Stat666 Homework 4

Arthur Lui

5 November 2014

1 PSA

1.1 Is there a significant relationship between the chemical abundances (x's) and the source contributions(y's)? What is $\hat{B}$?

The Wilk's Lambda statistic, $\Lambda_{p=8,\nu_H=17,\nu_E=181} \approx 0$. The corresponding F-statistic of this Wilk's Lambda statistic is $F_{df_1=136,df_2=1281.57}$ =565, with p-value $\approx 0 < .05$. Since the p-value for the Wilk Lambda's statistic is below .05, we conclude that there is a significant relationship between the chemical abundances and the source contributions. Table 1 displays $\hat{B}$.

Table 1: $\hat{B}$

	SummerSecondary	WinterSecondary	Mobile	SteelMill	Soil	ZincSmelter	CopperSmelter	LeadSmelter
	0.0715	0.0676	-0.1314	-0.0612	0.0008	0.0647	0.0653	-0.0222
Na	0.1110	0.0579	0.2063	0.1416	-0.0232	-0.0840	-0.0427	0.0890
Mg	-0.6959	-0.5312	0.5515	-0.7521	0.1481	0.3522	0.0569	-0.7466
Al	3.5898	-1.4080	-3.1404	-0.6968	1.6528	1.4101	1.9372	-0.3153
Si	-3.2218	0.6449	-1.2583	-0.2887	3.5591	-0.6052	-0.8002	0.2578
P	-0.9760	-0.7544	0.6328	0.6868	-0.0778	-0.5533	-0.4208	0.4394
S	3.1278	0.0060	0.3484	-0.0292	-0.1221	-0.2226	-0.2942	0.0015
Cl	-0.0375	0.0425	0.0249	0.1020	-0.0117	-0.0493	-0.0345	0.1148
K	-1.4285	0.1054	0.0687	-0.1528	-0.0678	-0.0623	-0.0985	-0.1428
Ca	-0.7237	-0.9349	17.6133	-0.6176	-0.5061	0.2763	0.3239	0.0366
Fe	-0.9849	-0.1874	-0.6292	7.5202	-0.2413	-0.0872	-0.0810	-0.4486
Cu	-0.0893	0.0689	-0.6229	-0.3232	-0.0435	-0.7106	6.0879	-0.1707
Zn	-0.0619	-0.4596	-1.2877	0.2983	0.0699	7.2220	-0.0494	0.7305
Ba	-0.4546	-0.6543	-0.7849	1.5762	0.0656	0.2002	0.1206	1.0838
Pb	0.1238	0.2126	-4.8853	-0.9928	-0.4061	-0.5260	0.2593	6.9933
OC	0.0590	0.0374	0.0454	-0.0211	0.0141	0.0438	0.0520	-0.0137
EC	-0.1999	0.0372	0.4420	0.1545	-0.0567	-0.1147	-0.1309	0.1071
SO4	0.3888	-0.0623	-0.1539	-0.0050	0.0466	0.0796	0.1157	0.0029
NO3	-0.1532	1.4164	-0.0240	0.0035	-0.0065	-0.0154	-0.0188	-0.0020

1.2 What is the essential dimensionality of the relationship between the x's and y's? Is that dimensionality directly evident from an inspection of $\hat{B}$?

The sum of the eigenvalues $\lambda_1 \& \lambda_2$ (substantially) accounts for 94% of the sum of the eigenvalues. So, the essential dimentionality of the realtionship between X and Y is 2. This is not directly evident from inspecting B1 because B1 is large and seeing trends within the matrix is difficult. (See Table2.)

Table 2: Eigenvalues of $E^{-1}H$

	1	2	3	4	5	6	7	8
Eigenvalues	1787.3709	614.7275	56.5317	43.0534	25.5563	9.3331	2.1871	0.7133
Proportions	0.7038	0.2421	0.0223	0.0170	0.0101	0.0037	0.0009	0.0003
Cummulative	0.7038	0.9459	0.9682	0.9851	0.9952	0.9989	0.9997	1.0000

1.3 Use canonical correlation analysis to identify and interpret the important dimensions of the relationship between the xs and the ys.

