MATH 588

HW7

Md Ismail Hossain

4/10/2022

Question 1

a

```
library(Sleuth3)
fat = ex1708
dim(fat)
```

```
## [1] 12 14
```

There are 12 data points with 14 variables. So, number of variables is greater than number of observation, as a result we can not run the regression model.

b

```
names(fat) = casefold(names(fat))

aic1 = rep(NA,13)
for (i in 1:13){
  aic1[i] = AIC(lm(fat$fat~fat[,i+1]))
}
one = which.min(aic1)
cat("AIC with m", one, " = ", aic1[one], "\n", sep="")
```

```
## AIC with m8 = 56.91094
```

```
aic2 = rep(NA,13)
for (i in (1:13)[-one]){
  aic2[i] = AIC(lm(fat$fat~fat[,one+1]+fat[,i+1]))
}
two = which.min(aic2)
cat("+m", two, " = ", aic2[two], "\n", sep="")
```

```
## +m4 = 48.16104
```

```
aic3 = rep(NA,13)
for (i in (1:13)[-c(one,two)]){
  aic3[i]=AIC(lm(fat$fat~fat[,one+1]+fat[,two+1]+fat[,i+1]))
}

three = which.min(aic3)
cat("+m", three, " = ", aic3[three], "\n", sep="")
```

```
## +m2 = 46.67968
```

```
aic4 = rep(NA,13)

for (i in (1:13)[-c(one,two,three)]){
  aic4[i]=AIC(lm(fat$fat~fat[,one+1]+fat[,two+1]+fat[,three+1]+fat[,i+1]))
}

four = which.min(aic4)
cat("+m", four, " = ", aic4[four], "\n", sep="")
```

```
## +m11 = 47.03792
```

```
summary(lm(fat~m8+m4+m2, fat))
```

```
##
## Call:
## lm(formula = fat ~ m8 + m4 + m2, data = fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5118 -0.4802  0.1923  0.6961  1.8650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.02937    1.26496   5.557 0.000537 ***
## m8           0.51319    0.09706   5.287 0.000740 ***
## m4           0.44151    0.10354   4.264 0.002746 **
## m2          -0.10190    0.06210  -1.641 0.139442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.366 on 8 degrees of freedom
## Multiple R-squared:  0.9851, Adjusted R-squared:  0.9795
## F-statistic: 176.4 on 3 and 8 DF,  p-value: 1.204e-07
```

c

```
x = prcomp(fat[,-1])
cumsum( round( 100 * x$sdev[1:4]^2 / sum(x$sdev^2), 2) )
```

```
## [1] 86.50 90.88 94.58 96.59
```

```
round(x$rotation[,1:4], 2)
```

```
##       PC1   PC2   PC3   PC4
## m1  -0.26  0.35  0.08 -0.80
## m2  -0.21  0.80 -0.22  0.33
## m3  -0.36  0.01  0.07  0.25
## m4  -0.32  0.01  0.10  0.05
## m5  -0.29  0.01 -0.41  0.04
## m6  -0.29  0.07  0.33  0.07
## m7  -0.29 -0.12  0.55  0.02
## m8  -0.32 -0.14  0.00 -0.04
## m9  -0.33 -0.08  0.18  0.10
## m10 -0.26 -0.12 -0.25 -0.06
## m11 -0.29 -0.33 -0.40 -0.25
## m12 -0.16 -0.18  0.00  0.27
## m13 -0.15 -0.18 -0.29  0.15
```

d

As 2 variable is significant at b, so considering $k = 2$ here,

```
fat$PC1 = apply(fat[,-1],1,mean)
fat$PC2 = 2.35*fat$m2+fat$m1-fat$m11
summary(lm(fat~PC1+PC2, fat))
```

```
##
## Call:
## lm(formula = fat ~ PC1 + PC2, data = fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69341 -0.73519  0.07348  0.89253  1.91939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.40588    1.51161   2.253   0.0507 .
## PC1          1.16728    0.06385  18.282    2e-08 ***
## PC2         -0.07291    0.02336  -3.121   0.0123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 9 degrees of freedom
## Multiple R-squared:  0.9804, Adjusted R-squared:  0.976
## F-statistic: 224.8 on 2 and 9 DF,  p-value: 2.077e-08
```

e

Looking at the adjusted R2, the two PCs explain about the same variance the three original variables chosen using the backward step wise procedure. So, two principal components are good enough to do the prediction.

