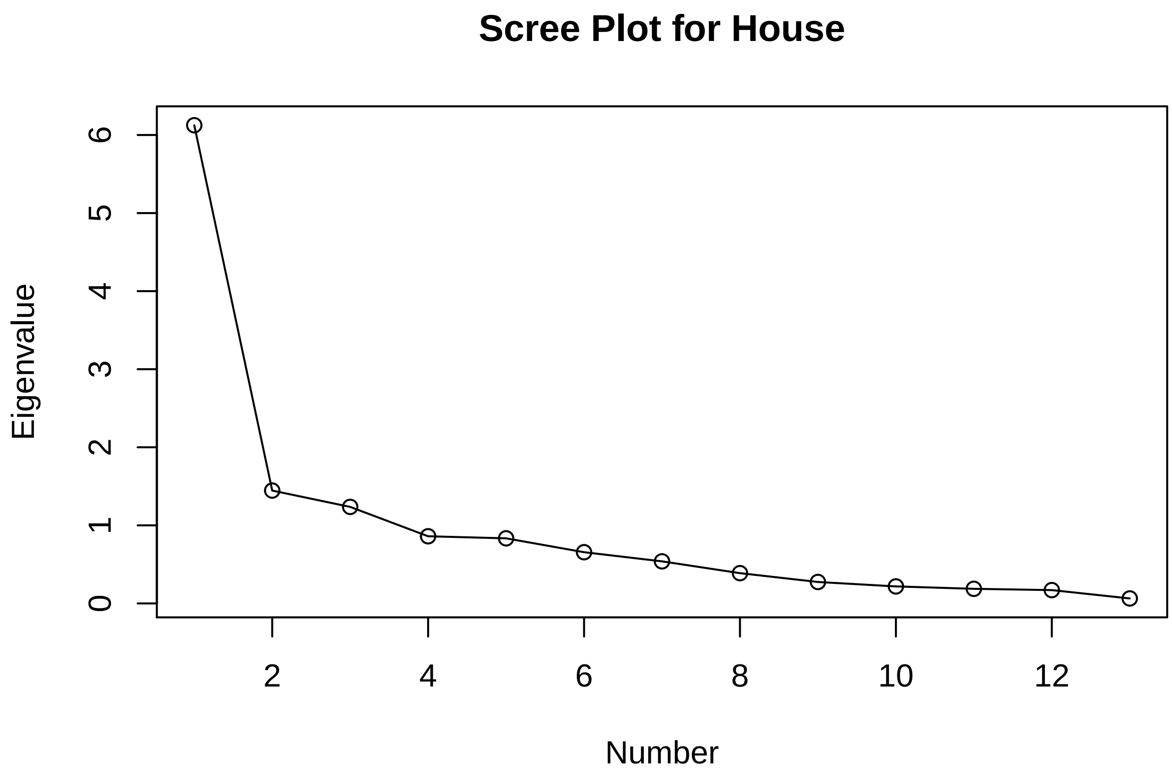
```
##         CRIM        PLAND         PBUS          OCE          NOC
## 7.562028e+01 5.366461e+02 4.755649e+01 6.608783e-02 1.351867e-02
```

```
##          ARM         PAGE         WDIS        INDEX         FTAX
## 4.986305e-01 7.703394e+02 4.492996e+00 7.642315e+01 2.873585e+04
##          PTR           BK          LSP
## 4.654971e+00 8.518767e+03 5.123659e+01
```

```
# Use correlation matrix because the variances in the data are not similar.
# Some have small variances and others have huge variances, which would effect the PC's.
# The variance for FTAX is the one that is too large.
```

B

```
R <- cor(house)
e.vec <- eigen(R)$vectors
e.val <- eigen(R)$values

plot(1:p,e.val, xlab="Number",ylab="Eigenvalue",main="Scree Plot for House", type="l")
points(1:p,e.val)
```




```
percentage <- rep(0,p)
for (i in 1:p){
  percentage[i] <- sum(e.val[1:i])/sum(e.val)
}
(percentage)
```

```
##  [1] 0.4711437 0.5823602 0.6774900 0.7436522 0.8078070 0.8582815 0.8997431
##  [8] 0.9295898 0.9507359 0.9675108 0.9819264 0.9950744 1.0000000
```