From Tables 3 and 4, we conclude that as Sulphur increases, Summer Secondary and Soil increase. Also, as Sulphur and Silicon increase, Soil and Summer Secondary increase.

Table 3: Standardized Canonical Coefficients for Y

	c1	c2
SummerSecondary	-1.0466	-0.3965
WinterSecondary	-0.0317	-0.0218
Mobile	-0.0401	0.0259
SteelMill	-0.0555	0.0514
Soil	-0.5479	0.7700
ZincSmelter	-0.1500	-0.1814
CopperSmelter	0.3269	0.1709
LeadSmelter	0.0085	-0.0694

Table 4: Standardized Canonical Coefficients for X

	d1	d2
Na	-0.0005	-0.0002
Mg	-0.0004	-0.0006
Al	-0.0138	-0.0171
Si	-0.0630	-0.1243
P	0.0005	-0.0007
S	-0.1553	0.0730
Cl	0.0005	0.0003
K	0.0632	-0.0165
Ca	0.0007	0.0010
Fe	-0.0047	-0.0095
Cu	0.0532	-0.0309
Zn	-0.0216	0.0276
Ba	-0.0005	0.0019
Pb	0.0079	0.0108
OC	-0.0008	-0.0004
EC	0.0006	0.0008
SO4	-0.0267	0.0002
NO3	0.0033	0.0009

1.4 **Traditionally, experts have assumed that heavy metals are important to the understanding of some pollution sources. In terms of the linear regression considered here, is the heavy metal Pb an important factor in the overall prediction of pollution source emissions? That is, can we drop Pb from the list of measured predictors without losing significant predictive ability?**

The Wilk's Lambda statistic for this full-reduced model is $\Lambda_{p=8,\nu_H=1,\nu_E=181} = 0.09731$. Its corresponding F-statistic is $F_{8,174} = 201.7671$. This statistic has a p-value $\approx 0 < .05$. So, we reject the null hypothesis that Pb (lead) is not an important factor in the overall prediction of pollution source emissions.

2 Body Fat

2.1 **Although the two ys are very highly correlated, there is some concern that each of these two measures of body composition may have different relationships with the xs. Is there a reason to be concerned about the ys being different in some sense, or are they responding similarly to the xs?**

Although the two ys are very highly correlated, the eigenvalues of $E^{-1}H$ reveals that there is only one essential dimension. So, the ys are responding similarly to the xs. (See Table 3.)

Table 5: Eigenvalues of $E^{-1}H$

	1	2
Eigenvalues	2.7462	0.0277
Proportions	0.9801	0.0199

2.2 For quick and inexpensive assessment of body fat and body density, it is not practical to measure all 13 of the predictors. Alternatively, we would like to have a subset of predictors that adequately predicts the bivariate response. Use a backward selection to obtain a subset of the predictor variables that adequately predicts the body fat related measurements (y1 and y2).

The backward selection algorithm chose x_2 (weight), x_6 (abdomen circumference), x_{11} (biceps circumference), and x_{13} (wrist circumference) as adequate predictors for body fat related measurements (y_1 and y_2).

2.3 Using canonical correlation analysis, describe and interpret the multivariate relationship between the body fat related measurements and your reduced set of predictors.

From Tables 6 and 7, we see that there is only one significant eigenvalue. As only one eigenvalue is twice that of the others. So, we conclude that body fat (y2) increases as abdomen circumference (x6) increases.