Question 2

a

```
ins = ex1716
names(ins) = casefold(names(ins))
head(ins)
```

```
##   zip fire theft  age income race vol invol
## 1  26  6.2    29 60.4  11744 10.0 5.3   0.0
## 2  40  9.5    44 76.5   9323 22.2 3.1   0.1
## 3  13 10.5    36 73.5   9948 19.6 4.8   1.2
## 4  57  7.7    37 66.9  10656 17.3 5.7   0.5
## 5  14  8.6    53 81.4   9730 24.5 5.9   0.7
## 6  10 34.1    68 52.6   8231 54.0 4.0   0.3
```

```
set1 = c("fire","theft","age","income","race")
set2 = c("vol","invol")
CCAins = cancor(ins[,set1], ins[,set2])
names(CCAins)
```

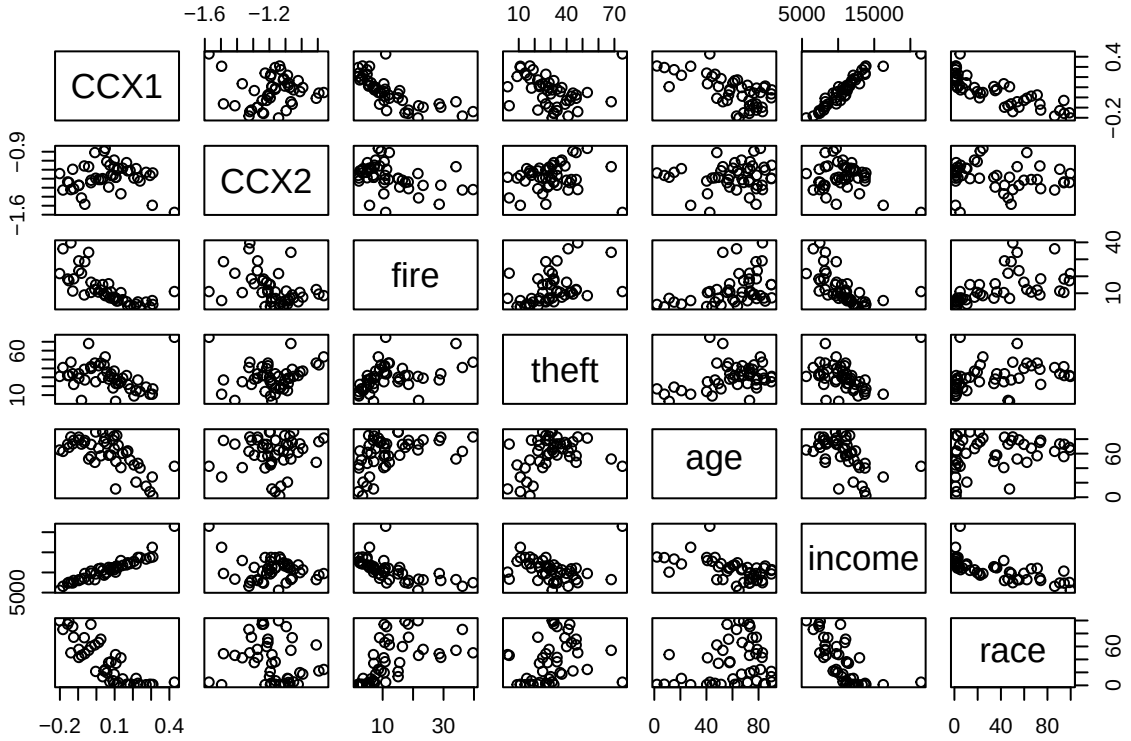
```
## [1] "cor"     "xcoef"   "ycoef"   "xcenter" "ycenter"
```

```
library(CCP)
p.asym(CCAins$cor, nrow(ins), 5, 2)
```

```
## Wilks' Lambda, using F-approximation (Rao's F):
##                stat    approx df1 df2     p.value
## 1 to 2:  0.07980847 20.318190  10  80 0.000000000
## 2 to 2:  0.65684513  5.354896   4  41 0.001455323
```

b

```
CCX1 = as.matrix(ins[,set1]) %*% as.matrix(CCAins$xcoef[,1])
CCX2 = as.matrix(ins[,set1]) %*% as.matrix(CCAins$xcoef[,2])
pairs(cbind(CCX1, CCX2, ins[,set1]))
```



Describe the canonical variables in terms of "vol" and "invol" characteristics associated with extreme ends of the scale:

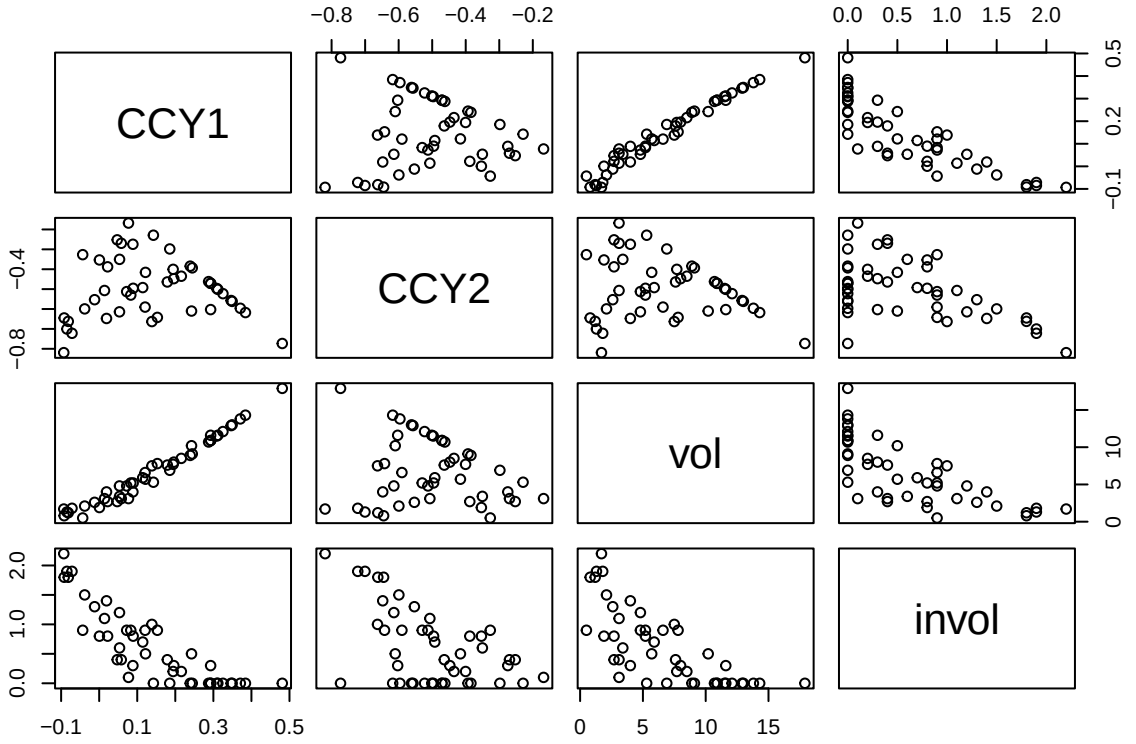
The first canonical correlation variable for set 1 contrasts zip codes with older, higher fraction of minority residents with lower income that have high rates of fire and theft with zipcodes that have younger, lower fraction of minority resident with higher incomes and low rates of fire and theft, with emphasis on income (high value is higher income).

The second CC variable seems to contrast high fire and low theft with low fire and high theft (high value is more theft than fire).

c

```
CCY1 = as.matrix(ins[,set2]) %*% as.matrix(CCAins$ycoef[,1])
CCY2 = as.matrix(ins[,set2]) %*% as.matrix(CCAins$ycoef[,2])
```

```
pairs(cbind(CCY1, CCY2, ins[,set2]))
```



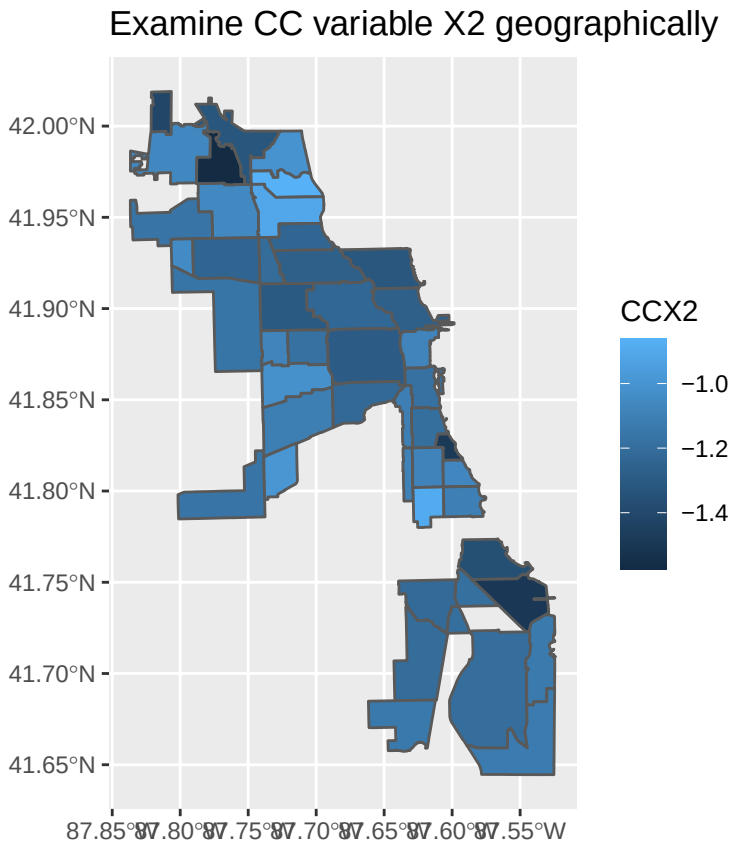
The first canonical correlation variable for set 2 contrasts zip codes with high voluntary insurance rates with low voluntary insurance rates (high value is more voluntary), while the second CC variable reflect just the mean insurance rate for both types (high value is low for both voluntary and involuntary).