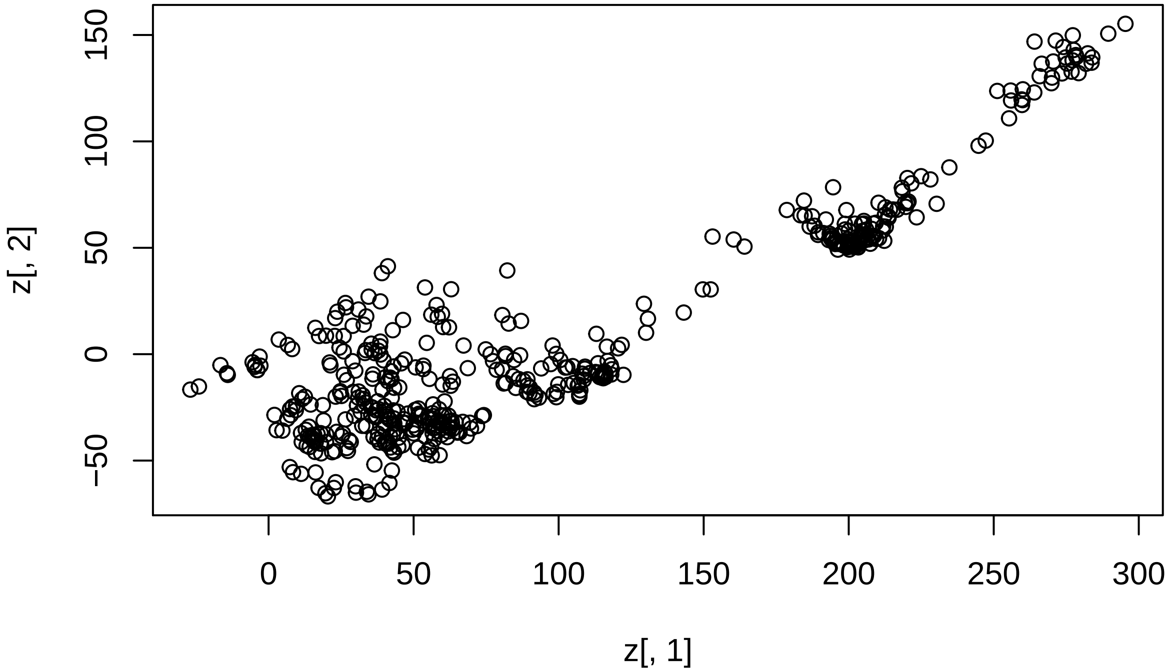
```
# I would use the first 5 PC's so that we retain 80% of the variability in the data.
# This is also where the scree plot really levels off.
```

C

```
# 80%
```

D

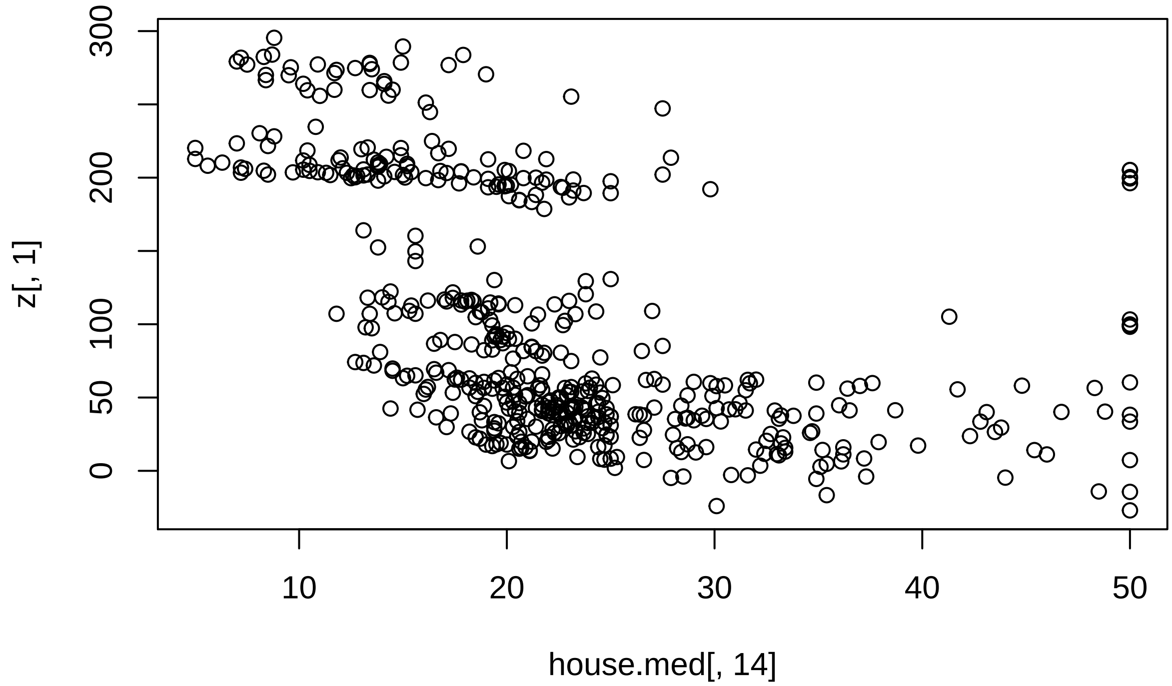
```
z<-as.matrix(house)%*%e.vec
plot(z[,1],z[,2])
```



