36-402/608 Homework #8 Solutions 3/18

1. Pig fat (50 points)

   (a) Look at Sleuth problem 17.08 and the answer on page 528. Load the data from ex1708.csv. Verify that you cannot run `step(lm(fat ., data=fat), direction="backward")` (even after correcting for whatever you called your data.frame and the upper/lower case of your variables). Although R could have sensible carried out this analysis, why do you think it did not?

   There are more variables than cases, which doesn't make sense in regression.

   (b) Starting with my code, add enough code to carry out the forward selection to the point where the AIC starts to rise. Turn in the `summary(lm())` of the best model using forward selection with AIC. (Note: this code is corrected from the originally supplied code, which had an error. You will not be penalized for replicating the error.)

```
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 = 48.16103
```

   The next two steps:

```
aic3 = rep(NA,13)
for (i in (1:13)[-c(one,two)]) aic3[i]=AIC(lm(fat$fat~fat[,one+1]+fat[,two+1]+f
three = which.min(aic3)
cat("+m", three, " = ", aic3[three], "\n", sep="")
# +m2 = 46.67967
aic4 = rep(NA,13)
for (i in (1:13)[-c(one,two,three)])
  aic4[i]=AIC(lm(fat$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 = 47.03791
```

   lower then raise the AIC, so we need 3 variables, m8, m4, and m2.

```
summary(lm(fat~m8+m4+m2, fat))
#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 0.139441
#
#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
```

It's OK to include a variable like m2 that is "non-significant" because the AIC model select procedure finds that it does add to the model fit beyond the level of noise.

(c) Carry out PCA on the pig fat dataset. Turn in the cumulative variance percents and the loadings (rotation) for the first 4 PCs.

```
pcfat = prcomp(fat[,-1]) # Need -1 to not include the outcome!!
#
cumsum( round( 100 * pcfat$sdev[1:4]^2 / sum(pcfat$sdev^2), 2) )
# [1] 86.50 90.88 94.58 96.59
#
round(pcfat$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) Let k be the number of variables chosen in part b. Create k new variables that are suggested by the first k PCs, but are simpler and more interpretable. Turn in the R code to create the variables and the `summary(lm())` that regresses the pig fat on the k new variables.

With $k = 2$, I chose the mean of all variables and twice m2 plus m1 minus m11 as reasonable surrogates for the exact principal components.