d

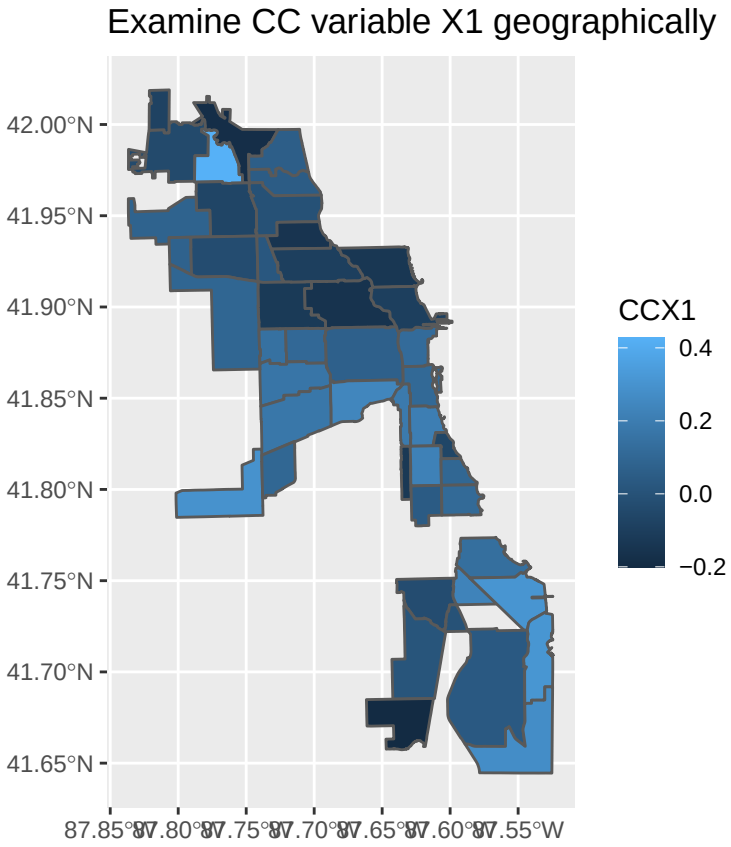
These exercises provide small insight into the second canonical variable from Set1. To see that this variable has some spatial consistency,categorize zip codes into quartiles based on this variable and color the zip codes in a reproduction of the map in Display 17.23. Repeat the color coding for the quartiles of the first canonical variable from Set1 to show its spatial pattern.

```
ins$CCX1 = CCX1
ins$CCX2 = CCX2


library(sf)
library(ggplot2)

chi_map <- read_sf("https://raw.githubusercontent.com/thisisdaryn/data/master/geo/chicago/Comm_Areas.ge

map_ins <- ins[,c("zip","CCX1","CCX2")]
ins_map2 <- merge(chi_map,map_ins, by.y = "zip", by.x = "area_numbe")
```

```
{ggplot(data = ins_map2, aes(fill = CCX2)) +
  geom_sf() +
  scale_fill_continuous() +
  ggtitle("Examine CC variable X2 geographically")}
```




```
{ggplot(data = ins_map2, aes(fill = CCX1)) +
  geom_sf() +
  scale_fill_continuous() +
  ggtitle("Examine CC variable X1 geographically")}
```



Both variable are showing strong geographical correlation!

e

```
p.asym(CCAins$cor, nrow(ins), 5, 2)
```

```
## Wilks' Lambda, using F-approximation (Rao's F):
##                 stat    approx df1 df2     p.value
## 1 to 2:  0.07980847 20.318190  10  80 0.000000000
## 2 to 2:  0.65684513  5.354896   4  41 0.001455323
```

According to the p-value from the test we found that the first pair of canonical variable is highly significant because the p-value is very low (very close to zero, 0.000000), which indicates a correlation of zipcodes with more poor, minority, low income residents and higher fire and theft rates with a relactive preponderance of involuntary (FAIR) insurance policies relative to voluntary policies. But the we can not that sure about the second pair of canonical variables because the p-value is relatively higher (0.0014) than the first one.

This is harder to understand the pattern of the insurance rate, but the suggestion is that the lowest overall insurance rates are central and in the northwest.