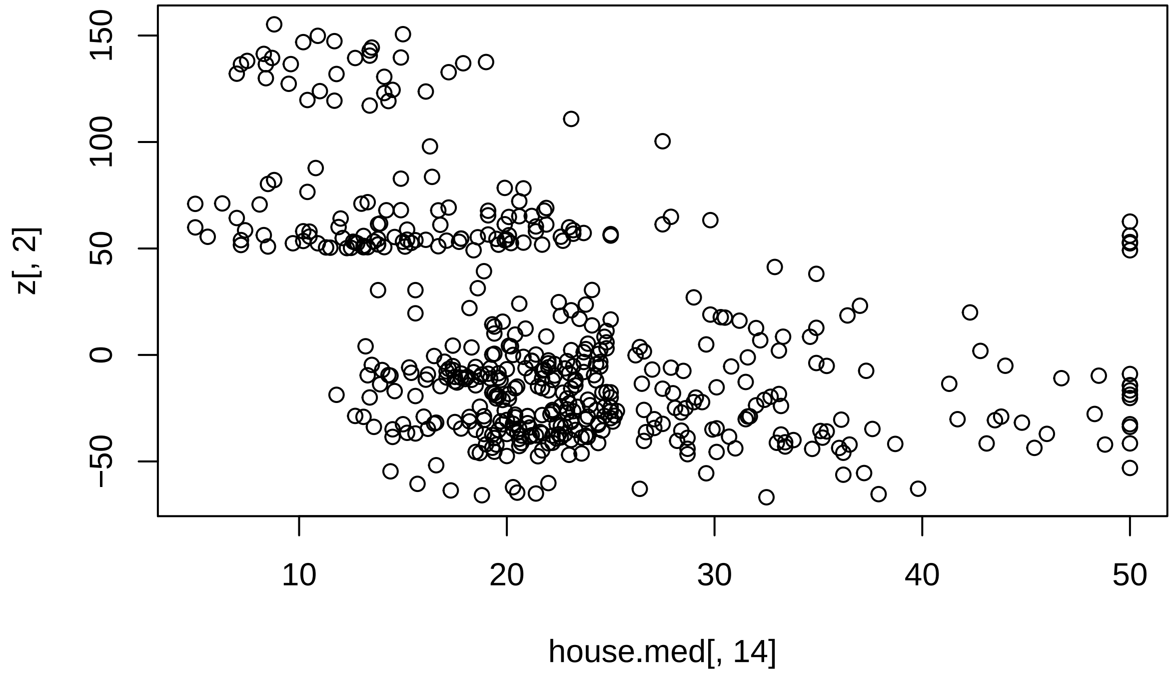

```
# Yes there are a few groups
```

E

```
plot(house.med[,14],z[,1])
```




```
plot(house.med[,14],z[,2])
```




```
# It does not appear these two PC's are very useful in predicting median home prices.
# There are big groups horizontally, which means that for one value of our PC, there
# is a big range of potential house prices.
```

4

A

```r
# y = mu + L %*% f + error
# Where L is a p by m matrix of loadings and f is the vector of common factors
# Assume: E(f)=0, E(error)=0, var(f)=I, var(error)=Psi, and f and error are independent.
```