Table 6: Standardized Canonical Coefficients for Y

	c1	c2
y1	0.0190	-0.0190
y2	-0.1317	-0.0189

Table 7: Standardized Canonical Coefficients for X

	d1	d2
x2	2.4492	1.7200
x6	-6.1575	-0.9160
x11	-0.5968	-1.2145
x13	0.7398	0.8494

3 PSA Code:

```
options("width"=150)
library(xtable)
# PSA
X <- as.matrix(read.table("PSAData.txt",header=T))
X <- cbind(1,X)

Y <- as.matrix(read.table("PSAContributions.txt",header=T))

#1
n <- nrow(X)
r <- ncol(X)
q <- r-1
p <- ncol(Y)

B <- solve(t(X) %*% X) %*% t(X) %*% Y
XB <- X %*% B
Y.XB <- Y-XB
E <- t(Y.XB) %*% Y.XB

xtab.B <- xtable(B,digits=4,caption="$\\hat{\\bm B}$")
y.bar <- apply(Y,2,mean)
H <- t(Y) %*% Y - n * y.bar %*% t(y.bar)

lam <- det(E) / det(E+H)

# lam can also be calculated this way:
s <- min(p,q)
l <- eigen(solve(E,H))$values[1:s]
#lam <- prod(1/(1+l))
V <- sum(l/(1+l)) # Pillai


lam.to.F <- function(lam,p,vh,ve) { # Approximation
  t <- sqrt( (p^2*vh^2-4) / (p^2+vh^2-5) )
  L <- lam^(1/t)
  w <- ve + vh - (p+vh+1)/2

  df <- c(p*vh, w*t-(p*vh-2)/2)

  F.stat <- (1-L) / L * df[2]/df[1]
  p <- pf(F.stat,df[1],df[2],lower.tail=F)
  out <- c(F.stat,df,p)
  names(out) <- c("F.stat","df1","df2","p.val")

  out
}

V.2.F <- function(V,s,N,m) { # Approximation
  F.stat <- (2*N + s + 1) / (2*m + s + 1) * V/(s-V)
  df <- s * (c(m,N) + s + 1)
```

```
  p <- pf(F.stat,df[1],df[2],lower.tail=F)

  out <- c(F.stat,df,p)
  names(out) <- c("F.stat","df1","df2","p")

  out
}

F.L <- lam.to.F(lam,p=p,vh=q-1,ve=n-r)
F.V <- V.2.F(V,s=min(p,q),N=(n-q-p-2)/2,m=(abs(p-1)-1)/2)

# p.val < .05 => Yes, there is a significant contribution.
#                H_0 B1 = 0 is rejected.

#2
S <- E / (n-r)
var.B.vec <- S %x% solve(t(X)%*%X)
B0 <- B[1,]
B1 <- B[-1,]

# Another.Lambda <- det(cov(cbind(Y,X[,-1]))) / (det(var(X[,-1])) *
# det(var(Y))). But why is it different?  We reject H0, and the first
# eigen value does not completely dominate the other eigen values.
# (See: l/sum(l)). l1+l2 accounts for 94% of the eigen values, which
# substantially dominates the other eigen values.  So, the essential
# dimensionality of the relationship between X and Y is 2.  This was
# not directly evident from inspecting B1, because B1 is large and it
# is difficult to see trends.

e <- l-1
eig <- e/sum(e)
cumm <- apply(matrix(1:length(l)),1,function(x) sum(eig[1:x]/sum(eig)))

#3: Canonical Correlation
# I don't know what to do for this problem YET.
#   OBJECTIVE: Summarize the linear relationship betweem
#              the two groups of variables, Y & X.
Syy <- var(Y)
Sxx <- var(X[,-1])
Syx <- var(Y,X[,-1])
Sxy <- var(X[,-1],Y)
A <- solve(Syy) %*% Syx %*% solve(Sxx) %*% Sxy
r2 <- eigen(A)$values[1:s]
R2 <- prod(r2) # Should be the same as det(A)
#R2 <- det(A) # Should be the same as product of the eigen values of A
# The first 6 r2's?

L.m <- apply(matrix(1:s),1,function(m) prod(1-r2[m:length(r2)]))
F.m <- t(apply(matrix(1:s),1,function(m)
         lam.to.F(L.m[m],p=q-m+1,vh=p-m+1,ve=n-m-p)))
```

```
#r2.l <- r2/(1-r2)
#r2.l / sum(r2.l)

A <- solve(Syy,Syx) %*% solve(Sxx,Sxy)
B <- solve(Sxx,Sxy) %*% solve(Syy,Syx)

a <- eigen(A)$vector
b <- eigen(B)$vector

(c <- diag(sqrt(diag(Syy))) %*% a)
(d <- diag(sqrt(diag(Sxx))) %*% b)

c <- Re(c[,1:2])
d <- Re(d[,1:2])

rownames(c) <- colnames(Y)
rownames(d) <- colnames(X[,-1])
colnames(c) <- paste0("c",1:ncol(c))
colnames(d) <- paste0("d",1:ncol(d))

c <- c[,1:min(ncol(c),ncol(d))]
d <- d[,1:min(ncol(c),ncol(d))]
# Why is my answer different???

#4:
Xr <- X[,-which(colnames(X)=="Pb")]
Br <- solve(t(Xr) %*% Xr) %*% t(Xr) %*% Y
Y.XBr <- Y - Xr%*%Br
Er <- t(Y.XBr) %*% Y.XBr

L.full.red <- det(E) / det(Er)
h <- 1
F.f.r <- lam.to.F(L.full.red,p=p,vh=h,ve=n-r)

# p.val < .05 => Pb is important in overall
#                prediction of pollution source emissions.
```

