Math 588 Quiz 7 Fall 2012 Name_____

```
out<-prcomp(state[,1:4],scale=T)
summary(out)
Importance of components: PC1    PC2    PC3     PC4
Standard deviation     1.6040 0.8803 0.6879 0.42318
Proportion of Variance 0.6432 0.1937 0.1183 0.04477
Cumulative Proportion  0.6432 0.8369 0.9552 1.00000

> out<-rprcomp(state[,1:4])
summary(out$out)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.6705 0.8216 0.59362 0.42645
Proportion of Variance 0.6977 0.1688 0.08809 0.04546
Cumulative Proportion  0.6977 0.8664 0.95454 1.00000

Rotation:PC1          PC2         PC3         PC4
gdp     0.4525021  0.688328888 -0.5429877 -0.1631243
povrt  -0.5563898 -0.016929402 -0.2468286 -0.7932335
unins  -0.4442238  0.725197372  0.5076082  0.1381588
lifexp  0.5369706  0.002347129  0.6217506 -0.5701607

out <- lda(state[,1:4],state[,5])
1-mean(predict(out,state[,1:4])$class==state[,5])
[1] 0.3
```

1) The PCA and LDA output above is for the Minor (2012) state data where gdp = GDP per capita, povrt = poverty rate, unins = 3 year average uninsured rate 2007-9, and lifexp = life expectancy for the 50 states.

a) How many principal components are needed? Use a 0.9 threshold.

3

b) Which principal component corresponds to 9 gdp −9 unins −11 povrt +11 lifeexp?

PC1

c) The fifth variable was a 1 if the state was not worker friendly and a 2 if the state was worker friendly. With these two groups, what was the apparent error rate (AER) for LDA?

0.3

60

```
> out <- lda(x,group)
> 1-mean(predict(out,x)$class==group)
[1] 0.02
>
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
[1] 0.04
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
>
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

2) The above output is for LDA on the famous iris data set. the variables are x_1 = sepal length, x_2 = sepal width, x_3 = petal length and x_4 = petal width. These four predictors are in the x data matrix. There are three groups corresponding to types of iris: setosa versicolor virginica.

a) What is the AER using all 4 predictors?

0.02

b) Which variables, if any, can be deleted without increasing the AER in a)?

X1

40