B

```r
library(psych)
R <- cor(house)

FA.ML4 <- factanal(covmat=R, factors=4, rotation="varimax")
Psi.ml4 <- diag(diag(R-FA.ML4$loadings%*%t(FA.ML4$loadings)))
FA.ML5 <- factanal(covmat=R, factors=5, rotation="varimax")
Psi.ml5 <- diag(diag(R-FA.ML5$loadings%*%t(FA.ML5$loadings)))
FA.ML6 <- factanal(covmat=R, factors=6, rotation="varimax")
Psi.ml6 <- diag(diag(R-FA.ML6$loadings%*%t(FA.ML6$loadings)))
FA.ML7 <- factanal(covmat=R, factors=7, rotation="varimax")
Psi.ml7 <- diag(diag(R-FA.ML7$loadings%*%t(FA.ML7$loadings)))

#Loglikelihood
m=4
FA4 <- (FA.ML4$loadings%*%t(FA.ML4$loadings)+Psi.ml4)
ll4 <- -n/2*(log(det(FA4))+sum(diag(solve(FA4)%*%R)))
(AIC4 <- -2*ll4+2*(p*(m+1)-m*(m-1)))
```

```
## [1] 2318.423
```

```r
m=5
FA5 <- (FA.ML5$loadings%*%t(FA.ML5$loadings)+Psi.ml5)
ll5 <- -n/2*(log(det(FA5))+sum(diag(solve(FA5)%*%R)))
(AIC5 <- -2*ll5+2*(p*(m+1)-m*(m-1)))
```

```
## [1] 2191.268
```

```r
m=6
FA6 <- (FA.ML6$loadings%*%t(FA.ML6$loadings)+Psi.ml6)
ll6 <- -n/2*(log(det(FA6))+sum(diag(solve(FA6)%*%R)))
(AIC6 <- -2*ll6+2*(p*(m+1)-m*(m-1)))
```

```
## [1] 2166.213
```

```r
m=7
FA7 <- (FA.ML7$loadings%*%t(FA.ML7$loadings)+Psi.ml7)
ll7 <- -n/2*(log(det(FA7))+sum(diag(solve(FA7)%*%R)))
(AIC7 <- -2*ll7+2*(p*(m+1)-m*(m-1)))
```

```
## [1] 2147.765
```

```r
# We should use 7 factors as this minimizes the value of AIC

e.val <- eigen(R)$values
```

```r
percentage <- rep(0,p)
for (i in 1:p){
  percentage[i] <- sum(e.val[1:i])/sum(e.val)
}
(percentage)
```

```
##  [1] 0.4711437 0.5823602 0.6774900 0.7436522 0.8078070 0.8582815 0.8997431
##  [8] 0.9295898 0.9507359 0.9675108 0.9819264 0.9950744 1.0000000
```

```r
# We retain 90% of the variability in the data
```

C

```r
FA.ML7$loadings
```

```
##
## Loadings:
##       Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## CRIM   0.659   0.121   0.128   0.131                  -0.143
## PLAND -0.107  -0.811  -0.193  -0.158  -0.206
## PBUS   0.464   0.402   0.299   0.243   0.159   0.531   0.114
## OCE                                                    0.267
## NOC    0.550   0.411   0.269   0.307           0.301   0.336
## ARM           -0.134  -0.817          -0.138           0.142
## PAGE   0.309   0.429   0.170   0.794           0.115   0.190
## WDIS  -0.399  -0.652  -0.110  -0.329          -0.230  -0.238
## INDEX  0.890   0.123                   0.403           0.109
## FTAX   0.799   0.106   0.166   0.146   0.361   0.296
## PTR    0.246   0.251   0.203           0.599          -0.340
## BK    -0.512  -0.125                                   0.128
## LSP    0.430   0.246   0.616   0.350                  -0.146
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## SS loadings      3.166   1.799   1.389   1.096   0.758   0.553   0.506
## Proportion Var   0.244   0.138   0.107   0.084   0.058   0.043   0.039
## Cumulative Var   0.244   0.382   0.489   0.573   0.631   0.674   0.713
```

```r
# considering loading >.6 significant.
# Factor 1: CRIM,INDEX,FTAX. Adding crime rates with closeness to highway and property tax rate
# Factor 2: PLAND,WDIS. Adding closeness to jobs and average size of lots.
# Factor 3: ARM,LSP. Contrasting average rooms with % lower status.
# Factor 4: PAGE
# Factor 5: PTR
# Factor 6: Trivial factor
# Factor 7: Trivial Factor
```