```
fat$PC1 = apply(fat[,-1],1,mean)
fat$PC2 = 2*fat$m2+fat$m1-fat$m11
summary(lm(fat~PC1+PC2, fat))
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  3.46266    1.52256   2.274   0.0490 *
# PC1          1.16013    0.06273  18.494 1.81e-08 ***
# PC2         -0.08160    0.02632  -3.100   0.0127 *
# Residual standard error: 1.484 on 9 degrees of freedom
# Multiple R-squared: 0.9802,     Adjusted R-squared: 0.9758
# F-statistic: 223.2 on 2 and 9 DF,  p-value: 2.144e-08
```

(e) Write a paragraph explaining what you think the pig farmers should do in practice to predict pig fat from MRI.

Looking at the adjusted R^2, the two PCs explain about the same variance the three original variables chosen using the backward stepwise procedure. In practice, we need to know whether making multiple MRI measurements is expensive. If not, making all of the 13 measurements and using the two PCs is a good method for the farmers. But more likely, it is cheaper to make one (m8) or two (m8+m4) measurements and use regression to get nearly the same results.

2. Insurance (50 points)

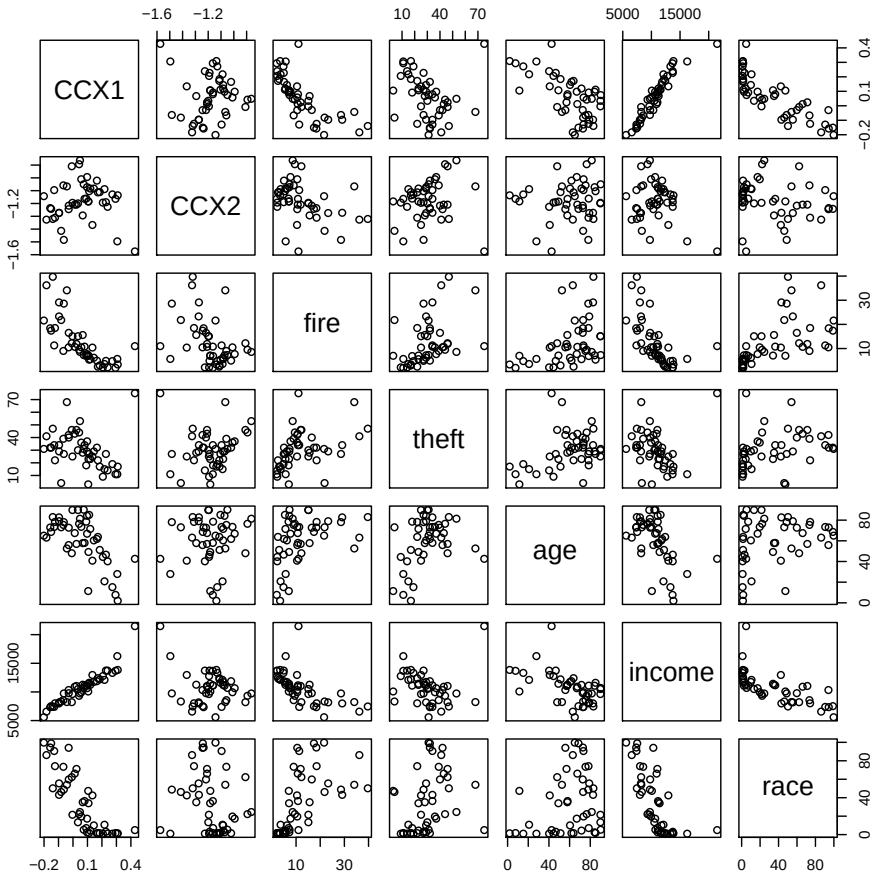
Do Sleuth problem 17.14.

(a) Perform CCA

```
ins=read.csv("ex1714.csv")
names(ins)=casefold(names(ins))
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:  0.65684515  5.354896   4  41 0.001455324
```

(b) Pairs plot of X and the new X canonical correlation variables

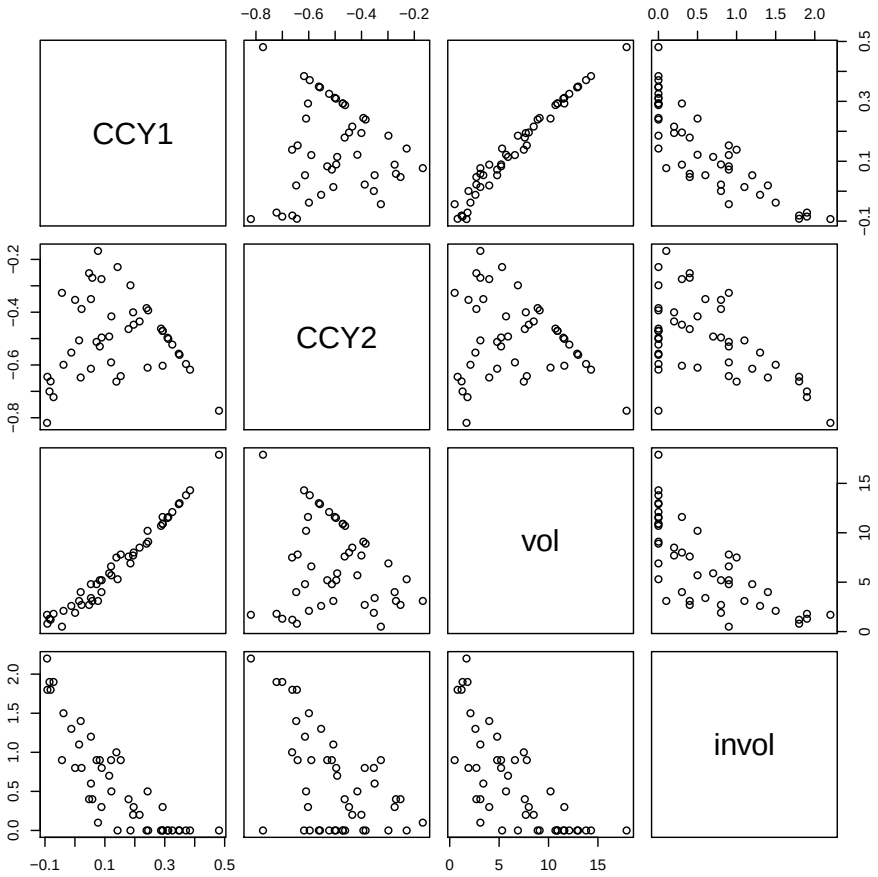
```
CCX1 = as.matrix(ins[,set1]) %*% as.matrix(CCAins$xcoef[,1])
CCX2 = as.matrix(ins[,set1]) %*% as.matrix(CCAins$xcoef[,2])
pairs(cbind(CCX1, CCX2, ins[,set1]))
dev.copy(pdf, "HW8p2b.pdf"); dev.off()
```



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).

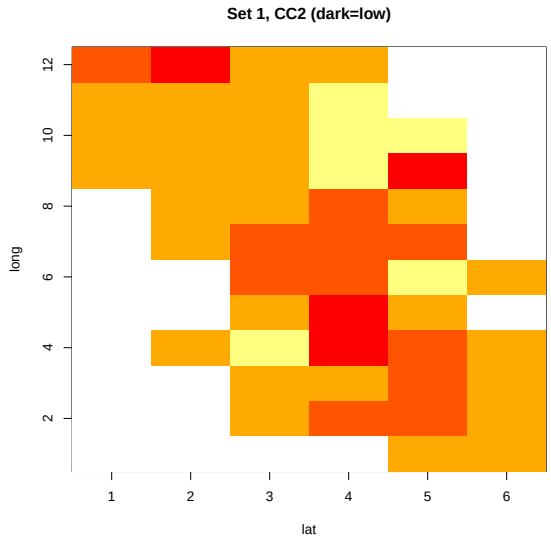
(c) Pairs plot of X and the new X canonical correlation variables