4 Body Fat Code:

```
# I have questions at #5,#6,#7
#options("width"=150)

lam.to.F <- function(lam,p,vh,ve) { # Approximation
  t <- sqrt( (p^2*vh^2-4) / (p^2+vh^2-5) )
  L <- lam^(1/t)
  w <- ve + vh - (p+vh+1)/2

  df <- c(p*vh, w*t-(p*vh-2)/2)

  F.stat <- (1-L) / L * df[2]/df[1]
  p <- pf(F.stat,df[1],df[2],lower.tail=F)
  out <- c(F.stat,df,p)
  names(out) <- c("F.stat","df1","df2","p.val")

  out
}

V.2.F <- function(V,s,N,m) { # Approximation
  F.stat <- (2*N + s + 1) / (2*m + s + 1) * V/(s-V)
  df <- s * (c(m,N) + s + 1)
  p <- pf(F.stat,df[1],df[2],lower.tail=F)

  out <- c(F.stat,df,p)
  names(out) <- c("F.stat","df1","df2","p")

  out
}

#5:
dat <- as.matrix(read.table("bodyfat.txt",header=F))
colnames(dat) <- c(paste0("y",1:2), paste0("x",1:13))

#y1  = Density determined from underwater weighing
#y2  = Percent body fat
#x1  = Age (years)
#x2  = Weight (lbs)
#x3  = Height (inches)
#x4  = Neck circumference (cm)
#x5  = Chest circumference (cm)
#x6  = Abdomen 2 circumference (cm)
#x7  = Hip circumference (cm)
#x8  = Thigh circumference (cm)
#x9  = Knee circumference (cm)
#x10 = Ankle circumference (cm)
#x11 = Biceps (extended) circumference (cm)
#x12 = Forearm circumference (cm)
#x13 = Wrist circumference (cm)

Y <- dat[,1:2]
```

```
X <- dat[,-(1:2)]

n <- nrow(Y)
p <- ncol(Y)
q <- ncol(X)
r <- q+1

X <- cbind(1,X)
B <- solve(t(X)%*%X) %*% t(X)%*%Y

E <- t(Y-X%*%B) %*% (Y-X%*%B)
S <- E / (n-r)
y.bar <- apply(Y,2,mean)
H <- t(Y) %*% Y - n*y.bar%*%t(y.bar)
lam <- det(E) / det(E+H)


s <- min(p,q)
l <- eigen(solve(E,H))$values[1:s]
#lam <- prod(1/(1+l))
V <- sum(1/(1+l))

F.L <- lam.to.F(lam,p=p,vh=q-1,ve=n-r)
F.V <- V.2.F(V,s=min(p,q),N=(n-q-p-2)/2,m=(abs(p-1)-1)/2)


Syy <- var(Y)
Sxx <- var(X[,-1])
Syx <- cov(Y,X[,-1])
Sxy <- t(Syx)
A <- solve(Syy) %*% Syx %*% solve(Sxx) %*% Sxy
r2 <- eigen(A)$values[1:s] # Squared Canonical Correlation in SAS
                           # the sqrt of r2 is the canonical corr in SAS
R2 <- prod(r2) # Should be the same as det(A)
#R2 <- det(A) # Should be the same as product of the eigen values of A

L.m <- apply(matrix(1:s),1,function(m) prod(1-r2[m:length(r2)]))
F.m <- t(apply(matrix(1:s),1,function(m) lam.to.F(L.m[m],p=q-m+1,vh=p-m+1,ve=n-m-p)))

r2.l <- r2/(1-r2) # This is the "eigenvalues" of E^(-1)H in SAS = l-1. Why?