```
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]))
dev.copy(pdf, "HW8p2c.pdf"); dev.off()
```

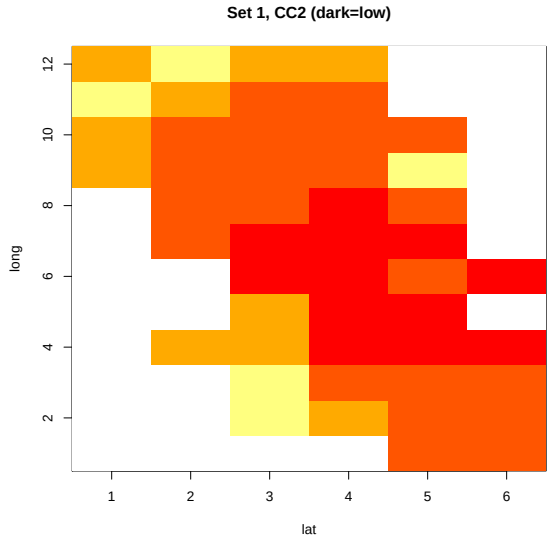


The first canonical correlation variable for set 2 contrasts zip codes with high voluntary insurance rates with low voluntary insurance rates (high value is more voluntary), while the second CC variable reflect just the mean insurance rate for both types (high value is low for both voluntary and involuntary).

(d) Examine CC variable X2 geographically



Then for CCX1:



Both show strong geographical correlations.

(e) Summary

The first pair of canonical variables has a correlation of 0.84 ($p < 0.0001$, F=20.3 with 10,80 df). This represents a correlation of zipcodes with more poor, minority, low income residents and higher fire and theft rates with a relative preponderance of involuntary (FAIR) insurance policies relative to voluntary policies. This supports, but does not prove (because it is a correlation study rather than a randomized experiment) the idea that poor, older and minority residents are less likely to obtain insurance policies voluntarily.

The second pair of canonical variables has a correlation of 0.58 ($p = 0.0014$, F=5.35, df=4,41). This is harder to understand, but the suggestion is that the lowest overall insurance rates are central and in the northwest.