prop.r2.l <- r2.l / sum(r2.l) # 98% => essentially one dimension

#6:
back.sel <- function(Y,X,a=.05) {
  n <- nrow(Y)
  p <- ncol(Y)
  X <- cbind(1,X)
  YTY <- t(Y) %*% Y

  get.L <- function(ve) {# note: p=p, vh=1, a=a=.05
    F.stat <- qf(1-a,p,ve-p+1)
```

```
    L <- 1 / (F.stat*p/(ve-p+1)+1)
    L
  }

  get.L.xi.rest <- function(i,x,E.full) {
    x <- x[,-i]
    b <- solve(t(x)%*%x, t(x)%*%Y)
    E.red <- YTY - t(x%*%b) %*% Y
    L <- det(E.full)/det(E.red)
    L
  }

  L.end <- get.L(n-r)
  L <- L.end + 1

  while (L > L.end) {
    r <- ncol(X)
    B <- solve(t(X)%*%X, t(X)%*%Y)
    E.full <- YTY - t(X%*%B) %*% Y
    #L.s <- apply(matrix(2:r),1,function(i) get.L.xi.rest(i,X,E.full))
    L.s <- apply(matrix(1:r),1,function(i) get.L.xi.rest(i,X,E.full))
    i <- which.max(L.s)
    L <- L.s[i]
    L.end <- get.L(n-r)
    if (L > L.end) {
      #X <- X[,-(i+1)]
      X <- X[,-i]
    }
  }

  X
}

new.X <- back.sel(Y,X[,-1])
#head(new.X)

# Question: Do I need to consider removing the intercept, or do I always
#           keep the intercept?

#7:
Sxx <- var(new.X[,-1])
Syx <- cov(Y,new.X[,-1])
Sxy <- t(Syx)
A <- solve(Syy) %*% Syx %*% solve(Sxx) %*% Sxy
r2 <- eigen(A)$values[1:min(ncol(new.X),ncol(Y))] # Squared Canonical Correlation in SAS
                          # the sqrt of r2 is the canonical corr in SAS
R2 <- prod(r2) # Should be the same as det(A)
#R2 <- det(A) # Should be the same as product of the eigen values of A

L.m <- apply(matrix(1:s),1,function(m) prod(1-r2[m:length(r2)]))
F.m <- t(apply(matrix(1:s),1,function(m) lam.to.F(L.m[m],p=q-m+1,vh=p-m+1,ve=n-m-p)))
```

```
r2.l <- r2 / (1-r2)
r2.l / sum(r2.l) # 99% in eig[1]

B.new <- solve(t(new.X) %*% new.X, t(new.X)%*%Y)



Syy <- var(Y)
Syx <- var(Y,new.X[,-1])
Sxy <- var(new.X[,-1],Y)
Sxx <- var(new.X[,-1])

A <- solve(Syy,Syx) %*% solve(Sxx,Sxy)
B <- solve(Sxx,Sxy) %*% solve(Syy,Syx)

a <- eigen(A)$vector
b <- eigen(B)$vector

(c <- diag(sqrt(diag(Syy))) %*% a)
(d <- diag(sqrt(diag(Sxx))) %*% b)

rownames(c) <- c("y1","y2")
rownames(d) <- c("x2","x6","x11","x13")
colnames(c) <- c("c1","c2")
colnames(d) <- c("d1","d2","d3","d4")

c <- c[,1:min(ncol(c),ncol(d))]
d <- d[,1:min(ncol(c),ncol(d))]
# Why is my answer different???
```